

**SOCIAL ONTOLOGY AND AGENCY**  
**METHODOLOGICAL HOLISM NATURALISED**

A dissertation by

**ANTTI JUSSI SAARISTO**

Department of Philosophy, Logic and Scientific Method

London School of Economics and Political Science

Presented to the

**UNIVERSITY OF LONDON**

for the degree of

**DOCTOR OF PHILOSOPHY**

March 2007

UMI Number: U615510

All rights reserved

INFORMATION TO ALL USERS

The quality of this reproduction is dependent upon the quality of the copy submitted.

In the unlikely event that the author did not send a complete manuscript and there are missing pages, these will be noted. Also, if material had to be removed, a note will indicate the deletion.



UMI U615510

Published by ProQuest LLC 2014. Copyright in the Dissertation held by the Author.  
Microform Edition © ProQuest LLC.

All rights reserved. This work is protected against  
unauthorized copying under Title 17, United States Code.



ProQuest LLC  
789 East Eisenhower Parkway  
P.O. Box 1346  
Ann Arbor, MI 48106-1346

THESES .

F.

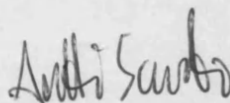
9527 .



1278542

I, Antti Jussi Saaristo, hereby confirm that the work presented in this thesis is my own. Where information has been derived from other sources, I confirm that this has been indicated in the thesis.

Berkeley, 23<sup>rd</sup> February 2007

  
Antti Saaristo





## ABSTRACT

Contemporary philosophy of the social sciences is dominated by methodological individualism. Intentional agency is assumed to be conceptually and explanatorily prior to social facts and social practices. In particular, it is generally thought that denials of methodological individualism are bound to include ontologically unnatural and, thereby, unacceptable views. This dissertation provides a comprehensive criticism of this orthodoxy.

Part I argues that social facts do not have to be understood as aggregates of actions and attitudes of essentially asocial individuals. Rather, the construction of social facts requires that acting as a member of a group rather than as a disparate individual is a fundamental building block of social reality and social facts. This idea is explicated in the anti-individualistic terms of the theory of collective intentionality.

Part II tackles the accusation that the theory of collective intentionality is indefensibly anti-naturalistic in the sense that its picture of humans is essentially incompatible with evolutionary biology. This accusation is answered in terms of detailed analyses of evolutionary models of human sociality and empirical studies of the nature of social action. Part II concludes that it is actually the methodologically individualistic picture of social action as strategic individual action that is unacceptable. The theory of collective intentionality is compatible with and supported by scientific naturalism.

Part III, then, defends full-blown methodological holism. It is argued that intentional action and agency as we know them actually require that individual agents (qua agents and not qua physical objects) are essentially constituted by social practices. Intentional action must be explained and understood in terms of social practices. However, this view is argued to be perfectly naturalistic both in the sense of not assuming any ontologically suspect entities and in the sense of being supported by the natural sciences. Indeed, it is the individualistic orthodoxy that has to apply unnatural notions.

## TABLE OF CONTENTS

Acknowledgements.....	6
Introduction.....	8
<b>PART I: SOCIAL ONTOLOGY.....</b>	<b>17</b>
Introduction to Part I.....	18
I.1 The Objectivity of Social Facts.....	20
I.1.1 Searle on the Objectivity of Social Facts.....	21
I.1.2 Problems with Searle’s Account.....	22
I.1.3 Towards a Durkheimian Account of the Objectivity of Social Facts.....	25
I.1.4 Explicating the Objectivity of Social Facts.....	29
I.1.5 Are Social Facts <i>Sui Generis</i> ?.....	32
I.1.6 Conclusion.....	39
I.2 Collective Intentionality as the Driving Force of Social Reality.....	41
I.2.1 Searle’s Building Blocks of Social Reality.....	41
I.2.2 Searle on Collective Intentionality.....	46
I.2.3 Gilbert on Collective Intentionality.....	48
I.2.4 Tuomela on Collective Intentionality.....	60
I.3 Conclusion.....	71
<b>PART II: THE EVOLUTION AND PSYCHOLOGY</b>	
<b>OF <i>HOMO SOCIOLOGICUS</i>.....</b>	<b>78</b>
Introduction to Part II.....	79
II.1 Evolutionary Egoism and Altruism.....	81
II.1.1 Group Selection.....	81
II.1.2 The Ontological Status of Group Selection.....	90
II.2 The Psychology of Social Action.....	95
II.2.1 An Evolutionary Perspective.....	96
II.2.2 Social Dilemmas: The Utility Transformation Rules.....	98
II.2.3 Beyond Philosophical Egoism.....	110
II.2.4 Collective Intentionality and Co-operation.....	115
II.3 Collective Intentionality and Empirical Social Science.....	123
II.3.1 Social Sanctions.....	123
II.3.2 The Discovered Preference Hypothesis.....	125
II.3.3 The Social Identity Approach.....	129
II.4 Conclusion.....	135

<b>PART III: ACTION AND AGENCY</b> .....	<b>139</b>
Introduction to Part III.....	140
III.1 Intentional Action and Its Explanation.....	145
III.1.1 The Standard View.....	145
III.1.2 The Logical Connection Argument.....	150
III.1.3 Davidson’s Alleged Refutation of the Logical Connection Argument.....	155
III.2 The Multiple Realisability Argument for Mental Causation.....	171
III.2.1 Introduction.....	171
III.2.2 The Multiple Realisability Argument.....	172
III.2.3 Kim on Explanatory Exclusion.....	185
III.2.4 Conclusion.....	193
III.2.5 Postscript: The Causalist’s Last Hope.....	197
III.3 What Is Intentionality Anyway? The Problem of Rule-Following.....	202
III.3.1 Introduction.....	202
III.3.2 Meaning and Rules.....	207
III.3.3 Kripke’s Scepticism, the Naïve Communitarian View and McDowell’s Quietism.....	217
III.4 Digging below the Bedrock.....	229
III.4.1 Introduction.....	229
III.4.2 The Indispensability of Collective Agency.....	231
III.4.3 Social Practices and Implicit Normativity.....	237
III.5 Explanation of Action and Collective Intentionality.....	246
III.5.1 The True Form of Intentional Explanations.....	246
III.5.2 The Status of Collective Intentionality.....	251
III.5.3 Tying Up Loose Ends.....	254
III.6 Conclusion.....	264
<b>APPENDIX: DURKHEIM’S GOD       or THE SOCIOPHILOSOPHY MANIFESTO</b> .....	<b>266</b>
Introduction.....	267
A.1 When the Community is Mistaken.....	268
A.1.1 Social Relativism and Talking <i>about</i> the World.....	268
A.1.2 The Brandom – Kusch Dissension: A Dissolution.....	273
A.2 Durkheim’s God: Concrete Practices and Direct Realism.....	276
<b>REFERENCES</b> .....	<b>287</b>

## ACKNOWLEDGEMENTS

Studying for a Ph.D. is a lonely journey. There is no escape from hard work, feelings of inadequacy and, frankly, sheer desperation. I could never have endured the project without the aid and encouragement I have received.

First, I wish to thank my family and friends in Finland for their endless support. Had they not been there for me, my project would have failed before it even began.

The Department of Philosophy, Logic and Scientific Method and the Centre for Philosophy of Natural and Social Science (CPNSS) at LSE, together with the philosophy departments of the other University of London colleges, form a philosophical community which remains, I believe, without an equal in terms of activity and quality even in the world scale. For a student of the foundations of the social sciences LSE in general and the CPNSS in particular offer an ideal intellectual environment. Similarly, my extended stays at UC Berkeley and the University of Helsinki have helped me enormously.

My supervisors J. McKenzie Alexander and Nancy Cartwright have read, commented and discussed with me different drafts and versions of my thesis and guided my studies also otherwise. Thank you.

My supervisors are not the only people whose thoughtful comments and arguments have influenced my work. Words cannot adequately express how much I have learned from discussions, debates and e-mail correspondence with Caroline Baumann, Philipp Beckmann, Lorenzo Bernasconi, Gregor Betz, Richard Bradley, John B. Davis, Emrah Cevik, Philipp Dorstewitz, Hanspeter Fetz, Margaret Gilbert, Raul Hakli, Jussi Haukioja, Tomi Kokkonen, Jaakko Kuorikoski, Steven Lukes, Edward F. McClennen, Eleonora Montuschi, Pekka Mäkelä, Uskali Mäki, Janne Mäntykoski, Philip Pettit, Mauro Rossi, David Rönnegard, Hans Bernhard Schmid, Armin Schulz, John R. Searle, Elliott Sober, Georg Theiner, Raimo Tuomela, Bruno Verbeek, Stuart Yasgur, Petri Ylikoski and indeed with all my friends, students, teachers and colleagues at LSE, UC Berkeley, the University of Helsinki and numerous conferences. Moreover, practically all the ideas in this dissertation have first been developed in long discussions with Damien Fennell in different cafes, restaurants and pubs around LSE. I am obliged. This final version has also benefited greatly from the criticisms and comments by my two examiners, Professors John Dupré and Martin Kusch, and it would most likely have remained unwritten without Maria Kreander's caring support.

Finally, although the intellectual and emotional support and encouragement by my wonderful friends has been absolutely necessary for my research, in our actual world it unfortunately is not sufficient. Thus I gratefully acknowledge also the financial support I have received from the Helsingin Sanomat Centenary Foundation, the Jenny and Antti Wihuri Foundation, the Emil Aaltonen Foundation, the Academy of Finland and the London School of Economics.

## INTRODUCTION

The thesis argued for over and over again in this dissertation is that the social world in general and meaningful social actions in particular cannot exhaustively be explained and understood as consisting of conceptually prior intentional actions of disparate, ultimately asocial individuals seeking to fulfil their private goals. Rather, we must admit that often individual actions are accurately understandable only when the acting individual is seen essentially as a member of a group, seeking to realise the goals of the group. In truly social action individual roles are derived from the more fundamental collective project. What is more, participation in social practices must ultimately be seen as constitutive of individual agency, and in this sense the social really is conceptually prior to the individual. However, this priority of the social is argued to be completely naturalistic both in the sense of building on a strictly naturalistic ontology (the world is governed by blind causal laws and consists entirely of physical particles in fields of force, as John R. Searle likes to put this in) and being supported by our best understanding of the natural sciences (again, as Searle voices this, our theory of agency and society must fit into the same picture with our fundamental and evolutionary biology). In sum, this dissertation is a defence of naturalised methodological holism *vis-à-vis* social ontology and agency.

To understand the proper scope of the argumentation it should be kept in mind that all the time my sole aim is to argue for naturalised methodological holism in the above sense. This concentration in focus results in the following ambivalence. On the one hand it could be said that the focus of my dissertation is too narrow: I dedicate a book-length study to an issue many writers address in the introductory chapter of a journal article. On the other, however, the scope is extremely wide: a full-blown defence of my main thesis requires me to visit several debates and arguments that are traditionally seen as belonging to other fields of philosophy, and within the limits of this dissertation I of course cannot address all the aspects of those debates and arguments relevant to those other fields. Rather, I draw from such debates only to the extent it is necessary for a comprehensive treatment of my main thesis. Consequently, a reader with core interests different from mine might occasionally think that I fail to enter some interesting debates connected to the themes I discuss.

The answer both to the worry concerning narrowness and to the worry about wideness of my scope is this: Methodological individualism is so generally seen as the obviously correct approach to the social sciences that an extended defence of holism is

certainly called for. Moreover, precisely because holism is so generally thought of as an obviously unacceptable and unnaturalistic position, the construction of a comprehensive and coherent account of naturalised methodological holism and the explication of the connections the position has to certain other philosophical debates is a worthy philosophical accomplishment in itself, even if I have to omit some issues that are crucial for other philosophical projects. After all, the aim of this dissertation is to defend naturalistic methodological holism *vis-à-vis* social ontology and agency, not to provide a series of contributions to other fields of philosophy. However, I appreciate that it is probably rather difficult, or at least tiring, to follow an argument that requires a whole book to spell out. Hence in this introduction I explain shortly how the different chapters of my dissertation contribute to the main argument. This should make it evident that the scope of the dissertation is rather closely focused indeed. However, I also point out how, while developing the main line of thought, the different chapters contain several independent philosophical results that are also important as such and not only as part of my overall defence of naturalised methodological holism.

Part I of this dissertation sets the stage for the more substantial and more controversial arguments of Parts II and III. Chapter I.1 opens this task by asking in what sense the social world can be said to be objective when it is unquestionably also a human construction. Building on Searle's well-known distinction between ontological and epistemological objectivity (Sections I.1.1 and I.1.2), I construct a Durkheimian (I.1.3 and I.1.4) analysis of social facts where what matters is that social facts are *collectively* constructed and reproduced and hence objective from any *individual* point of view. This is the very starting point of my methodological holism. The most central philosophical results here are the novel criticisms of Searle's influential account and its Durkheimian improvement, which also removes the unfortunate metaphysical burden traditionally but nonetheless somewhat unfairly attached to Durkheim's notion of social facts. However, the exact nature of the essentially *collective* construction of social facts in the improved account is found to be open to several interpretations.

Thus, Section I.1.5 analyses the major ways of understanding collective construction by introducing three ways to understand social facts and discussing their interconnections. I call these accounts the Individualistic Account, the Wittgensteinian Account and the Durkheimian Account. The Individualistic Account holds that the actions required for the construction of objective social facts are collective merely in the sense of involving a number of essentially asocial individuals and their intentional interactions. The Wittgensteinian and Durkheimian Accounts reject this kind of straightforward



methodological individualism for different reasons. The Wittgensteinian Account holds that the Individualistic Account is false because individual intentional action already presupposes social practices, and thus individual action cannot be a fundamental building block of social reality that includes social practices. The Durkheimian Account, in contrast, holds that the Individualistic Account must be rejected because the construction and reproduction of sufficiently objective and robust social facts require that individuals act together in a more fundamental sense than the conceptual resources of the Individualistic Account allow. Thus, Wittgensteinianism and Durkheimianism are mutually logically independent: accepting one view does not require one to adopt also the other. Thus, the two ways of rejecting individualism must be discussed separately. The present dissertation defends full-blown methodological holism (*i.e.*, comprehensive rejection of the Individualistic Account) in the strong sense of defending both Durkheimianism (Part II) and Wittgensteinianism (Part III). The main philosophical result in Section I.1.5 is the explication of the two different ontologically naturalistic ways to challenge mainstream methodological individualism that are examined and defended in the rest of the dissertation.

Before Part II gives a full defence of Durkheimianism, an account of what, exactly, a contemporary version of Durkheimianism could look like is required. Thus, Chapter I.2 presents the so-called collective intentionality approach to social ontology as the major contemporary Durkheimian analysis of action sufficiently *collective* for constructing and reproducing social facts. More precisely, Section I.2.1 offers a critical reconstruction of Searle's theory of social reality and social facts that builds on collective intentionality. The anti-individualism in Searle's theory boils down to his theory of collective intentionality, analysed in I.2.2, as the irreducible *we-mode* of certain intentional states and intentional actions. The accounts of collective intentionality of the other major theorists of collective intentionality in the Durkheimian sense, Margaret Gilbert and Raimo Tuomela, are critically analysed in I.2.3 and I.2.4, respectively. The main philosophical result here is the detailed analysis of the three main accounts of collective intentionality, including their individual weaknesses and mutual compatibility. I argue that my final analysis of Tuomela's reason-based *we-intention* in I.2.4 captures also the anti-individualistic core of Searle's and Gilbert's notions of collective intentionality and defines what I mean by collective *we-intentionality*. Chapter I.3, then, sums up the collective intentionality view of social ontology, *i.e.*, the contemporary form of Durkheimianism in the sense I.1.5.

Normally the question of whether or not we really need to assume the reality of irreducible we-intentionality in this sense is addressed in terms of testing our intuitions about social action in different imaginable cases (see, in particular, Searle 1990, Gilbert 1989, Miller 2007 and indeed Chapter I.2). Thus the debate proceeds such that the Durkheimians suggest counterexamples to Individualistic Accounts of collective actions, and the individualists answer by arguing that in fact the Individualistic Account can handle the alleged counterexamples (in my view Bratman 1999 & Miller 2007 have a particularly strong case here). Instead of joining the industry of presenting my intuitions concerning hypothetical cases, I want to give a stronger, *conceptual* argument (see II.3 and III.5.2 for an argument that empirical research indeed cannot settle the issue) to show that the assumption of anti-individualistic collective intentionality is indeed *required* if we wish to account for collective action in general. Thus, Part II provides a badly-needed general argument<sup>1</sup> in favour of the reality of *sui generis*, irreducible we-intentionality, *i.e.*, contemporary Durkheimianism in the sense of Part I.

Part II begins this task by considering a possible attack on the naturalisticness of collective intentionality in terms of its apparent evolutionary implausibility. The basic evolutionary dynamics of competition and survival of the fittest seem to dictate that the kind of group-centred social solidarity implied by the theory of collective intentionality is likely to have been selected against in the course of evolutionary history (Chapter II.1). Section II.1.1 presents this problem in a precise form and offers the group selection theory of Elliot Sober and David Sloan Wilson (1998) as a solution to the problem. However, this dissertation is not a study in biology. Nor am I doing armchair evolutionary theory. Accordingly, Section II.1.2 explains that my argument is not tied to the specific view of biological evolution favoured by Sober and Wilson. Rather my aim is to show that evolutionary biology does not pose a conceptual or theoretical obstacle for the collective intentionality view. Thus, my arguments show that the view of social behaviour implied by the collective intentionality theory, *i.e.*, individuals acting essentially *qua* group members and the instrumental rationality of their actions being revealed primarily when rationality is addressed in collectivistic terms of what the group is doing and what is optimal for the group, is fully compatible with and even supported by our best understanding of evolutionary dynamics. This is a strong philosophical result as such and, moreover, a result that features prominently in the arguments of Part III.

---

<sup>1</sup> In contrast to most other writers in the field, Tuomela (*e.g.*, 2000, 2002, 2007) shares my preference for general, conceptual arguments over intuitive discussions concerning particular thought experiments. Accordingly, Part II builds largely on Tuomela's work, although the main argument is rather different from Tuomela's treatment of the issue.

However, Chapter II.1 talks only about the evolutionary plausibility of social patterns of bodily behaviour, not about forms and modes of psychological mechanisms. The question concerning the psychology associated with social behavioural patterns is addressed in Chapter II.2 that presents my general argument for the Durkheimian we-mode account and against the Individualistic Account. The animating idea is to adopt an evolutionary perspective that guides us to address the following evolutionary design problem (Section II.2.1): what kind of psychological mechanism – in particular, individual-mode or we-mode – is evolutionarily the most optimal mechanism to have evolved to produce the kind of social behaviour we observe every day and the evolutionarily plausibility of which was addressed in Chapter II.1? Section II.2.2 follows a prominent tradition in the study of social action and interprets this question as the question concerning the possibility of rational co-operation in social dilemma situations, such as the Prisoner's Dilemma. Sections II.2.3 and II.2.4 show that this framing of the question yields a very strong, general argument for the reality of collective we-intentionality by demonstrating the functional superiority of we-mode over individualistic I-mode. In my view this argument forms a major contribution to the theory of collective intentionality.

Note, however, that my arguments and problems are all the time conceptual and not empirical. After all, this is a philosophical study. I am not doing speculative evolutionary biology or psychology. This is especially clear when it is kept in mind that my defence of we-mode intentionality in Part II remains neutral regarding Wittgensteinianism, the topic of Part III. Chapter II.2 discusses evolutionary dynamics in general, *i.e.*, independently of whether the selection process behind collective intentionality is thought to be materialised in biological (non-Wittgensteinianism) or cultural processes (Wittgensteinianism).

Chapter II.3, then, makes a move towards empirical considerations by asking whether empirical social science is able to either falsify or verify the results of the conceptual argumentation in Chapter II.2. There are approaches in empirical social science that are directly hostile (II.3.2) to the collective intentionality theory and approaches that strongly support it (II.3.3). However, Chapter II.3 concludes that such approaches alone cannot settle the issue. Thus, we have to go with the general, conceptual arguments I have offered. Finally, Chapter II.4 concludes the defence of the Durkheimian we-intentionality of Part II.

Part III, then, takes up the issue of Wittgensteinianism. As in Part II, I motivate the discussion by starting from a fundamental problem: is the framework of intentional

agency, assumed in Parts I and II, justified in the first place? Chapter III.1 opens the discussion by analysing the nature and explanation of intentional actions. The easiest solution to my problem would be to hold simply that the framework does not form any kind of philosophical problem, because intentional states and meaningful actions simply belong to the fundamental furniture of the natural world and feature non-problematically in causal chains (intentionality, be it collective or individual, would be seen as a biologically primitive feature of the world, as Searle puts this). I call this common-sense view the Standard View (III.1.1).

However, the Standard View faces grave problems, particularly the so-called Logical Connection Argument (III.1.2) which seeks to demonstrate that intentional states are causally inefficacious and intentional explanations are not causal explanations. Donald Davidson is routinely cited as the philosopher who took up this challenge and refuted the Logical Connection Argument, rescuing thereby the Standard View. Section III.1.3 analyses Davidson's argumentation in detail and concludes, contrary to the prevailing view in the philosophical community, that Davidson's alleged refutation fails. Moreover, in demonstrating Davidson's failure, Section III.1.3 also shows that Jon Elster's influential interpretation of Davidson as the champion of the Standard View is inescapably incoherent. Although these are very strong and important results in contemporary action theory as such, the implication they have for my overall argument is that one cannot follow Davidson to the Standard View and simply assume that the framework of intentional agency (including collective intentionality) is not a problem for the strong naturalism I have committed myself to. Something else is required.

Chapter III.2 analyses another important attempt to defend the naturality of the framework of intentional agency, the essentially non-Davidsonian version of the Standard View which holds that mental states are causally efficacious qua mental – but in a way that does not challenge general naturalism (III.2.1). A very widely accepted view in the philosophical community is that the so-called Multiple Realisability Argument can deliver this kind of naturalistic defence of the Standard View and, thereby, the framework of intentional agency. However, Sections III.2.2 and III.2.3 argue that this popular programme is bound to fail in its own terms. Section III.2.4, then, suggests that this argumentation implies that to defend the framework of intentional agency we must reject the causalist Standard View, since both the Davidsonian argumentation and the Multiple Realisability Argument fail to defend it. Section III.2.5 generalises this result to show that in fact the Standard View consists essentially of five theses that cannot all be coherently accepted. One must go, and I suggest that it is the commitment to the causal

nature of the framework of intentional agency. Moreover, III.2.5 argues that the only real option to non-causalism is, *e.g.*, Nancy Cartwright's and John Dupré's metaphysical pluralism. In addition to revealing the failures of the celebrated Multiple Realisability Argument, the striking conclusion of Chapter III.2 is that to defend the framework – and, thus, collective intentionality – one must accept either non-causalism or metaphysical pluralism. Both are unacceptable positions to the vast majority of contemporary philosophers.

The arguments in Chapters III.1 and III.2 presuppose that we pre-theoretically understand the nature of the intentional framework and then show what one must accept to be able to hold to the framework. To motivate the choice between metaphysical pluralism and non-causalism, Chapter III.3 adopts a different approach and scrutinises the very nature of the intentional framework itself (III.3.1). This scrutiny leads me to address the very nature of meaning and, in particular, its relation to rules (III.3.2). Here my argumentation draws heavily from the extensive literature on the so-called problem of rule-following. I argue that the Standard View of seeing psychological states as intrinsically meaningful (including Platonism that treats meanings as abstract objects in need of contentful interpretation) is unacceptable, as is the causal dispositionalism that equates meaning with our causal dispositions. These views cannot deliver a coherent defence of the intentional framework. III.3.3 makes the problem even worse by showing that straightforward communitarian solutions cannot help us to understand meaning and intentionality either. Thus, Chapter III.3 concludes that the quietist position, according to which we simply cannot analyse the foundations of the intentional framework, is a viable option. Again, this is a strong philosophical result – although obviously highly unattractive for my task of defending naturalistic methodological holism and collective we-intentionality.

However, in this dark hour Chapter III.4 comes to the rescue by introducing the so-called social solution to the problem of rule-following as a way to defend the intentional framework in a thoroughly naturalistic way (III.4.1). Section III.4.2 argues that an essentially social version of the dispositional view manages to avoid the problems that were (III.3) fatal for psychologism, Platonism and dispositionalism, resulting in the Wittgensteinian view of I.1.5. In particular, Section III.4.3 argues that a pre-intentional version of the analysis of we-mode action given in I.2.4 can give us the kind of picture of pre-intentional social practices that is, according to the Wittgensteinian view on rule-following and the social solution, required for an acceptable analysis of the intentional framework. Moreover, these practices correspond to the kind of social behaviour the

evolutionarily plausibility of which was defended in II.1. Although this improved version of the social solution to the rule-following problem developed in III.4.2 and III.4.3 is of course an important philosophical result in itself, from the perspective of my main argument the value of Chapter III.4 is seen only in Chapter III.5, which explicates the role and implications of the argument of III.4 for my main thesis by applying the improved social solution to the major questions of the dissertation (III.5.2 in particular).

First, Section III.5.1 applies the improved social solution to the question concerning the nature and explanation of intentional action. It is argued that the rule-following considerations imply that of our two possible ways to defend the intentional framework, non-causalism and metaphysical pluralism (III.2.5), it is non-causalism that grounds the intentional framework (this, of course, is not to say that metaphysical pluralism is wrong but only that it is not required for the intentional framework). The improved social solution gives the philosophical justification for the view of intentional agency and its explanation of the Logical Connection Argument of III.1.

Section III.5.2, then, applies the improved social solution of III.4 and the view of intentional action and its explanation of III.5.1 to the question concerning the status of collective we-mode intentionality as discussed in Parts I and II. The result is a social constructivist view of collective intentionality. Thus, the argument of Section III.5.2, which brings together the results of the whole preceding dissertation, is perhaps the most important philosophical result of my study: III.5.2 offers a constructivist alternative to Searle's account of a biologically primitive collective we-intentionality currently dominating the theory of collective intentionality. And whereas Searle's account is explicitly based mainly on Searle's intuition that the biological primitiveness of (collective) intentionality is obvious (*e.g.*, Searle 2007b), my constructivist alternative is based on series of detailed philosophical arguments and analyses. Thus, to defend the Searlean orthodoxy against my constructivism it is not enough to appeal to intuitions. Rather, one must challenge the philosophical arguments this dissertation consists in.

Finally, Section III.5.3 applies the view of the intentional framework developed in the earlier sections to some issues that were left open in Parts I and II when I wanted to keep my argumentation neutral regarding the question of Wittgensteinianism. Most of these issues have notable importance to other fields of philosophy (*e.g.*, my arguments against evolutionary psychology and the co-operative virtue theory of cooperation, or my discussion concerning Ian Hacking's doctrine of conceptual practices making up people), but since they do not directly contribute to my main argument but

rather cash out certain implications of it, I do not explain them further here. Chapter III.6 closes Part III by summing up the results.

Although Part I, II and III jointly form my main argument for naturalised methodological holism *vis-à-vis* social ontology and agency and thus this dissertation, I have decided to add a rather longish Appendix to my main argument. The appendix has two goals. First, following in the footsteps of Searle (1995), I too worry that my constructivism may be mistaken as implying an unacceptably strong form of relativism and hence, like Searle, I want to add a further discussion on the connection between realism and my form of constructivism – especially since, unlike Searle, my dissertation subscribes to social constructivism also concerning agency and meaning and defends a non-causal theory of action. Appendix argues that there is no need to worry: realism worthy of the name (and our support) actually presupposes the kind of constructivism I defend.

Second, since in developing my main argument I have been forced to borrow rather extensively from other heated philosophical debates without addressing all the nuances that feature in the literature concerning those debates, I want – if only as an example – connect my main argument explicitly to at least one big philosophical debate not directly relevant for my argument to show how the view defended in the thesis travels to other fields of philosophy. The issue of realism seems to be well-suited for this task, and my discussion in Appendix serves to indicate what forms further research that accepts the main argument of my dissertation could take in other philosophical debates.

However, I also believe that the specific arguments given in Appendix are philosophically significant in their own right. A.1.1 connects the discussion of Chapters III.3, III.4 and III.5 to a defence of non-referential, direct realism. A.1.2 takes up a disagreement between Martin Kusch and Robert Brandom and provides a philosophical house cleaning this dispute in my view clearly needs. A.2, finally, revisits Quine's and Davidson's attempts to bridge the conceptual gap between their holistic conception of all things meaningful and the non-holistic material world that threatens to turn into an unbridgeable epistemic gap lethal for the notion of thinking and talking about the world. Particular attention is given to the way John McDowell seeks to resolve the Quine-Davidson problem. My main argument here is, as can be anticipated, that the problem McDowell identifies cannot be resolved in a satisfactory way unless the kind of naturalistic holism I have constructed in Parts I-III is accepted. Thus, in my view the arguments in Appendix give further evidence for my main argument, *i.e.*, for the importance of naturalised methodological holism *vis-à-vis* social ontology and agency.

**PART I:**

**THE BUILDING BLOCKS OF SOCIAL REALITY**



## INTRODUCTION TO PART I

We all live in societies. Yet this platitude does not imply that we are thoroughly familiar with the nature of social entities and social facts. Their philosophical nature is hardly an issue we pause to contemplate. However, since we are clearly able to act and live in social reality, we must understand at least approximately how social reality works. It seems to me, though, that the knowledge we possess as participants in social life is mainly practical “*know how*” type of knowledge that consists of a largely tacit, inexplicable capability to *act* more or less successfully in a social setting. Especially, our practical social knowledge typically does not include explicit understanding of the nature of social reality.<sup>2</sup>

An unfortunate result of our primarily implicit familiarity with the social realm is that also our scientific and philosophical understanding of social reality is all too often fuzzy and confused. The views of sociologists, economists, anthropologists and philosophers range from direct denials of the importance of the social (extreme individualism) to views that it is the social that determines the essence of individuals (full-blown social holism). Consensus does not appear to be forthcoming.

Hence, insofar as we are not completely satisfied with mere practical knowledge, there is true need for theoretical – indeed, philosophical – social theory that seeks to *explicate* the crucial features of our tacit understanding of the social. Social ontology as I see it is about making explicit issues that we already master in practice. This, as we shall see, does not mean that the results of such explications will not be surprising. Sometimes explications of implicit mastery will require us to correct some of our more uncritical views concerning the social world.

However, since the starting point of social ontology is our tacit knowledge regarding social reality, our theory of social ontology must also acknowledge what we know for certain about the social world and how we experience it. In particular, in our everyday life we presuppose a great number of essentially *social facts*, such as laws, norms, the existence of money and universities and so on. Philosophical social ontology must take this into account.

Thus, instead of presenting a philosophical theory of the ontology of the social world right at the outset I begin my journey by looking at what, exactly, we require – and, as importantly, what we do not require – from a theory of social reality and social facts. In particular, I wish to examine in what sense the most ordinary social facts can be

---

<sup>2</sup> Part III argues that this is what our knowledge of fundamental social practices ultimately *must* be like.

said to be *objective*. Moreover, I will initially rely on a common-sense understanding of what social facts are like. The idea is that the discussion of the objectivity of social facts will guide us to refine the common-sense notion and set the stage for further discussions, for it pinpoints the issues we need to understand in the ontological structure of social facts and thereby builds a bridge from commonsense understanding to high theory.

## CHAPTER I.1:

THE OBJECTIVITY OF SOCIAL FACTS<sup>3</sup>

Are social facts objective? To get a clear grasp of the problem, let us look at a simple example that John R. Searle (1995), whose work I will largely concentrate on, treats as a paradigmatic case of social facts:<sup>4</sup> the pieces of paper in my pocket are money. The monetary status of these pieces of paper is quite objective: which pieces of paper in my possession *are* money and which pieces are not is, unfortunately, beyond my control.

However, just as obviously the pieces of paper that are money do not have their monetary status directly in virtue of their physical properties. It is not the physical structure of the pieces of paper *qua* physical entities that *causes* them to function as media of exchange. Rather, as for example Searle (1995) and Tuomela (1995, 2002) emphasise, to be money is a social *status* based on the psychological facts that people in general *think* of certain pieces of paper as money and *accept* them as media of exchange, as money. In other words, the pieces of paper function as media of exchange in virtue of the social status assigned to the pieces, and the assignment and constitution of this status is due to collective acceptance (see Searle 1995, Tuomela 1995, 2002). Remove humans and their practices, and social facts and social reality disappear as well. In such a situation the pieces of paper would not be money anymore, even though they would still have all their physical properties. This seems to suggest that the fact that the pieces of paper in my pocket are money is not an objective fact after all.

The following definition captures this latter notion of objectivity:

(1) A fact is objective iff it obtains independently of us.<sup>5</sup>

The fact that the pieces of paper in my pocket are money cannot be objective in the sense of Definition (1), since the monetary status of the pieces obviously depends on human practices and, hence, the obtainment of the fact depends on us. But Searle's example of money also shows that any notion of objectivity which implies that social facts

---

<sup>3</sup> This Chapter is largely based on Saaristo (2003).

<sup>4</sup> Within the limits of this dissertation I of course cannot discuss the distinguishing features of all the different social facts. Instead my goal is to explicate the core elements of the objectivity of paradigmatic social facts. To use Hacking's (1997) terminology, I am in the business of explicating the *constitutive logic* of social facts, not of classifying different possible empirical scenarios.

<sup>5</sup> This definition is analogous to some standard textbook definitions of metaphysical scientific realism. Hence, even though I have chosen to conduct my discussion in terms of objectivity, my arguments are easily translatable into a discussion on the prospects of scientific realism in the social sciences (cf. Mäki 1996).

are not objective, cannot be fully satisfactory. An acceptable theory of the objectivity of social facts ought to explain also the intuition that there is something perfectly objective in social facts.

Searle thinks that this ambiguity concerning the objectivity of social facts captures the core puzzle in understanding social reality. In short, the fundamental problem is to explain how it is possible that “there are portions of the real world, objective facts in the world, that are only facts by human agreement” (Searle 1995, 1).<sup>6</sup> This means that in order to understand properly the objectivity of social facts we cannot be satisfied with Definition (1). Let us see how Searle seeks to resolve the ambiguity in the objectivity of social facts.

### 1.1.1 SEARLE ON THE OBJECTIVITY OF SOCIAL FACTS

The starting point of Searle’s (1995, 7 ff.) discussion concerning the objectivity of social facts is the idea that a universal notion of objectivity, such as the one expressed by Definition (1), cannot capture the different senses of objectivity relevant for a theory of social facts. According to Searle, two such senses are crucial here. In order to capture both of the mentioned intuitions we must distinguish between *ontological* and *epistemic* objectivity.

Ontological objectivity and subjectivity, Searle tells us, are “predicates of entities and types of entities” (Searle 1995, 8). An ontologically subjective entity is dependent on a perceiver or a mental state. Searle’s example is pain. Brute natural objects, such as mountains, are ontologically objective entities. I will conduct my discussion mainly in terms of facts rather than entities, and by ontologically objective facts I mean facts whose obtainment is independent of human activity.

Epistemic objectivity and subjectivity, in turn, are properties of statements, not of entities. An epistemically subjective statement is such that its “truth or falsity cannot be settled ‘objectively,’ because the truth or falsity is not a simple matter of fact but depends on certain attitudes, feelings, and point of view of the makers and hearers of the judgment” (Searle 1995, 8). Typical examples of such subjective statements are statements that are based on the taste and attitudes of the maker of the statement. Searle’s example is the statement that Rembrandt is a better artist than Rubens. Epistemically

---

<sup>6</sup> It is not very clear what kind of agreement Searle means here. It seems to me that often explicit agreement is not needed; implicit, tacit agreement will be perfectly sufficient in many cases (Part III of this study argues that the very possibility of explicit agreement actually *presupposes* quite a lot of implicit agreement).

objective statements are such that their truth-values are not dependent on attitudes or tastes. Searle's example is the statement that Rembrandt lived in Amsterdam during the year 1632. When in what follows I occasionally talk about epistemically objective (or subjective) facts, I use the expression as an abbreviation of this notion of the nature of the sentences describing the facts.

Searle's conceptual apparatus appears to meet the abovementioned requirements very well. The ontological notion of objectivity captures the common-sense intuition behind Definition (1), allowing us to conclude that social facts are not ontologically objective. The problem with Definition (1), however, was that it could not account for the other intuition we have: social facts appear to be objective in the sense that their obtainment is immune to our personal attitudes, desires, feelings etc. This, of course, is not a problem for Searle since he can say that although social facts are ontologically subjective, the facts (or, rather, the statements describing them) are nonetheless epistemically objective.

Thus Searle's solution manages to avoid the obvious problem with Definition (1), and by doing so it also succeeds in doing justice to the core puzzle of social reality. Notwithstanding these indisputable successes, I nonetheless think that we can give a better account of the objectivity of social facts than the one given by Searle. Searle's account, although not mistaken, is nonetheless somewhat misleading or at least *incomplete* since it fails to capture some crucial features of social facts, and thus we should not be content with it. The next section explains what limitations I see in Searle's account.

### I.1.2 PROBLEMS WITH SEARLE'S ACCOUNT

There are at least three major (interconnected) shortcomings in Searle's account. (i) It is not clear how the two distinctions are related to one another (and indeed to the structure of social facts), and this poses problems for attempts to construct a Searlean taxonomy of facts. (ii) The account fails to highlight the distinctive character of social facts. (iii) The account fails to increase our understanding as to *why* social facts are ontologically subjective but epistemically objective. The last two problems are the most important problems here, but let me start with the first since it sheds light on the more serious problems.

Searle appears to think that the two distinctions – ontological and epistemic objectivity/subjectivity – enable us to construct a taxonomy of facts. In other words, Searle

seems to think that we can classify facts to those that are (a) both ontologically and epistemically objective, (b) ontologically and epistemically subjective, (c) ontologically objective but epistemically subjective and (d) ontologically subjective but epistemically objective. It is not clear that this way of classifying facts will result in anything interesting, largely because the epistemic distinction concerns, as Searle (1995, 8) emphasises, predicates of judgements, whereas the ontological distinction deals with predicates of entities. It is not obvious that the marriage of these two distinctions, operating with completely different categories, will reveal anything valuable about different kinds of facts in general and about social facts in particular. After all, neither of the distinctions is defined in terms of facts.

And indeed the role of some of the Searlean classifications remains somewhat perplexing. The class (a) is clear enough, but what should we think about, say, the class (c). Again, Searle's intuition is respectable enough: "For example, the statement 'Mt. Everest is more beautiful than Mt. Whitney' is about ontologically objective entities, but makes a subjective judgment about them" (Searle 1995, 8-9). The problem is, though, that it is not clear whether this teaches us anything new about facts. The fact involved seems to be simply that the maker of the statement prefers Mt. Everest to Mt. Whitney in aesthetic terms. Hence the entities most intimately associated with this example – preferences, mental states – are by Searle's own definition ontologically subjective entities, even though this was supposed to be an example of ontologically objective entities (what mountains of course are). I do not think Searle says anything obviously false here, but I also fail to see why this kind of classification should be of great importance when addressing the objectivity of different kinds of facts.

Similarly for the class (b): Supposedly we should find here value judgements about entities whose mode of existence depends on mental states. An example might be the following: I dislike the pain I am now feeling. Again, it is not clear whether we learn anything new about facts by learning to classify statements like this as belonging to the class (b). The facts involved seem to be simply the two ontologically subjective facts that I am in pain and that I find it unpleasant. Searle's point seems to be to emphasise that statements about pain are epistemically objective in the sense that their truth-values are independent of my attitudes, whereas my dislike is an attitude and hence the same cannot be said about it. However, the truth-value does depend on my feelings (*i.e.*, whether I do feel pain or not) and hence Searle's (1995, 8) explicit definition seems to render it epistemically subjective. Similarly, when we move beyond first appearances the epistemic status of the judgements about my preferences becomes problematic as

well. Preferences are by Searle's definition ontologically subjective, but in what sense can we say that the judgements about them are epistemically subjective? Surely the truth-values are dependent on my tastes etc., but nonetheless there seems to be something objective in them, since many of my preferences are just given to me in introspection or action, and I cannot change them at will. It seems to me that learning to place facts such as the ones discussed here into the class (b) raises more questions than it answers.

Finally, even if we sidestep these problems with epistemic objectivity and subjectivity, there are still other confusing issues left. Searle defines the epistemic objectivity and subjectivity of a given fact in terms of whether the truth conditions of a sentence describing the fact depend on the person uttering the statement or not. Hence the same statement can be epistemically subjective or objective depending on who actually utters it. For example, the statement "John Searle likes beer" is epistemically subjective when uttered by Searle, but epistemically objective when uttered by someone else. As such this feature of Searle's classificatory system can be seen as an advantage, since it allows for a clear distinction between the first and third-person approaches to the mental facts cognitive psychology deals with (see Dennett 1991a for the importance of this feature and Sections I.1.2 and I.1.3 below for what this implies for the present discussion).

However, the price Searle's theory pays for this success is that by shifting the focus from the facts themselves onto their descriptions, the ultimate structure of social facts remains unaccounted for. For example, a peculiar implication of Searle's characterisations is that the statements "I am in pain" and "The pieces of paper in my pocket are money" have exactly the same status. These statements are about entities whose "mode of existence" is ontologically subjective, since the existence of the entities depends on us. However, the statements reporting the facts are epistemically objective, since each statement is "made true by the existence of an actual fact [albeit ontologically subjective] that is not dependent on any stance, attitudes, or opinions of observers" (Searle 1995, 9).<sup>7</sup>

---

<sup>7</sup> Searle's claim that the epistemic objectivity of social facts is based on their independence of human attitudes seems to be in tension with what Searle says about the ontology of social facts. For Searle, social facts "are only facts by human agreement" (Searle 1995, 1). Now surely human agreement is tied to human attitudes. However, Searle can hold that a social fact is epistemically objective due to its independence of the attitudes of the person observing the fact here and now, even if the fact is not independent of attitudes in general. In the following sections this distinction is explicated and its importance is made clear. Searle's failure to be clear on these issues forms one more reason why we cannot be completely satisfied with his account.

The other example is problematic too. No doubt pain is independent of attitudes, but it seems to be dependent on feelings, which was earlier (Searle 1995, 8) said to imply epistemic subjectivity.

Philosophers who accept the impossibility of private languages and advocate a social theory of agency, meaning and mental content (see, for example, Williams 1999 and Part III of this study) might want to argue that actually the implication of Searle's classifications that mental and social facts have the same epistemic status and refer to entities of the same ontological status should be seen as an advantage of Searle's account. However, this would be inappropriate for two reasons. First, Searle wants to see an essential difference between mental and social facts, and hence his account's failure to reflect what he thinks this difference is must be seen as a shortcoming of his account. Second, the message of the communalists is that the structure of certain facts is, contrary to pre-theoretic intuitions, irreducibly social rather than individualistic. Hence a satisfactory classificatory system of different kinds of facts should reveal what it is in the structure of social facts that sets them apart from individualistic facts, and then the communalist could explain why she thinks we should see the facts under scrutiny as social facts rather than as individualistic mental facts. Searle's framework is conceptually too deprived to offer such clarification.

To conclude, Searle's analysis of the objectivity of different types of facts fails to explain what features characteristic of social facts actually make the sentences describing the facts epistemically objective despite the ontological subjectivity of the facts in question. This shortcoming makes, in my view, Searle's account unsuitable as an analytic tool for shedding light on many disputes in the philosophy of the human sciences, such as the quarrel over the communalist and individualist theories of meaning and mental content.

What is needed is an account of the objectivity of social facts that can explain *why* the judgements about both individualistic mental facts and social facts are epistemically objective despite the ontological subjectivity of the facts, while also explicating precisely what differentiates social facts from individualistic mental facts (if they indeed are different). This is something Searle's framework cannot offer.

### 1.1.3 TOWARDS A DURKHEIMIAN ACCOUNT OF SOCIAL FACTS

Searle seeks to capture the different senses of objectivity by pointing out that we can understand "independence" in Definition (1) either as ontological independence or as epistemic independence. Thus, the Searlean picture replaces Definition (1) with the following definitions.



(2) A fact is ontologically objective iff its obtainment is independent of us (if it could obtain independently of humans).

(3) A statement is epistemically objective iff the truth-value of the statement is independent of the attitudes etc. of the maker of the statement (if a representation of the fact is made true or false by something independent of the attitudes of the person constructing the representation, be the fact represented ontologically objective or subjective).

The fundamental reason why Searle's account led to the problems discussed above is that (2) does not actually add anything to (1), and (3) does not talk primarily about *facts* anymore, but about sentences. Hence, we need new candidates for replacing (1), (2) and (3) that are about features of facts. At this point it is time to turn to the *locus classicus* of the philosophy of social facts, Emile Durkheim's *The Rules of Sociological Method* (Durkheim 1895/1982).

Durkheim seeks to capture the objectivity of social facts with his notorious maxim that we must treat *social facts as things*. Naturally, this principle leads us to ask what does it mean to say that social facts are to be treated as things. Notice that the claim is *not* "that social facts are material things, but that they are things just as are material things" (Durkheim 1895/1982, 35). For Durkheim, things include "all that which the mind cannot understand without going outside itself" (Durkheim 1895/1982, 36). I read this as saying that a description or representation of a thing must satisfy Definition (3).

Hence, in Durkheim's picture the defining feature of a thing is that the descriptions of it are *epistemically* objective in Searle's sense. Consequently, where Searle's notion of epistemic objectivity captures a feature of neither entities nor facts, but of sentences and representations, also for Durkheim "[t]o treat facts of a certain order as things is therefore not to place them in this or that category of reality; it is to observe towards them a certain attitude of mind" (Durkheim 1895/1982, 36). This is good to keep in mind, for all too often Durkheim's account is dismissed on ontological grounds.

However, precisely because the maxim to treat social facts as things does not go further than the Searlean definition (3) and hence points out only what the representations of social facts have *in common* with the representations of many other facts, being thus completely silent about the *distinctive* features of social facts, Durkheim cannot leave his discussion on the objectivity of social facts here. But it is not clear how Durkheim could proceed from here coherently. Durkheim's dilemma is that on the one hand

he thinks that in the name of ontological naturalism we must accept that “society comprises only individuals” (Durkheim 1895/1982, 39), but on the other he is also committed to the view that the study of social facts is an autonomous discipline, the objects of which are not studied by non-social human sciences or natural sciences.

For Durkheim, the characterising feature of social facts is that regardless of which individual perspective they are observed from, social facts are *always* treated or observed as things, whereas, *e.g.*, individualistic psychological facts are treated as things only when observed from the third-person perspective. Thus, in Durkheim’s view the crucial question for the possibility of *sui generis* social science is this: Is there something in the nature of social facts that makes them things, *i.e.*, epistemically objective, in this full-blown sense and thus to differ qualitatively from their constitutive parts (such as individualistic psychological facts), even when in the strict ontological sense society consists of nothing but individuals? Durkheim thinks that if he can give an affirmative answer to this question, he has provided an account of *methodological holism* (for social facts are then qualitatively different from their constitutive parts) which is nonetheless *ontologically naturalistic* (since society comprises only individuals). A defence of this kind of naturalistic holism is also my goal in this dissertation.

As we know, Durkheim’s answer is that the epistemically objective status of social facts is due to the fundamental structure of social facts: “What constitutes social facts are the beliefs, tendencies and practices of the group [of individuals] *taken collectively*” (Durkheim 1895/1982, 54, my italics). This suggests that in order to do justice to Searle’s line of reasoning about ontological and epistemic objectivity so that we also manage to capture the distinctive features of social facts, we should concentrate on the notion of “us” in Definition (1), rather than on the notion of “independence” as Searle does. We can understand “us” here in two different senses. First, “us” might be taken to mean something like “humans in general” or, in some cases, “the members of the social group in question”.<sup>8</sup> However, “us” may also be taken to mean “any particular individual”. Consequently, the Durkheimian replacement of Definition (1) would explicate two different senses of objectivity with the following pair of definitions.

---

<sup>8</sup> For the present purpose of explicating the structure of social facts the difference between “humans in general” and “the members of the social group in question” is not crucial. What matters is that both formulations bring in *intersubjectivity*. However, for some other fundamental problems of the philosophy of social science the difference can be central (*e.g.*, when tackling the challenge of cultural relativism) and should accordingly be explicated in the definitions of objectivity. I thank Steven Lukes for highlighting the importance of this point to me.

(4) A fact is objective<sub>1</sub> iff it obtains independently of human attitudes and actions.

(5) A fact is objective<sub>2</sub> iff it obtains independently of the attitudes and actions of any particular individual.

Now Definition (4) captures the common-sense intuition of objectivity as expressed by Definitions (1) and (2), and hence also Searle's notion of ontological objectivity. Obviously social facts are not objective<sub>1</sub> in the sense of Definition (4). However, what we required was a notion of objectivity that captures also the intuition about the epistemic objectivity of social facts, as highlighted by Searle's example of money and as aspired after by Definition (3). Objectivity<sub>2</sub> in the sense of Definition (5) can offer this and, moreover, it points out *why* social facts are epistemically objective in the sense of Definition (3), and does it in a way that sets social facts clearly apart from facts about individuals that are objective neither in the sense of Definition (4) nor in the sense of (5) – a feature emphasised by both Searle and Durkheim. Furthermore, Definition (5) can provide all this without confusing social facts with brute natural facts that are objective in the sense of Definition (4) (and, hence, also in the sense of (1) and (2)), which, obviously, is another requirement an acceptable view must meet.

It could be argued that it is to some extent misleading to call this kind of account of social facts *Durkheimian*, since Durkheim's view, as Steven Lukes (1982, 7-8) emphasises, characteristically includes metaphysically demanding notions such as independent macro-level causal forces, and I do not wish to commit myself one way or another regarding the existence of such forces.<sup>9</sup> However, I believe that the importance I lay upon Definition (5) justifies the label "Durkheimian". In the following section I will nonetheless leave Durkheim in the background and explicate the distinctive nature of social facts in terms of Uskali Mäki's (1996) discussion concerning the modes of existence of economic entities.

---

<sup>9</sup> One might wonder why I do not mention Durkheim's notorious notion of the collective mind as the example of Durkheim's ontologically dubious notions. The reason is that arguably that notion is of great importance, and the naturalisation of it is indeed the key for understanding social ontology (my use of the theory of collective we-mode intentionality in this dissertation is meant to be precisely that kind of naturalisation). Of course also Durkheim's macro-causation may turn out to be indispensable – see, in particular, III.2.5.

#### I.1.4 EXPLICATING THE OBJECTIVITY OF SOCIAL FACTS

This section presents a classification of different kinds of facts basing on the internal structure of the facts themselves. I have largely adopted the terminology of Mäki (1996), whose line of thought my presentation for the most part agrees with. However, the aim of Mäki's discussion is to examine the prospects of scientific realism in economics, and he states his definitions in terms of the modes of existence of physical, economic and psychological objects. The present discussion differs from Mäki's in two ways. First, I generalise the account to apply to social science and social reality in general and, second, I continue to talk about facts rather than objects. Hence, in what follows the definition of independent facts is based on Mäki's (1996, 432) definition of independent existence, the definition of external facts on Mäki's (1996, 432) external existence and the definition of objective facts on Mäki's (1996, 433) objective existence. I should also add that Mäki is not committed to the idea that the account would be in any sense Durkheimian in nature.

Let us call facts that are objective in the strong sense of Definitions (1), (2) and (4) *independent facts*.

A fact is an *independent fact* (it obtains independently) iff it obtains independently of the human mind.

It should not confuse us that "independence" appears both in the *definiens* and in the *definiendum*. After all, this is not a semantic analysis of a concept, but rather an explication of *independently of what* a fact must obtain if the fact is said to obtain independently (this remark applies *mutatis mutandis* also to the definitions of external and objective facts below).

Brute natural facts are independent facts,<sup>10</sup> whereas social (and psychological) facts are not. We can define the full-blown *ontological* version of methodological individualism in the philosophy of social science in terms of independent facts as follows.

(OI) Social facts do not obtain independently.

---

<sup>10</sup> Actually, Appendix argues that the whole category of independent facts is under a serious threat. However, in this Part I will go along with the common-sense understanding of independent natural facts.

On the basis of Section I.1.3 above it is easy to see that Durkheim, who is generally seen as the ultimate sociological holist, nonetheless accepts (*OI*). Of course the acceptance of (*OI*) leaves room for many anti-reductivist forms of methodological holism; since (*OI*) simply says that social facts are neither based on any novel Cartesian substance nor are they parts of the natural world that exist independently of human activity.

Let us call the category of facts that are epistemically objective in the sense of satisfying Definition (3), but nonetheless intersubjective<sup>11</sup> in the sense of Definition (5), *i.e.*, excluding individualistic facts, the category of *external facts*.

A fact is an *external fact* (it obtains externally) iff it obtains independently of (and external to) any individual human mind.

As a stronger notion independence clearly implies externality, and thus independent brute natural facts are also external facts. However, also social facts in Durkheim's sense, *i.e.*, facts that depend on human attitudes *taken collectively*, can be external facts. It is the externality of social facts that explains their ontological subjectivity and epistemological objectivity. The pieces of paper in my pocket are money because they are collectively accepted as money. Hence their status as money is not an independent fact. Despite its non-independence, however, the monetary status of the pieces is quite immune to changes in my attitudes; the fact that the pieces are money is a fact external to and independent of me (or of any particular individual for that matter).

At this point we can see why it is more intuitive to speak about obtaining facts than about existing objects. Although in the light of what has been said we can understand the complex ontological status of the statement that the British Society for the Philosophy of Science is an externally existing social entity, we nevertheless tend to associate the terms "entity" and "object" so strongly with independently existing material things we face in our everyday lives that the talk about social objects may make us feel a bit uneasy. But if we instead state that it is an external fact that the President of the British Society for the Philosophy of Science is Professor Steven French, it is easy to understand that this fact obtains only due to several agreements and conventions.

However, this concentration on facts rather than objects does not allow us to avoid considerations of the status of social entities, since, as Ruben (1985, 34) reminds us, if a singular *de re* sentence "*x* is *P*" is true, it follows that *x* exists. Hence, if we be-

---

<sup>11</sup> This of course suggests that Searle's solution to call social facts ontologically *subjective* can be misleading in some contexts, since the essence of social facts seems to rest precisely in their *intersubjectivity* as opposed to *subjectivity*.

lieve that some fact about a social institution obtains externally, we are also committed to external existence of the institution in question – but as emphasised, this does not tell against ontological individualism in the sense of (*OI*). We can still argue, as I – following Searle – do below, that the external existence of social institutions comes down to collectively accepted patterns of behaviour and collectively held attitudes, beliefs *etc.*

It is possible, however, that when analysing social facts we sometimes must take into account also facts that are neither independent nor external. For example, there might be a society in which only those pieces of paper are money that are pronounced to be money by one particular individual (a bit as English bank notes that include the signature of the chief cashier of the Bank of England). The acceptance of a single individual is of course dependent on that individual, and hence it is not an external fact.

Nonetheless, the involvement of individual facts like this does not necessarily destroy the external objectivity of social facts, since also in this case the fundamental fact that *renders* the non-external fact socially significant is itself an external fact. The non-external fact is socially important only insofar as people collectively accept its role (perhaps unwillingly or tacitly by just going along with a perceived custom). It is the external fact of collective acceptance that makes the non-external fact socially significant. Even in a society where only those pieces of paper are money that the king declares to be money, the king's declaration can assign the status of money to the pieces of paper only insofar as the members of the society accept (perhaps tacitly) this procedure as the correct way to issue new tokens of money, *i.e.*, grant the king a certain status first. Social statuses always require collective acceptance.

However, even if we grant that in order to have social significance a non-external fact presupposes an underlying external fact, it remains true that there are situations in which social explanations must make use of non-external facts such as the acceptance of a particular individual. Hence, to complete<sup>12</sup> the framework in which the objectivity of different kinds of facts can be discussed, let us also examine in what sense non-external facts can be objective. The key insight here, I think, is Dennett's (*e.g.*, 1991a, 71) insistence that the perspective of scientific objectivity – and in Dennett's view the perspective of all science – is the third-person point of view. Dennett thinks that if the human sciences are to be mature sciences they must be able to represent individual facts from the objective third-person point of view.

---

<sup>12</sup> Actually, this completes only the account inspired by Mäki's argumentation. Below I give one more definition, that of social facts proper.

Let us follow Dennett and call facts that are capable for being so represented *objective facts*.

A fact is an *objective fact* (it obtains objectively) relative to a given representation iff it obtains unconstituted by that particular representation.

Once again, Durkheim is in full agreement:

The facts of individual psychology themselves [...] must be considered in this light [*i.e.*, as objective facts]. Indeed, although by definition they are internal to ourselves, the consciousness that we have them reveals to us neither their inmost character nor their origin.<sup>13</sup> [...] This is precisely why during this century an objective psychology has been founded whose fundamental rule is to study mental facts from the outside, namely as things.  
(Durkheim 1895/1982, 36-37.)

In other words, whereas material and social facts are essentially objective in the above sense, facts of individual psychology are objective only insofar as we represent them from the third-person perspective (to use Dennett's terminology), *i.e.*, only insofar as we treat them as things (in Durkheim's terminology).

As the final point of this section I should add that strictly speaking social facts as external facts are not completely independent of any particular individual. External facts depend on individuals qua group members, but the impact of each individual alone is quite negligible – if the impact were not negligible, then the fact in question would not be an external fact, but, by definition, (presentable as) a mere objective fact (cf. Barnes 2002, 251).

### 1.1.5 ARE SOCIAL FACTS *SUI GENERIS*?

The last remaining clarification that needs to be done before the framework for understanding the nature of social facts can be regarded as complete is to explicate the different ways of understanding the status of external social facts or, more precisely, in what sense we collectively construct and reproduce them. Until such an account is provided we have not moved beyond traditional *emergentism*, which simply holds that social facts are emergent features of groups of individuals, and thus does not say anything helpful about the precise structure of social facts. I call the different accounts to be dis-

---

<sup>13</sup> Indeed, the theory of the "character and origin" of contentful mental states I defend in Part III shows that Durkheim is correct: I suspect some people might find the view I defend as challenging their pre-theoretical views on this matter.

cussed below the Individualistic Account, the Wittgensteinian Account and the Durkheimian Account. In this Section I do not defend any of the three views. Rather, at this point I merely want to make the competing views explicit, and an informed choice between them is the topic of the rest of this dissertation.

### The Individualistic Account

The first candidate for a theory of the nature of social facts as external facts is to say that they do not form a *sui generis* category of facts in any substantial sense. Rather, social facts as external facts are essentially aggregates of individual psychological facts. In the present context paradigmatic representatives of this account<sup>14</sup> are, for example, D. H. Mellor (1982) and Peter Abell's (2000) rational choice approach to social theory. Abell, for example, defines this account so that it "invites us to understand individual actors [...] as acting, or more likely interacting, in a manner such that they can be deemed to be doing the best they can for themselves, given their objectives, resources, and circumstances, as they see them [*i.e.*, whether "their best" is based on egoistic, altruistic, or group-directed motivations, cf. II.2]" (Abell 2000, 223) and social facts as nothing but aggregates of such individual actions. External social facts, then, are significant only in the sense that when aiming to maximise one's utility function part of the circumstances a rational agent must take into account is the behaviour of others. Sociality reduces to strategic interaction.

In short, an advocate of the Individualistic Account of external social facts would hold that there are only physical (independent) facts and individualistic psychological (objective) facts. In addition to these, we may for pragmatic reasons wish to talk about aggregates of objective psychological facts as external social facts. However, just like a collection of physical facts does not form a new basic category in comparison to individual physical facts, a social fact as an aggregate of facts about individuals does not belong to any category substantially different from psychological facts. By the same token, though, this view implies eliminativism about social facts no more than modern physics implies that macroscopic physical entities do not exist. According to this view, social facts as external facts are real and consist of systems of interlocking beliefs and intentions of fundamentally asocial individuals (cf. Bratman 1999 or Miller 2001). The animating idea behind the Individualistic Account is that the acceptance of (*OI*) implies

---

<sup>14</sup> See also the classical formulations of methodological individualism by Hayek, Popper, Watkins and others in O'Neill (1973).



that social facts must be instantiated by an individualistic mechanism, and today it is often either implicitly assumed (*e.g.*, Elster 1989) or even explicitly stated (*e.g.*, Cowen 1998) that the individualistic mechanisms are to be defined in terms of individual rational choice and equilibrium, and accordingly Part II argues mainly against this interpretation of the account.

In sum, the Individualistic Account is based on the anti-Durkheimian assumption that since society consists of interacting individuals, the fundamental level of social theory is that of individual agents who make choices from their individual perspectives. Hence this account is a methodologically individualistic account *par excellence*. Social facts as external facts are not *sui generis*.

### The Wittgensteinian<sup>15</sup> Account

Philosophers who think that participation in social practices is what makes the human form of life, including individual intentional agency, possible, take a substantial step away from the methodological individualism of the Individualistic Account.<sup>16</sup> For the Wittgensteinians social practices are conceptually prior to individual psychological facts, and hence external social facts are *sui generis*. Indeed, they think that social facts cannot be aggregates of psychological facts, because it is social facts in the sense of objective social practices that make meaningful individual thoughts and actions (including language), *i.e.*, psychological facts, possible in the first place.

The idea is that just as pieces of paper cannot be money merely in virtue of their independent physical properties, but require social practices within which the pieces are accepted as media of exchange (and thus as money), also a brain state or an expression cannot represent or refer to anything (*i.e.*, be intentional, about something) in virtue of its independent properties. For the Wittgensteinians, meaning in the sense of *conceptual content* presupposes that one participates in social practices that assign the statuses of meaningfulness and contentfulness to one's states and expressions.

In this manner, the Wittgensteinians think, thoughts and language really are meaningful (intentional, about something) only within social practices. This can be

---

<sup>15</sup> I have labelled this account "Wittgensteinian", since most adherents of this position locate its foundations in Wittgenstein's later philosophy (especially Wittgenstein 1953). As with Durkheim, I remain uncommitted regarding exegetical issues.

<sup>16</sup> Examples of Wittgensteinian philosophers in this sense – and whose discussions proceed in terms relevant to the present Chapter – include Barnes (1983, 1995, 2000), Bloor (1996, 1997), Brandom (1994, 2000), Esfeld (1999, 2001), Haugeland (1990), Kusch (1999), McDowell (1998a), Pettit (1993), Williams (1999) and Winch (1958).

seen as the very feature that distinguishes methodological holists from individualists (although below I argue that the Durkheimian Account introduces an alternative way of rejecting individualism):

For the individualist, concepts are individually held, mental entities. But for the non-individualist, these are irreducibly social. Even if it is accepted by the individualist that beliefs often involve social concepts, he cannot dispel the view that they exist only as internal states of individuals. The non-individualist contends that, whether or not they are attitudes of individuals, these have a constitutive relation with intrinsically social practices.  
(Bhargava 1992, 12.)

The arguments of the Wittgensteinians are examined in detail in Part III. Here it suffices to understand that although the Wittgensteinians accept the classification of facts as represented in I.1.4, they also emphasise that when the focus is on the human sciences the crucial category is that of external social facts, as opposed to the concentration on individualistic psychological facts, which is essential for individualistic accounts such as the rational choice approach. The conceptual order is not captured by the common-sense view that independent natural facts give rise to objective psychological facts, which in turn generate external social facts. Rather, independent natural facts have evolved into external, co-operative social practices (ultimately, language) that form the bedrock of the human form of life, and it is only in virtue of these practices that objective psychological facts are possible.

Thus, the Wittgensteinian Account turns the relation between the social and the individual upside down when compared to the Individualistic Account, which starts with asocial individuals whose actions constitute the social. According to the individualistic approach, individual actions are prior to social practices, for social practices are but aggregates of individual actions. The Wittgensteinian Account, in contrast, sees social practices as conceptually prior to individual actions, for it is the existence of social practices that constitutes the very possibility of individual actions. Thus, the Wittgensteinian position is a form of methodological holism. However, just like the Individualistic Account, also the Wittgensteinian Account subscribes to (*OI*).

### The Durkheimian Account

Another way of moving away from the methodological individualism of the individualistic approach without rejecting the ontological thesis (*OI*) is to concentrate on *anti-individualistic collective action*. The idea is that truly collective action, including collec-

tive acceptance that constitutes social facts, is based on a mode of psychology and agency different from individual agency. Durkheim held notoriously that social facts are constituted by the activities of a collective consciousness, and hence I have labelled the views subscribing to this understanding of external social facts the *Durkheimian Account*. This view is based on an explicit rejection of the *summative* view advocated by the Individualistic Account. As Durkheim puts it (to use emotions as an example of collective agency), “[a]n outburst of collective emotion in gathering does not merely express the *sum total* of what individual feelings share in common, but is *something of a very different order*” (Durkheim 1895/1982, 56; my italics).

This is of course the very aspect of Durkheim’s thought that has caused a lot of agitation among modern social theorists. Steven Lukes captures this problem well:

When writing of social facts as ‘external to individuals’ he [Durkheim] usually meant ‘external to any given individual’, but often suggested (especially to critical readers) that he meant ‘external to all individuals in a given society or group’: hence, the often repeated charge against him that he ‘hypostasised’ or reified society, a charge which is by no means unfounded.  
(Lukes 1982, 4.)

In my terminology, although Durkheim typically portrays the state of the collective consciousness (which constitutes social facts) as an external fact, he sometimes describes it as an independent fact. However, I have already argued that Durkheim accepts ontological individualism in the sense of (*OI*).

Hence the way to understand the strong externality Lukes talks about is the collective agency view sketched above, where the collective consciousness does not belong to a new holistic entity, but rather consists of a psychology *different in kind* from the psychology associated with individual action. Individuals can act qua individuals, or qua group-members. For Durkheim, these modes of agency are on a par; neither is more fundamental. Thus, *pace* the Individualistic Account, in Durkheim’s view social facts are not aggregates of individual-mode psychological facts. Of course, a modern theorist cannot be content with vague appeals to collective consciousnesses, and hence the challenge for modern Durkheimians is, as Lukes (1982) constantly emphasises, to provide a naturalistic explication of the micro-foundations of Durkheim’s concepts. In my view the theory of collective we-mode intentionality delivers precisely such a naturalisation.

The Durkheimian picture can be seen either as a further development of the Wittgensteinian view, or as an alternative to it. Philosophers who see it as a development of the Wittgensteinian position, include Barnes (2000) and, to some extent, Bloor (1996), Kusch (1999) and possibly Tuomela (especially 2002). The motivation for their

Durkheimian position is that they want to resist the anti-naturalist conclusion of some Wittgensteinians (notably McDowell (1998a), perhaps also Wittgenstein himself), according to which social practices that form the bedrock of the human form of life are so fundamental that we must accept them as primitive notions not to be analysed further. As the slogan goes, we should not even try to dig below the bedrock. The Durkheimian aspiration is that the theory of collective agency might offer the tool for such excavations. This research programme is analysed in Part III.

However, some Durkheimians do not accept the Wittgensteinian version of anti-individualism. For these philosophers intentionality, including collective psychology resulting in collective agency, is a biologically primitive feature of the human brain and not something that is constituted by social practices. Gilbert's (1989) theory of social facts and collective intentionality is explicitly both Durkheimian and anti-Wittgensteinian. Similarly, Searle holds that both the individual and collective mode of psychology and action are aspects of human biology (Bloor 1996 and Haugeland 1990 criticise perceptively this aspect of Searle's view in a context relevant for the present discussion). Tuomela (especially 2002) must be placed somewhere in between, since he argues that the Wittgensteinian view is a contingent truth about the actual world, but not a conceptually necessary condition for all possible meaningful action (including contentful thought), as the mentioned Wittgensteinians seem to think. According to Tuomela (and, e.g., Pettit 1993), in some possible worlds the view of Gilbert and Searle is true.

Despite the fundamental disagreement concerning the ultimate sources of intentionality (social or innate), in aspects relevant to the present task of explicating the status of social facts the non-Wittgensteinian version is remarkably similar to the Wittgensteinian version of the Durkheimian Account. The connecting idea is that the construction and maintenance of sufficiently stable social facts presuppose a stronger conception of sociality than the summative notion of the Individualistic Account of strategic interaction of asocial individuals (cf. Barnes 2001, 23). The Durkheimians think that social action worthy of the name must be based on essentially social psychology. Durkheim's claim that there are collective attitudes which are not sums of individual attitudes but "something of a very different order" can, the contemporary Durkheimians think, be acknowledged by admitting that there are two *sui generis* modes of human psychology and action. Sometimes we act essentially qua autonomous individual agents, *i.e.*, in the individual mode (or the I-mode), and sometimes we act essentially

qua group members, *i.e.*, in the we-mode (these notions are discussed at length in I.2 and Part II).

Thus, all the Durkheimians are committed to a fourth category of facts. This is the category of full-blown *social facts* and it can be defined as follows:

A fact is a *social fact* iff it is constituted by collective acceptance based on collective we-mode intentionality, psychology and action (or collective agency).

All social facts are external facts, but all external facts are not social facts (in particular, the external facts captured by the Individualistic Account cannot be social facts in this sense). Social facts are also external to any individual perspective, since individual-mode perspectives do not even participate in their construction and reproduction, for the facts are based on collective we-mode agency. Thus the Durkheimians avoid the problem mentioned in the end of Section I.1.4.

Where the Wittgensteinian and anti-Wittgensteinian versions of the Durkheimian Account differ from each other is in the way they understand the conceptual order of facts of different kind. The Wittgensteinians think that there are natural (i) *independent facts* that give rise to non-intentional collective agency. Collective agency, in turn, constitutes (ii) *social facts* (essentially, social practices). According to the Wittgensteinians, all this can be *explained* within the framework of evolutionary biology and other causal explanations of the natural sciences. The social practices, then, constitute the *normative* framework that makes psychological (iii) *objective facts* possible, opening thus the door for normative human sciences that aim to *understand* human action. The psychological facts in place, we can finally have (iv) *external facts* in the sense the Individualistic approach sees them, *i.e.*, as aggregates of objective, interlocking psychological facts.

The anti-Wittgensteinian view does not imply a similar distinction between the natural and the human sciences, or between causal explanation and normative understanding. However, the problem also the anti-Wittgensteinian Durkheimians see in most attempts for naturalistic explanations of the social world is the unquestioned commitment to methodological individualism in the sense of building all social notions on essentially asocial individual agency. The anti-Wittgensteinians hold that the ontologically fundamental level is that of natural (i) *independent facts*. But these natural facts are seen to have evolved to give rise to *two* modes of intentional agency that are irreducible to one another: collective we-mode agency and individual-mode agency. These

modes are both biologically primitive features of individual agents, and as such (ii) *objective facts*. Action based on we-mode intentionality constitutes truly (iiia) *social facts* whereas individual-mode actions constitute (iiib) *external facts* in the sense of the Individualistic Account.

### I.1.6 CONCLUSION

The distinctions made in this chapter help me to define more clearly what I am after in this dissertation. First, we need an account of how objective and external social entities and facts (non-eliminativism) are possible in our natural world (the acceptance of *OI*); or, in other words, how external and objective social reality is constructed out of naturalistic building blocks. In particular, we need to examine the fine structure of social and institutional facts as well as social practices and entities to find out whether the Individualistic, the Wittgensteinian or the Durkheimian Account offers the right way to understand the nature of external, social facts.

In the next Chapter I begin this task by examining the central arguments of the *locus classicus* of contemporary social ontology, namely Searle (1995). In short, Searle thinks that ontologically speaking social facts and entities boil down to collectively upheld patterns of behaviour. Thus, Searle's account remains faithful to the naturalistic principle (*OI*). However, Searle thinks that to secure epistemic objectivity we have to take very seriously the idea that social practices are maintained *collectively*. This amounts to favouring the Durkheimian reading of social facts. Chapter I.2, then, analyses the Searlean naturalisation of Durkheim's collective consciousness, namely the theory of collective we-intentionality.<sup>17</sup> I connect Searle's line of thought to the argumentation of the other two main theorists of collective intentionality, Margaret Gilbert and Raimo Tuomela, and seek to develop a view that solves the problems inherent in the accounts of Gilbert, Searle and Tuomela while nonetheless preserving the strengths of each account.

Although Chapter I.2 captures, I think, the intuitive plausibility of the collective intentionality theory (and thus the contemporary, naturalised version of Durkheim's methodological holism), it does not amount to a general argument establishing the real-

---

<sup>17</sup> I should add that Searle is not committed to the idea that the theory of collective intentionality would be in any sense Durkheimian or connected to the idea of a collective consciousness – indeed he categorically denies such connections (Searle 2006). However, I use the label to highlight a certain view of social action, not to suggest that the views of Durkheim and Searle are similar *tout court* (cf. Gross 2006 & Lukes 2007).

ity of collective intentionality. That task is left for Part II, which offers a detailed examination of the plausibility of the theory of collective intentionality.

Finally, both this Part and Part II, although defending Durkheimianism, remain neutral as to whether we should favour the Wittgensteinian (towards which Tuomela may be leaning) or anti-Wittgensteinian (Gilbert and Searle) reading of the Durkheimian position. The debate concerning Wittgensteinianism has huge consequences for the philosophy of mind, theory of meaning and the philosophy of the social sciences (especially the explanation versus understanding debate), and I do not wish to bind the arguments of Part I and II to any specific view on these issues. Part III, then, tackles the question of Wittgensteinianism and explicates its consequences. Accordingly, *my* defence of the theory of collective intentionality and, thus, naturalised methodological holism, is not complete until the end of Part III. However, arguments in earlier Parts should be important on their own right, *i.e.*, relevant also for those who fail to be impressed by my treatment of Wittgensteinianism in Part III.

CHAPTER I.2:  
 COLLECTIVE INTENTIONALITY  
 AS THE DRIVING FORCE OF SOCIAL REALITY

I.2.1 SEARLE'S BUILDING BLOCKS OF SOCIAL REALITY

According to Searle (1995), the fundamental building blocks of social reality are (i) the assignment of function, (ii) collective intentionality and (iii) constitutive rules. This Section discusses (i) and (iii), whereas (ii), collective intentionality, the heart of Durkheimianism in my sense, is analysed in the rest of this chapter by means of a critical examination of the definitions of the main collective intentionality theorists. As with the objectivity of social facts, also here my attitude towards Searle's views is somewhat ambivalent. Although I think that Searle's core intuitions are again more or less correct, I will nonetheless move quite quickly beyond his views.

Searle (1995, 9) points out that in addition to its intrinsic features, the world exhibits also features that are somehow relative to the intentionality of observers, users and other intentional, conscious agents. In my terminology, observer-relative features do not exist independently. Thus, the generation of observer-relative features does not add any new independent objects to the world, although it may add external and objective facts to the world. Searle's paradigmatic examples of observer-relative entities include objects such as screwdrivers and paperweights. Although both objects are no doubt independently existing material objects, their identities *qua* screwdrivers and paperweights are observer-relative non-independent facts. In Searle's view this is the first fundamental building block of social reality: as intentional agents humans *assign functions* to independently existing objects and to independent facts. Functions in Searle's sense are never intrinsic features of the world but are always observer-relative, assigned by intentional agents.

This definition allows Searle to be very explicit about when we are describing intrinsic features of the world and when we are describing observer-relative functions. For example, although the *causal processes* involved in the circulatory systems of humans are of course intrinsic features of the world, when we assert that the *function* of the heart is to pump blood, "we are doing something more than recording these intrinsic facts. We are situating these facts relative to a system of values that we hold." (Searle 1995, 14-15.) So the discovery of a "natural" function, such as the function of the heart, does not involve a discovery of any independent facts beyond the facts about causal



processes that are not functional as such. Functions are essentially normative and teleological, and one of the main achievements of modern natural science has been to clear away such notions. Thus, functional language amounts in Searle's view only to integrating non-normative causal facts into our value systems: "As far as nature is concerned intrinsically, there are no functional facts beyond causal facts" (Searle 1995, 16). To think otherwise amounts to nothing less than to the naturalistic fallacy of seeing certain natural processes as intrinsically normative.

Biologists and philosophers of biology by and large agree with Searle's point. Elliott Sober (1993, 82), for instance, observes that although biologists disagree on whether we should assign adaptive functions on the basis of what we know about the causal history of a trait or about its role in current ecological environment, biologists agree that in the world there are only non-normative causal processes. The normative talk in terms of functions is merely a convenient way for us to talk about the world. Consequently, the term *function* "does not occur ineliminably in any [biological] theory" although we use it "to talk about theories" (Sober 1993, 83). Sterelny and Griffiths (1999, 224) express this by saying that biology employs exclusively a "causal role conception of function".

A further distinction Searle draws is between *agentive* and *non-agentive* functions. Searle calls a function agentive if it consists of a fact that we as agents put (intentionally) an object to use, such as in the case of using a stone as a paperweight. Non-agentive functions are, then, assigned to natural objects and causal processes as part of a theoretical account of the phenomenon in question, such as in the case of saying that the function of the heart is to pump blood. The distinction between agentive and non-agentive functions, as Searle (1995, 20-21) admits, is not a clear-cut dichotomy. However, the distinction emphasises aptly that some functions are, so to speak, more dependent on continuous human activity than others. Agentive functions typically require that agents assigning the function continue to use the object in question in that function. A stone is a paperweight only insofar as it is used as one, whereas hearts keep pumping blood regardless of whether we describe that as their function or not.

Finally, from the point of view of social ontology perhaps the most interesting subclass is formed by those agentive functions, the functional component of which is not based on the purely causal capacities of the object to which the function is assigned. In the case of such functions it is quite natural to talk about the *meaning* or *status* assigned to the object in question. The fact that certain pieces of paper function as money, that we can use them in buying and selling *etc.*, is not based on the causal capacities of

those pieces in the same way as our ability to use stones as paperweights is due to the causal properties of stones. The clearest example of such an agentive status function might be the linguistic meaning assigned to certain marks on paper.

The possibility of assigning agentive status functions is clearly not enough to explain the ontology of money: even if I sincerely intend to treat randomly chosen pieces of paper as money – an objective fact about me – I cannot expect others to accept them as money. The status of money is a fact independent of my personal intentions; it obtains externally to me. Yet the status does not depend in any law-like way on the physical properties of the things to which the status is assigned either. The independent properties of the pieces of paper in question are largely irrelevant here; what matters is that the pieces are collectively *accepted* as money. In terms of Chapter 1.1, the fact that certain pieces of paper are money is an external fact that depends, not on any individual attitude as such, but on collective acceptance. Before examining the nature of such acceptance, let us first enrich the constructivist toolbox by redescribing some agentive status functions in terms of rules of acceptable behaviour.

For example, when we agree to treat certain pieces of paper as money the structure of the assignment of the function of money to the pieces can be expressed, according to Searle (1995, 28), by the following formula:

“X counts as Y” or “X counts as Y in context C”.

Searle’s example is that “Bills issued by the Bureau of Engraving and Printing (X) count as money (Y) in the United States (C)” (Searle 1995, 28). Since this formula is meant to capture a fundamental building block of social reality, the obvious problem here is that the formula seems to commit us to the existence of the very social institution the construction of which we are trying to analyse. In order for anything to be able to count as *Y*, it is obvious that *Y*, or at least the concept of *Y*, must already be defined satisfactorily.

Thus, if this is the final word Searle is able to say on this matter, we must conclude that Searle’s formula can at best explicate the way in which pre-existing social institutions get new instances or, in other words, how social reality and social institutions are reproduced and, perhaps, transformed. But then the theory could not be an account of how social reality is ultimately constructed. In order for some pieces of paper to count as money we must already understand, *prior* to assigning the function of

money to the pieces, what it is to be money, indeed what money as a social institution is.

Fortunately, Searle has a solution to this problem of circularity. Let us look in more detail at the structure of the new status *Y* that Searle's formula assigns to the *X*-element. To say that certain pieces of paper count as money amounts, according to Searle, actually to saying simply that the person in possession of those pieces has certain conventional rights and duties, or *deontic powers*, in certain situations *qua* the possessor of those pieces of paper. Thus, Searle argues, the "primitive structure" of the social status *Y* in the formula "*X* counts as *Y* in *C*" is actually that the person *S* in the situation *C* is allowed (or required) to perform certain actions – to use Searle's second formula, "*S* is enabled(*S* does *A*)" (Searle 1995, 104). Or to put it in terms of Searle's favourite example, to say that "*X*, this piece of paper, counts as *Y*, a five dollar bill," is in fact just another way of expressing the underlying *normative rule*: "*S*, the bearer of *X*, is enabled (*S* buys with *X* up to the value of five dollars)" (Searle 1995, 105).

Searle's example is not the clearest possible, since "buying" and "the value of five dollars" are obviously also social notions. But that just points to another typical feature of social reality, namely that most social facts are closely connected to other social facts. This, however, does not mean that "buying" and "the value of five dollars" cannot be given analyses similar to Searle's original example. Thus, *X*, *S*'s behaviour, counts as *Y*, buying goods, in a shop *C* insofar as *S* behaves according to the rules governing trading, and similarly for "shop", "trading" and so on. In Searle's view the content of social facts – and social reality in general – boils down to *rule-governed behaviour*. This, I think, is the core of Searle's answer to the form of circularity that appears to threaten the idea of using assignments of functions as a fundamental building block of social reality.<sup>18</sup>

Actually, in order to secure the non-circularity of his account Searle needs to introduce one more distinction, the one between *regulative* rules and *constitutive* rules. Regulative rules, as the name suggests, regulate already existing activity. Searle's example of a regulative rule is the rule "drive on the right-hand side of the road" (Searle 1995, 27). This rule regulates driving, but driving can exist prior to and independently of the existence of the rule: an activity can count as driving even if it violates that rule. The rule "drive on the right-hand side of the road" does not *constitute* the possibility of

---

<sup>18</sup> Note that this very same idea answers also another criticism sometimes directed against Searle (e.g., Smith 2003 & Thomasson 2003), namely that in certain cases there seem to be no physical *Xs* on which the social status *Y* could be imposed, such as in the case of the social status of a corporation. Since the content of the social facts in any case ultimately boils down to rules governing the appropriate behaviour of people, this is not a problem for Searle's theory (cf. Searle 2007b).

driving but merely regulates it – one may even argue that regulative rules presuppose the prior existence of the activity they are regulating.

Constitutive rules are essentially different. These rules constitute the very possibility of certain activities. Let us look at Searle's example:

[T]he rules of chess do not regulate an antecedently existing activity. It is not the case that there were a lot of people pushing bits of wood around on boards, and in order to prevent them from bumping into each other all the time and creating traffic jams, we had to regulate the activity. Rather, the rules of chess create the very possibility of playing chess. The rules are *constitutive* of chess in the sense that playing chess is constituted in part by acting in accord with the rules. If you don't follow at least a large subset of the rules, you are not playing chess. (Searle 1995, 27-28.)

The formula "*X counts as Y*" is meant to express a constitutive rule. Searle's suggestion that the content of the formula comes down to acceptable patterns of behaviour by defining the proper way to act in certain situations highlights the constitutive nature of the formula. In this manner the rule in Searle's formula (or, as in the case of chess, the system of rules collectively) is indeed constitutive. Constitutive rules create new types of action, such as playing chess. Thus, I conclude that Searle's theory succeeds in avoiding this form of circularity by appealing to constitutive rules.<sup>19</sup>

So far I have discussed functions and the corresponding constitutive rules neutrally relative to their status as facts. But surely the existence of such rules is not an independent fact in the terminology of Chapter I.1. As clearly they are objective or external *vis-à-vis* individual humans; I cannot just decide what are the rules constituting, for example, money. So the question is, where does the externality and objectivity come from? We have seen that Durkheim appealed to a collective consciousness that fixes social facts. In a sense, Searle's view is very similar. Social reality is based on *collective acceptance*: "There is exactly one primitive logical operation by which institutional reality is created and constituted. It has this form:

We collectively accept, acknowledge, recognize, go along with, etc., that (S has power (S does A))." (Searle 1995, 111.)

---

<sup>19</sup> Anthony Giddens (1984) argues that Searle's dichotomy between constitutive and regulative rules is misleading, since many, if not all, rules have both constitutive and regulative elements. For example, the seemingly purely constitutive rules of chess also regulate the behaviour of chess-players. Similarly, a seemingly purely regulative rule that all workers of a factory must clock in at 8.00 a.m., even if it does not constitute work as an institution, nevertheless "enters into the definition" "of a concept like 'industrial bureaucracy'" (Giddens 1984, 20) and is constitutive of that social phenomenon. However, within Searle's framework one may hold that the two aspects are often present in one rule and the clear-cut dichotomy is just a theoretical tool for highlighting the constitutive aspect that plays a crucial role in Searle's theory.

In Searle's view collective acceptance and other collective attitudes are based on *collective intentionality* qualitatively different from individual-mode intentionality. Moreover, Searle's ontological naturalism implies that mere acceptance, collective or individual, cannot bring into existence new independent causal factors, for that would be a form of ontological idealism. Thus, in Searle's view social facts are essentially normative practices that ontologically speaking boil down to collectively accepted norms of appropriate behaviour.

By integrating the notion of collective intentionality and collective assignment of functions into Searle's general framework, we get the following route from non-normative physical facts to normative social facts (cf. Figure 5.1 on p. 121 of Searle 1995). First, there are observer-independent physical facts, such as the fact that there is snow on Mt. Everest. Some facts, however, are observer-relative and thus non-independent, such as the fact that I am in pain. Further, some observer-relative facts are unlike the fact about my pain in the sense that they involve intentional, contentful mental states, such as my desire to see IFK Helsinki to win the Finnish Hockey League. Next, some of such intentional facts are *social* in the sense that they involve *collective intentionality*,<sup>20</sup> such as *our* intention to carry a table upstairs. Furthermore, certain social, intentional facts involve the assignment of an agentive function, such as the fact that we use this stone as a paperweight.

Finally, some of such collectively assigned agentive functions are not based on the causal capacities of the object the function is assigned to. Rather, sometimes we collectively assign *status functions*, such as that these pieces of paper are – or count as – money. And, as we saw, the monetary status is analysed in terms of norms of acceptable and required actions. Thus, in Searle's system social facts proper – *collectively assigned status functions* – are, ontologically speaking, collectively created, maintained and reproduced *norms* of action.

## 1.2.2 SEARLE ON COLLECTIVE INTENTIONALITY

A central aspect of Searle's theory of the epistemic objectivity of social facts is the claim that sometimes we assign functions collectively, *i.e.*, *together* in the strongest sense of the word. This kind of assignment is, according to Searle, based on *collective*

---

<sup>20</sup> This step, namely the Durkheimian requirement that the sociality involved must be something stronger than the sum of individual intentions, is what ultimately separates Searle's account from methodological individualism.

*intentionality*. Examples of collective intentionality are situations “where *I* am doing something only as part of *our* doing something” (Searle 1995, 23). Collective intentionality is for Searle a mode of intentionality qualitatively different from normal individual intentionality. This can be seen in the fact that in collective action collective intentionality appears to be conceptually and factually prior to individual intentionality, and hence, ultimately, Searle’s account is a Durkheimian – and not Individualistic – Account in the sense of I.1.5.

In collective action “the individual intentionality that each person has is derived from the collective intentionality that they [in ideal cases] share” (Searle 1995, 25). If I adopt collective we-mode intentionality, I figure out what is *our* goal in the situation at hand and what is *our* best means for achieving that end, and then form a we-intention of the form “*We* will do *X*”. Only when I have reached this stage I *derive* my individual-mode intention from the we-intention. Hence my individual-mode intention to do *Y* is subordinate to the we-intention in the sense that I first figure out that *we* should do *X*, and only then that *in order for us* to do *X*, I should do *Y* as my part of *X*, and hence I set out to do *Y*. In sum, collective intentionality is not in Searle’s picture reducible to individual intentionality. Collective intentionality is *sui generis*.

Unfortunately, Searle does not analyse his notion of collective we-mode intentionality much further. He merely emphasises that we should not think that collective intentionality commits us to the existence of “some Hegelian world spirit, a collective consciousness, or something equally implausible” (Searle 1995, 25). Searle makes it very clear that in his view individuals are all the *agents* there are.<sup>21</sup> He thinks that sometimes the psychology and action of individuals is irreducibly in the we-mode, *i.e.*, based on collective intentionality in the sense that the actions and attitudes are appropriately conceptualised only such that the collective “we” is the formal subject of the attitudes and actions.

For Searle, collective intentionality is perhaps the most fundamental building block of social reality – Searle even goes so far as to stipulate that “social fact” refers to any fact involving collective intentionality (Searle 1995, 26). Although I think that Searle’s account of collective intentionality is basically correct, I also think that we ought to say more explicitly how we understand collective intentionality to work. In a sense, Searle’s account does not add anything more to Durkheim’s puzzling statements

---

<sup>21</sup> In fact, Searle says that all intentionality must be in the heads of individuals. However, as I have explained, I want this and the next Part to be neutral regarding Wittgensteinianism or indeed any other specific theory in the philosophy of mind, and thus I prefer my formulation which leaves room for *externalism* that says that no intentionality – be it individual or collective – is strictly speaking in the heads of individuals.

about collective consciousness than the naturalistic conviction that, whatever collective intentionality amounts to, it involves only individuals. I believe that by analysing Margaret Gilbert's and Raimo Tuomela's work on this topic we can illuminate the fine structure of collective intentionality and, thereby, the Durkheimian rejection of methodological individualism.

### 1.2.3 GILBERT ON COLLECTIVE INTENTIONALITY<sup>22</sup>

The animating idea in Margaret Gilbert's work on collective intentionality is her claim that a precondition of all social facts proper is the ability to create a collective *group will* out of individual wills. Gilbert's main thesis is that since naturalism compels us to accept (*OI*), the ultimate subject matter of social science cannot be social entities as independently existing objects, but social action that constitutes social facts. Furthermore, for action to be truly social or collective, it must in her view be based on collective attitudes<sup>23</sup> attributable to a social group. However, as we shall see, collective intentionality in Gilbert's group-centred sense presupposes collective intentionality in the we-mode sense.

Gilbert subscribes to the view that our account of collective intentionality must not commit us to the existence of group minds or anything ontologically equally suspect. Gilbert hypothesises that the acceptance of this naturalistic principle is the main reason why many philosophers and sociologists have a tendency to be sceptical about the very possibility of collective intentionality. Gilbert thinks that these anti-collectivists accept the thesis of *psychologism about intentionality*. According to this thesis, "in order for the English predicate '... believes' [or '... intends'] to apply to something, that thing would have to have a mind" (Gilbert 1989, 238).

---

<sup>22</sup> In what follows I concentrate largely on Gilbert (1989), for I am interested in the very foundations of her views and her later works (in particular 1996, 2000 & 2006) build mainly on the theory of 1989 in discussing moral and political philosophy and other issues I cannot include in this study. Moreover, Gilbert's (e.g., 2002, see also Hakli 2006 for a review) interests have, in line with Velleman's (1997) criticism, turned more and more to the problem of how several individuals can share the same intentional attitude so that we can talk literally about the group's attitude. This is very different from the aims of Searle and Tuomela, who are mainly analysing the possibility of assigning we-mode attitudes to *individuals* and deny the literal attribution of attitudes to groups. In a sense this difference need not be terribly deep, for – as I explain below – a widespread we-mode attitude can be said to be the we-group's attitude (see Tollefsen (2004) for further discussion). Moreover, below I argue that Gilbert would do better by explicitly building on Searle and Tuomela here. This is not to say that securing that the individuals indeed share the *same* intention (cf. Velleman 1997) is not a difficult problem (and most certainly it is an *important* problem for applications of the collective intentionality theory, e.g., when analysing collective responsibility), but simply that this dissertation is not the place to study that problem (see, for example, Pettit 2003 and Pettit & Schweikard 2006 that concentrate largely on this issue).

<sup>23</sup> Gilbert discusses mainly specific collective attitudes, but her discussion generalises *mutatis mutandis* to intentionality in general.

The thesis of psychologism about intentionality is then combined with the naturalistic common-sense thesis of *anti-psychologism about social groups*, which states that groups do not have minds of their own (cf., however, the provocatively-titled Pettit 2003). Hence, it is thought, if collective intentionality is to be a meaningful concept at all, it must come down to the sum of individual intentions of a set of people, in which case it is actually not a *sui generis* phenomenon. Gilbert calls this kind of view the *summative account* of collective intentionality, of which the paradigmatic example is Quinton (1975-76). The Individualistic Account of I.1.5 would have no quarrel with this kind of collective intentionality. Gilbert's goal is to create a non-summative but ontologically naturalistic theory of collective intentionality that goes beyond the Individualistic Account. In what follows I analyse her argumentation and argue that although Gilbert is successful in highlighting many important aspects of collective intentionality, in the end her theory remains somewhat unsatisfactory, because in my view the only real option between individualistic, summative accounts and rejections of (*OI*) is the we-mode account of collective intentionality, and Gilbert is reluctant to build explicitly on it.

Gilbert motivates her discussion by pointing out that the summative account hardly captures the collectivity we want from a theory of collective intentionality. Even if all – or most of – the brown-haired people in London happen to have the attitude *X*, this is obviously insufficient to create a *collective* attitude *X* in any interesting sense. Gilbert emphasises that it will not help the summative account to add the requirement that the members of the set in question must know that each member has this intention. Most normal adults will know if they are brown-haired persons living in London, and although brown-hairedness supposedly does not carry with it any special tendencies or attitudes, it is reasonable to assume that as normal adult humans the brown-haired Londoners will have some typical intentional attitudes in common (at least in the sense of conditional or situation-relative dispositions – if it makes the case for summative account stronger, we may even suppose that these attitudes are related to the fact that they have brown hair). Given all this, the members of the set will also know that each member of the set of brown-haired Londoners has the attitude *X*. Still, argues Gilbert (1989, 271), our intuitions about social action tell us that we should not say that the brown-haired Londoners have a *collective* attitude *X*. To share the same attitude and know this about each other does not suffice for having the attitude collectively and together (cf. Searle 1990).



The primary problem with the example presented above appears to be that the set of brown-haired Londoners is not a proper social group in the sense of having any sort of social significance. The members of the set are defined to belong to the same set by introducing a completely random criterion they happen to satisfy. Thus, the set is a set only relative to the arbitrary classification system, which, moreover, is not based on the social behaviour (or something similar) of the members and, hence, the set does not correspond to any social group. Intuitively it seems to be essential that a collective attitude is such that it is associated with a social group.

However, I am examining the claim that collective intentionality is a fundamental building block of social reality and of social facts – including the fact that some individuals form a social group. Hence, to say, for example, that a collective intention to do  $X$  is the sum of the individual intentions of the members of a social group  $G$  to do  $X$  would be circular. Similarly, we saw above that Gilbert thinks that collective action constituting social facts must be based on collective intentions, and for collective intentions to be truly collective they must be held by social groups. So here is the problem: Either collective intentionality presupposes some social facts (about social groups), in which case collective intentionality cannot be a fundamental building block of social facts, or collective intentionality is not a real phenomenon, in which case it by definition cannot be a fundamental building block of social facts.<sup>24</sup> This is one of the main reasons why, for example, David-Hillel Ruben (1985) thinks that all explications of social reality in the spirit of (*OI*) fail and we must accept social entities as ontologically primitive.

In order to avoid this dilemma, Gilbert thinks that all summative accounts of collective intentionality must be rejected, for to make the sum of individual attitudes socially significant seems to require a social criterion for picking out the relevant aggregations. Thus, we need an account of a process which is simultaneously an account of the construction of a social group and an account of forming a collective attitude. The central notion here is that of collective action (Gilbert 1996, 178) based on collective intentionality: a set of people will form a social group capable of collective action if and only if they are able to construct a collective intention to perform a collective action.<sup>25</sup> The notions of a social group and forming a collective attitude belong together.

---

<sup>24</sup> I should emphasise that the looming circularity is a problem also for Searle, for in his view “collective intentionality seems to presuppose some level of sense of community before it can ever function” (Searle 1990, 413). Searle, however, does not address the circularity problem in satisfactory detail – and thus, for example, Jennifer Hornsby (1997, 431) sees it as a major problem for Searle’s theory.

<sup>25</sup> In fact, it seems that forming any (non-summative) collective attitude suffices for construction of a social group. For example, agents who believe *collectively* that  $P$  form a social group in the sense a set of individuals who all happen to believe individually that  $P$  does not.

Gilbert's problem is how a number of individuals can join together to construct a unit that is capable of constructing a collective attitude, constituting thereby a social group – or, in other words, how a set of individuals can form a *plural subject* of an attitude or action, since “a set of people constitute a social group if and only if they constitute a plural subject” (Gilbert 1989, 204). What she needs is a non-summative theory of social action that explains what it is for two or more agents to act together in the strong sense of forming a plural subject, to which the authorship of actions may correctly be ascribed.

Gilbert's point of departure is Georg Simmel's view, according to which the unity required for social action is constituted by the agents being conscious that they constitute a unity (Simmel 1908/1971, 7). Again, this sounds rather circular – the existence of a group is a precondition for anyone to *know* that the group exist.<sup>26</sup> However, Gilbert emphasises that the claim is *not* that “a social group is wholly constituted by people knowing that they form a social group with certain others” (Gilbert 1989, 147). Instead, Gilbert suggests the following statement – which she calls the Simmelian schema – to be the correct reading of Simmel's view: “a social group's existence is basically a matter of the members of a set of people being conscious that they are linked by a certain special tie” (Gilbert 1989, 148-149). Being conscious of the “special tie” is the internal criterion a set of people must satisfy in order to form a plural subject and to act together in the strong sense Gilbert is after.

Gilbert's first, rather obviously unsatisfactory hypothesis for an analysis of acting together – and the special tie – is the following:<sup>27</sup>

(H1) Agents  $A_1, \dots, A_n$  in situation  $S$  satisfy the internal criteria for doing  $X$  together in the strong sense if and only if each of  $A_1, \dots, A_n$  is intending to do  $X$  in the situation  $S$ .

---

<sup>26</sup> This problem is the core of Ruben's (1985) rejection of (OI). Ruben argues that the attitudes of members of a social group cannot constitute the group, since plausible candidates for such propositional attitudes are, according to him, always *about* the group, and hence all such accounts are bound to be circular. I think Ruben is partly correct: if we are to claim that the attitudes of individuals constitute social institutions, then the attitudes must be about something else than the institution in question. As in 1.2.1, also here I argue that ultimately the attitudes are about acceptable and required patterns of behaviour, not about institutions as such. This is a possibility Ruben does not discuss.

<sup>27</sup> In what follows the hypotheses (H1), (H2), (H3) and (H4) are my reconstructions and explications of Gilbert's discussions about the conditions of “sharing in action” and the criteria for the pronoun “we” to be used properly to refer to a we-group, the archetype of a plural subject (Gilbert 1989, 154-203). Since I am here primarily interested in the problem of collective intentionality, I discuss exclusively what might be called the “internal criteria” of action, *i.e.*, the relevant intentions and beliefs. However, satisfying the internal criteria may not suffice for collective action, which often involves also external components such as observable behaviour.

(H1) requires only that two or more agents are engaged in the same activity in the same time and place, and this is clearly insufficient for acting together in the strong sense. Anyone who has ever suffered a trip in an overcrowded London underground train during the rush hour can testify that the passengers, who are all travelling situated far too close to one another, are nonetheless not *travelling together* in the sense of forming a social group of any sort. They are all preoccupied with their own personal projects that just happen to share the feature of involving an underground ride at that particular time. In order for agents to be doing *X* together they need to have as their goal that they do *X* together.

This suggestion can be formulated as the following hypothesis:

(H2) Agents  $A_1, \dots, A_n$  satisfy the internal criteria for acting together in the strong sense if and only if each of  $A_1, \dots, A_n$  is willing to share in action (or a range of actions) with  $A_1, \dots, A_n$  in circumstances of the type at hand.

Note that this definition is “objective” in the sense that it does not require that the agents know or even believe anything about the views of the others. Interestingly, this objectivity also means that (H2) in fact fails to capture the sufficient criteria for collective action in the strong sense we are aiming to analyse. Gilbert’s example of this is a case of two persons, both of whom hope and intend to travel together, but who are for some reason too shy to communicate this desire to the other (Gilbert 1989, 157-158). This lack of communication leads them to fail to travel *together* in the strong sense, since although they satisfy (H2), they do not explicitly share the goal of travelling together. However, in my view the main problem here is not the lack of communication, but rather that although the goals the persons have are directed towards the other and their mutual co-operation, the goals are, just as in the first case, still nevertheless personal goals in the sense that they are held individually by disparate agents.

As Gilbert (1989, 160-161) emphasises, the crucial point is not the characterisation of the goal, but rather *whose* goal it is.<sup>28</sup> If the goal is held by an individual agent, it motivates individual action. But if the goal is held by a collective, it can be the goal of collective action, pursued together by acting together. Recall, however, that we cannot

---

<sup>28</sup> This distinction between the directedness and content of goals and attitudes on the one hand and the agent to which the goal or attitude is ascribed on the other plays a crucial role in the main argument of Part II, and thus it is important to be clear of the distinction already in the present context. However, where I wish to account for this distinction ultimately in terms of distinction between the we-mode and the I-mode applied to the attitudes of the relevant individuals, Gilbert (*e.g.*, 2002) often talks in terms of the the group really having the attitude.

appeal to a pre-existing social group here. Rather, the task is to explain the construction of a social group in virtue of a set of agents satisfying the criteria for acting together. In Gilbert's view this requires that we include a condition concerning knowledge about the views of the others to the hypothesis. Let us next look at the following hypothesis (*H3*):

- (*H3*) Agents  $A_1, \dots, A_n$  satisfy the internal criteria for acting together in the strong sense if and only if
- (i) each of  $A_1, \dots, A_n$  is willing to share in action (or a range of actions) with  $A_1, \dots, A_n$  in circumstances of the type at issue
  - (ii) each of  $A_1, \dots, A_n$  knows that (i).

Moreover, in order to really share in action, the clause (i) in (*H3*) should be *common knowledge* within the relevant set of agents  $A_1, \dots, A_n$ . Hence, we need to add a clause (iii):

- (iii) each of  $A_1, \dots, A_n$  knows that (ii), and so on with higher-order knowledge as far as one cares to go (Gilbert 1989, 161 ff.).

However, Gilbert does not believe that a hypothesis such as (*H3*), even when strengthened with the clause (iii), can be an adequate analysis of the internal criteria for collective action. The problem is that the agents are still acting individually, performing their own tasks that may be synchronised and other-regarding, and the fact about this is common knowledge, but the tasks and goals – indeed the *intentions* – are, nevertheless, still personal tasks, goals and intentions of individual agents. The situation captured by (i)-(iii) of (*H3*) does not give rise to *normative attitudes* that characterise collective intentionality and collective action in the sense of acting together (Gilbert 1989, 162) – recall also Searle's view that ontologically speaking social facts boil down to norms of appropriate action. Gilbert thinks that what is missing here is the element of being *jointly committed* to reaching the goal, or *accepting the goal jointly* to be *their* goal. Or, as Gilbert likes to put it, “each must manifest his willingness to constitute with the other[s] a *plural subject* of the goal” (Gilbert 1989, 163). Thus, we need to add the “idea that each one is aware of each one's willingness by virtue of each one's expression of his willingness to the others” (Gilbert 1989, 182).

Gilbert appears to think that this amounts to something like (*H4*):

- (H4) Agents  $A_1, \dots, A_n$  satisfy the internal criteria for acting together in the strong sense if and only if
- (i) each of  $A_1, \dots, A_n$  is willing to share in action (or a range of actions) with  $A_1, \dots, A_n$  in circumstances of the type at hand
  - (ii) each of  $A_1, \dots, A_n$  knows that (i) as a result of each one's having, in effect, expressed this willingness to each of the others
  - (iii) each of  $A_1, \dots, A_n$  knows that (ii), and so on as far as one cares to go.

In sum, (H4) says that in order for two or more agents to do X together in the strong sense, the agents must (i) each be willing to share in action with the others, (ii) express this willingness to the others, and (iii) all this must be common knowledge among the participants. Gilbert thinks that the clauses (i)-(iii) in (H4) are necessary and sufficient conditions for the agents  $A_1, \dots, A_n$  to satisfy the internal criteria for acting together. Moreover, in Gilbert's view this is equivalent to saying that if the clauses (i)-(iii) of (H4) are satisfied, the agents  $A_1, \dots, A_n$  are able to create a collective intention to perform a collective action, or, as this can be summed up, they form a social group in the strong sense of forming a plural subject of a collective intention. Gilbert appears to think that the expression of willingness to participate to the pursuit of a joint goal amounts to a commitment that brings in the rights and duties that a joint goal places upon the participants in the plural subject in question. Here I think Gilbert is after the idea that, as Searle puts it, in the context of collective action individual action-intentions (individual roles in a collective task) are derived from an irreducible collective intention and thus collective considerations are conceptually prior to individual-mode considerations. But it is very difficult to hold this in Gilbert's framework.

In particular, (H4) appears, *contra* Searle's theory of collective intentionality, to employ standard individual-mode attitudes. The attitudes in (H4) are intentions to form (with the others) the kind of plural subject that can adopt a goal of its own, and the individual action-intentions are then derived from this collective intention. In short, (H4) is not an analysis of a collective we-mode action-intention, but an analysis of what it takes to form a plural subject that can have such intentions. However, if this is the correct reading of Gilbert's account, then Gilbert nonetheless fails to deliver what she explicitly wants to deliver, namely of a theory of social action that goes beyond the Individualistic Account of social facts by accepting an ontologically naturalistic version of Durkheimianism (I.1.5). A card-carrying methodological individualist would have no problems in accounting for (H4); she would only say that of course individuals may have the

kind of pro-social I-mode attitudes that (*H4*) talks about. The joint acceptance and joint commitment cannot be mere matters of explicit expressions if we are to exceed the Individualistic Account.

As we know (I.1.5), Gilbert is explicitly committed to Durkheimianism and, consequently, she is forced to de-individualise her analysis afterwards. This is what she seeks to do in her later writings (cf. Gilbert 2003, 53 ff., where she admits that, although basically correct, the 1989 view must be socialised). Let us look, for example, at the analysis of group action in Gilbert (2000):

A group G performed an action A if and only if, roughly, the members of G were jointly committed to accepting as a body the relevant goal X, and acting in the light of this commitment, relevant members of G acted so as to bring X about.

(Gilbert 2000, 131.)

Here the notions doing the philosophical work are again “jointly committing to accept as a body” and “acting in light of this commitment”. Gilbert explicitly contrasts *joint* commitment with *personal* commitment and explains that it is precisely this notion that brings in the strong anti-individualism of her view (e.g., Gilbert 1996, 7 ff.; 2000, 3 ff. & 130-131; 2003, 54). However, Gilbert does not really analyse this crucial notion. Indeed, what Gilbert requires of a joint commitment even in her later works is merely that the participating individuals are ready to commit themselves to a joint enterprise in the sense of (i) being willing to share in action with the others and (ii) communicating this to others (Gilbert 1996, 366). This, however, does not add anything to (*H4*). Gilbert appears to think that the *jointness* of the attitudes of the individuals forming a plural subject guarantees that in fact (*H4*) operates with Durkheimian attitudes that go beyond the Individualistic Account of I.1.5. However, Gilbert does not tell us what this jointness of the relevant attitudes ultimately is. Requiring mutual communication is clearly insufficient. In my view, the best way to understand the jointness of the attitudes is in terms of Searle’s (I.2.2) and Tuomela’s (I.2.4) we-mode attitudes.<sup>29</sup>

---

<sup>29</sup> Hence, from the point of view of the present analysis the main difference between Gilbert on the one hand and Searle and Tuomela on the other is one of focus: Gilbert aims at analysing what she calls full-blown plural subject phenomena, whereas Searle and Tuomela are interested (also) in the psychological foundations of such phenomena. Gilbert (2007) accepts this division of labour but argues that Searle’s conceptual apparatus is insufficient for addressing what she sees as the crucial question, namely how different individuals can literally share the same intention in the strong sense that allows us to speak of a group’s intention that brings in morally relevant group responsibilities (cf. Velleman 1997). For the present essay, however, this is not crucial: my aim is to explicate the factor that makes the collective intentionality theory anti-individualistic (and not all the possible applications of the theory, no matter how important and interesting they are), and also in Gilbert’s view this factor is the “jointness” of the relevant participant attitudes, which I analyse in terms of the we-mode.

Hence, according to this reading of Gilbert's plural subject theory, the we-mode attitudes (bringing in "joint commitment to accept a goal as a body") of individuals guide the individuals to join together to pursue a collective goal. And when they have formed such a collective goal, the individuals can derive their individual roles in the collective task of realising the goal and in this sense "act in the light of their joint commitment". Thus, "[t]he goal of any joint action is seen by the participants as the goal of a plural subject" (Gilbert 1989, 164), and the sub-goals of the participants are derived from that goal.

Although this reading presents Gilbert's account as largely similar to Searle's view, the transparency of Gilbert's analysis allows us to address the circularity problem more explicitly than within Searle's theory. To recapitulate, the problem is that we do not want to end up saying that the existence of a social group  $X$  depends on anyone's consciousness that there is an  $X$ , since the "possibility of the latter appears to depend on the former" (Gilbert 1989, 222; cf. Ruben 1985). If we look at what Gilbert has actually said, I think we can conclude that the account is not circular in this sense. After all, Gilbert thinks that the existence of a social group comes down to the following features: (i) those who are to form a group are willing to jointly commit to uniting their individual wills to create a pool of wills (Gilbert calls this the *willed unity condition*), and, (ii), everyone must express the will mentioned in (i) to the others (the *expression condition*),<sup>30</sup> and, finally, (iii), this expression of willingness must be common knowledge (the *common knowledge condition*) (Gilbert 1989, 222-223).

Moreover, if it is indeed essential (as I think it is) that collective attitudes are irreducible to individually held attitudes, we have here further evidence for the inevitable failure of any summative account. As Gilbert puts it, it is "both logically necessary and logically sufficient for the truth of the ascription of group belief [...], roughly, that all or most members of the group have expressed willingness to let a certain view 'stand' as the view of the group" (Gilbert 1989, 289). This makes it possible for Gilbert to acknowledge another important feature of collective attitudes: the collective attitude of the group does not have to be an attitude that the majority of the members of the group hold personally (Gilbert 1989, 300). Indeed, the jointly accepted group attitude needs not be the personal I-mode attitude of *any* member of the group (Gilbert 1989, 298-299; 1996, 200 ff.).

---

<sup>30</sup> In fact, in my view implicit, tacit expression is often enough – sometimes even mere holding back resistance suffices. Gilbert thinks she needs explicit expression to ground unambiguously the deontic elements of joint action required, in particular, for applications in moral philosophy. However, as becomes clear below, I think one ought not to build this element into the theory of collective intentionality.

This is perhaps the main argument against summative accounts of collective group attitudes. A summative account that captures a widespread *individual* view may, for example, be a crucial piece of information for explaining why the *collective* came to change its view in the matter in question. The fact about how well the belief of a group corresponds to the summative majority view of the members, or to a given personal view, may be an important factor in social scientific explanations, but its importance does not affect the “logical” independence of a group belief from personal (individual) beliefs (Gilbert 1989, 310).

However, a surprising feature in Gilbert’s 1989 view is that she appears to think that a group  $G$ ’s “joint acceptance” of a given view  $P$  presupposes that it is common knowledge in  $G$  that the members of  $G$  *have expressed their willingness to accept  $P$*  to be the view of the group. I do not quite see why this is assumed. Could it not be sufficient to say that the members have expressed their willingness to find *a* view that can be adopted as the view of the group  $G$ , or their willingness to let *any* view that is selected through a certain process to stand as their view? What I have in mind is that in some cases the members of a group might want to accept a system of forming a group view which does not require all group members to participate actively in that process. For example, the citizens of a nation state might accept that some people (say, the members of the parliament) represent them in certain matters and form (*e.g.*, by discussing and voting) a view that will then count as the view of the whole group including those citizens who are not members of the parliament. In this example the citizens have not expressed (even implicitly) their willingness to accept  $P$  (as opposed to some other view  $Q$ ) as the view of their social group (*i.e.*, the nation), but their willingness to accept *whatever* view it is that the parliament decides to be the view of the whole social group.<sup>31</sup>

I think that this possibility is in fact consistent with (H4). Indeed, in a later work, Gilbert (1996, 209) explicitly accepts the kind of scenario I have described above, *i.e.*, that the citizens of a country accept as their view whatever view the government will adopt. There Gilbert simply states that this does not speak against the 1989 plural subject theory, although the original 1989 formulation was explicitly argued to require strengthening in terms of what she called *the members’ knowledge principle*. According to this principle, a necessary condition for a group  $G$  to have the view  $P$  is that the

---

<sup>31</sup> This principle of the division of social labour is highly analogous to Putnam’s (1975) famous principle of the division of linguistic labour. Indeed, in Part III and Appendix it is argued that in the case of one fundamental social institution, language, it is absolutely crucial that individuals can commit themselves to views the content of which they do not personally know.



members of *G* know that *G* has the view *P* (Gilbert 1989, 259). However, the members' knowledge principle is not necessary for what is valuable in Gilbert's theory. In order to form a plural subject, only the intention to find a collectively acceptable view – and the expression of this intention – is supposed to be common knowledge. In fact, I do not believe that the members' knowledge principle is even a desirable feature of a theory of collective attitudes including collective intentions, since it would make it much more difficult to deal satisfactorily with many complex features of social reality and group phenomena – and indeed also Gilbert (1996) is at least implicitly willing to modify the condition (Gilbert 2000 remains silent concerning this issue): Members must know that there is a way of forming a group belief, not what the belief is.

In sum, although Gilbert's account takes us in a sense further than Searle's theory (especially in addressing the circularity problem explicitly), it nonetheless remains somewhat unsatisfactory in certain aspects – and in my view does indeed worse than Searle's account in many crucial points. For example, Gilbert's formulations are expressed in terms of the members of a collective being *in fact* willing to share in action with the other members and the members *knowing* this to be the case. But surely we would like to be able to account for cases where only some agents are willing to share in action because they *believe* – falsely, as it happens – that the others are willing to do so as well, whereas in fact the others do not share the willingness and are rather prepared to take advantage of the naivety of the believers.

Similarly, for Gilbert a collective intention really is a shared attitude of the plural subject, whereas for Searle a collective intention is a we-mode attitude (shared or not) of individuals. A natural way to resolve this difference is simply to say that where Searle (and Tuomela) are interested in the psychological preconditions of social action, Gilbert analyses social actions that already presuppose the we-mode psychology (in the guise of joint commitment). Gilbert, however, is not happy with this reading. Sometimes she writes as if full-blown group attitudes were required *prior* to we-mode attitudes to make conceptual room for the admittedly important ideas that (i) sometimes no individual personally holds the group view and (ii) in collective action individual action-intentions (roles) are derived from a collective intention.<sup>32</sup> However, as I demon-

---

<sup>32</sup> Gilbert (2007) also suggests that on conceptual grounds a Searlean must think that we-mode intentions presuppose prior group intentions, because in Searle's (1983) theory the content of an intention represents its conditions of satisfaction and, according to Gilbert, a central satisfaction condition of an individual's we-intention is that there really is a group-level collective intention. This argument, however, is based on a misguided reading of Searle's theory of intentions: in the Searlean view, the satisfaction conditions of intentions have the world-to-mind direction of fit (with some complications relating to causal self-referentiality that need not concern us here) and not the mind-to-world direction of fit that, e.g., beliefs have. It is not a satisfaction condition of intending to order a beer that one already has ordered a beer.

strate in I.2.4, the same conceptual room can be made in terms of the we-mode versus I-mode distinction without the ontologically dubious appeal to the attitude of the group, literally shared by group-members.

Moreover, there are reasons to favour the way in which Searle and Tuomela build collective attitudes in Gilbert's sense on we-mode attitudes. First, as we saw, the we-mode view allows us to be more explicit concerning the anti-individualistic *jointness* aspect of the constitutive attitudes of the group-members that also Gilbert relies on but which remains rather opaque in Gilbert's writings. Moreover, Gilbert's group attitude view does not leave room for the idea that only one individual can adopt the anti-individualistic stance – and switch between the we-mode and the I-mode. This idea plays a major role in Parts II and III of this dissertation, and Part II indeed provides what I think as the decisive argument in favour of the we-mode view, and thus the interpretation of Gilbert that sees her discussions as highly compatible with the we-mode theory is good for Gilbert too.

However, to get there we first need a more detailed analysis of collective we-mode intentionality that is able to accommodate these considerations foreign to Gilbert and not focused on by Searle. In other words, we need to analyse collective intentionality from the point of view of a single agent who is willing to share in action with others, because she believes – rightly or wrongly – that also the others are willing to form a plural subject. To do this, let us turn to the third major theorist of collective intentionality, Raimo Tuomela.<sup>33</sup>

---

<sup>33</sup> In Gilbert's view favouring the we-mode at the expense of the fundamentality of shared attitudes attributed to the group is not a goal worth pursuing. However, in line with my strictly naturalistic and individualistic ontology, I share Searle's and Tuomela's emphasis on the importance of the we-mode providing a coherent way of talking *as if* groups were full-blown agents with intentional attitudes, without the need to talk about literally shared attitudes of the group – and thus to get the results of Gilbert with less worrying ontology. Moreover, I fully agree with Searle (personal communication) that by explicitly assuming full-blown language in her analyses of shared intentions (in particular, the expression condition, cf. Footnote 30) Gilbert's discussion is not suitable for a general theory of social institutions, for surely language itself is a central institution indeed (however, in Part III it will become clear that my view of the construction of language is very different from Searle's, which builds on the notion of intentionality as a biologically primitive feature of the world (cf. Searle 2007a)). Thus, when Velleman (1997) and Gilbert base their accounts of securing the literal sharedness of an intention to public, linguistic commitments, this is by no means a strategy foreign to Searle or Tuomela (see especially Tuomela's (2000, 2002, 2007) Bulletin Board View); Searle and Tuomela just emphasise the importance of analysing the we-mode attitudes that make such language-based co-operation possible – and which I have argued also Gilbert implicitly relies on (by emphasising the "jointness" of the attitudes in (H4)) as the very source of her anti-individualism. Following Miller (2007), one could say that Gilbert's notion of commitment conflates two senses of "commitment": the non-normative sense of irreversibility in making up one's mind characteristic of intentions (e.g., jumping off a cliff; at some point the jumper commits herself, there is no turning back) and the institutional, normative sense of putting oneself under a public obligation. The alleged jointness of the former comes from the we-mode of the relevant attitudes and of the latter from applying the social institution of promising or other relevant linguistic tools. The present essay concerns primarily the former aspect and accordingly I prefer Searle's and Tuomela's framework, where it is easy enough to

## 1.2.4 TUOMELA ON COLLECTIVE INTENTIONALITY

Although the theory of collective intentionality allows us to talk of group agents in a more substantial sense than the summative view, the commitment to *(OI)* requires that ontologically speaking individual agents are the only agents there are in the social world. Accordingly, I think the question of how individual agents come to adopt collective we-mode attitudes – which, when shared, can then be seen as the attitudes of a group or a plural subject – is an even more important question than Gilbert’s analysis of when a set of agents forms a plural subject.

Recall that Searle thinks that when an agent’s intentionality takes the form of we-mode intentionality, the agent’s action is rationalisable only to the extent the agent is seen acting essentially as part of her group. Although I think Searle’s view is basically correct, in this section I will seek to define the view in more detail. I start with Raimo Tuomela’s (2000) definition, although I will abandon it almost immediately. However, although Tuomela’s explicit definition fails, his informal descriptions (and what I take to be his core intuitions) are, I argue, correct, and hence what follows ought to be seen as my explication of the core of Tuomela’s views rather than as a criticism of Tuomela’s position.

According to Tuomela, the most central notion in this context is that of a collective attitude. It is a general notion that has collective intention as a special case. Following Tuomela, let  $x$  stand for a member of a collective,  $B_x(q)$  for  $x$ ’s belief that  $q$ ,  $ATT_x(p)$  for  $x$ ’s attitude with the content  $p$ ,  $WATT_x(p)$  for a we-attitude with the content  $p$  held by  $x$ ,  $MB(p)$  for a mutual belief (in the relevant collective) that  $p$ , and, finally,  $ATT_E(p)$  for everybody in the collective having the attitude  $p$ .<sup>34</sup> With these symbols, Tuomela defines a collective we-attitude held by an individual  $x$  as follows (Tuomela 2000, 50; 2002, 23):

$$WATT_x(p) \leftrightarrow ATT_x(p) \ \& \ B_x(ATT_E(p) \ \& \ MB(ATT_E(p))).$$

---

keep these two different aspects explicitly separated. I am grateful to John Searle and the members of his research group on social ontology at UC Berkeley for discussing these issues with me.

<sup>34</sup> Of course it is not realistic to require literally *everybody* to hold the attitude. Usually it is sufficient that *almost* all the members, *most* of them or the *operative members* hold it. But these are complications required for particular applications that does not need to concern us when explicating the basic structure of a collective we-attitude.

Although as it stands Tuomela's definition does not require it, "[a]n intuitive idea related to we-attitudes is that a person has ATT in part *because* the others have ATT and this is mutually believed in the group" (Tuomela 2000, 50).<sup>35</sup> To see how this idea can be built in to the definition, let us move from the general level of collective attitudes to the special case we are primarily interested in, namely to that of forming a collective we-intention.

Let  $I_x(G)$  stand for  $x$  intending (in the individual mode)  $G$ . We are interested in a situation where the attitude in question is  $x$ 's intention, *i.e.*, where  $ATT_x(p) = I_x(G)$ . As an illustration we can think of Ruben's (1985, 119) example of  $x$ 's engaging in the British custom of drinking tea at breakfast. Thus, let  $I_x(G)$  be " $x$  intends (drink tea at breakfast)". The fact that  $x$  intends to *participate* in an existing custom means that her intention is not simply of the form  $I_x(G)$ , since it would not include the *social* element of her intention (Ruben 1985, 121-122). When  $G$  is  $x$ 's we-mode intention, such as the intention to participate in a social custom, let us write it as  $I_{we,x}(G)$ .

Note that  $I_{we,x}(\textit{to drink tea at breakfast})$  reads only as " $x$  we-intends (to participate in the custom of the relevant collective) to drink tea at breakfast"; her success in really participating depends of course on, for example, there really being such a custom in which she can participate. Thus, Tuomela's framework will immediately allow us to reach beyond Gilbert's account, which required the collective practice (or at least the shared willingness to construct one) to really *be* there and the agents to *know* this. With these clarifications, we can now write Tuomela's definition in the following form (see Tuomela 2000, 51):

$$I_{we,x}(G) =_{df} I_x(G) \ \& \ B_x(I_E(G) \ \& \ MB(I_E(G))).$$

When applied to Ruben's example this definition says that  $x$  intends to participate in the custom (of the relevant collective) to drink tea at breakfast ( $x$  we-intends to drink tea at breakfast) if and only if (i) she intends to drink tea at breakfast and (ii) she believes that (a) (more or less) everyone in the relevant collective (say, those interested in cherishing traditional British customs) intends to drink tea at breakfast and that (b) there is a mutual belief in the collective in question that (a).

---

<sup>35</sup> Instead of groups, I prefer to use socially more neutral terms to avoid accusations of circularity (the ongoing theme of this chapter). In accordance with I.2.3, the process of forming we-attitudes is simultaneously the process of forming a group in the socially significant sense. Hence the notion of a group must not be presupposed in the analysis.

This example points out nicely why Tuomela's intuition about  $x$ 's *reasons* for having the intention in question is so important. The definition, as it stands so far, leaves open the possibility that although  $x$  believes that there is a custom of drinking tea at breakfast, her reasons for intending to do so also herself have nothing whatsoever to do with her belief (or indeed the object the belief, the custom). Surely we would not say that in such a case  $x$  we-intends to participate, *qua* a member of the group of tea-drinkers, in the custom (see Tuomela 2000, 180-181 for a discussion on *correct* reasons for following a social norm). To exclude this possibility, Tuomela (2000, 51-52) represents a *reason relation* with  $/_r$  and, correspondingly, adds the index  $r$  to  $I_{we,x}(G)$  to refer to the intention being a reason-based notion. Thus we get the following (cf. Tuomela 2002, 26):

$$I_{we,x}^r(G) =_{df} I_x(G /_r B_x(I_E(G) \& MB(I_E(G))))).$$

This definition says that  $x$  we-intends  $G$  if and only if  $x$  intends  $G$ , and this is so at least partly for the reason that  $x$  believes that everyone in the relevant collective intends  $G$  and because  $x$  believes that there is a mutual belief in the relevant collective that everyone intends  $G$ . Or to use our example,  $x$  intends to participate in the British custom of drinking tea at breakfast insofar as her beliefs about the British tea-drinkers form (at least partly) the reason why she intends to drink tea at breakfast.

Here my reasoning has departed slightly from Tuomela's line of thought. In contrast to my *subjective* version of  $x$ 's reason-based we-intention, Tuomela thinks that the reason-based notion of  $x$ 's we-intention should be written in the *objective* form, *i.e.*, as follows (Tuomela 2000, 51):

$$I_{we,x}^r(G) =_{df} I_x(G /_r I_E(G) \& MB(I_E(G))).$$

In Tuomela's version it is not  $x$ 's *beliefs* about the intentions of her fellow collective-members that form  $x$ 's reason for her own intention, but the *facts* that (i) the others indeed do intend  $G$  and that (ii) there is a mutual belief that the first fact obtains.<sup>36</sup> In an ideal situation these facts no doubt *are* the reasons, but I believe it is important to pay attention to the subjective version of the reason-based notion of a we-intention.

---

<sup>36</sup> In fact, in Tuomela's definition the scope of  $/_r$  is not perfectly clear; I take it that the mutual belief is meant to be included in the scope. Tuomela (2002, 23 ff.), reformulates this definition in the subjective form. The scope of  $/_r$ , however, remains somewhat unclear in that work too.

The subjective version allows us to do justice to situations where  $x$  sincerely has a we-mode intention but the intention is based on *false* beliefs.  $x$  may be the only member of the collective who is in fact ready to form a collective intention and to engage in collective action in the situation in question. Tuomela almost makes this point himself when he writes that an agent must be required to believe that the objectively stated reason holds, and, according to Tuomela, this “‘subjective channelling’ property is assumed to have been built into the  $/_r$  relation” (Tuomela 2000, 52).

In Tuomela’s view the beliefs seem to be the proximate mechanisms that link the facts as *causes* to  $x$ ’s intention. I do not want to adopt this line of thought for two reasons. First, Part III examines whether the relation here is *causal* at all, and I do not want the present account to be committed either way. Second, I do not think the present analysis should require that the beliefs must be true. In contrast to my subjective version, Tuomela’s notion cannot take into account the possibility that the “subjective channelling” process *misrepresents* the facts behind the beliefs.

In addition to these considerations on subjectivity, there are a few more ambiguities in Tuomela’s discussions that must be clarified before we can be satisfied with the account. Tuomela tells us that his definition requires that the facts about the others are that “the others *we-intend*  $G$  and that there is a mutual belief about this” (Tuomela 2000, 51, my italics). But this is not what Tuomela’s definition in fact says, since the intention of the others is formalised as  $I_E(G)$ , which says that everyone intends  $G$ , not that everyone *we-intends*  $G$ . However, I think that Tuomela’s informal description is closer to what we want to say. Using Tuomela’s formal notation, Tuomela’s informal description of  $x$ ’s reasons for having the intention  $G$  should actually be expressed as follows:

$$I_{we,x}^r(G) =_{df} I_x(G /_r (I_{we,E}(G) \& MB(I_{we,E}(G))))).$$

On the first approximation, this way of writing the definition looks suspiciously circular, since the we-intentionality I aim to define appears both in the *definiens* and in the *definiendum*. To apply the definition to Ruben’s tea-drinking example, this definition appears to say that  $x$  we-intends to participate to the custom of drinking tea at breakfast if and only if the reason why  $x$  intends to drink tea at breakfast is that everybody in the relevant collective we-intends to do so and there is a mutual belief about it.

This would be acceptable if we were willing to restrict the scope of the analysis exclusively to those cases where  $x$  comes from outside a group to *join* it and its pre-

existing institutions. However, if we want, as I think we should, to consider the analysis to reveal something essential about the *construction* of social institutions, this kind of circularity cannot be accepted. The abstract definition should represent the situation from the point of view of any individual belonging to the relevant collective (or even more correctly, of any individual belonging to a set of individuals joining together to form a social group that does not exist prior to this), and then we cannot presuppose the fact about pre-existing we-intentions.

This problem of circularity shows why my subjective way of treating collective intentionality is preferable also for more fundamental reasons than the one discussed above. Perhaps it could be argued that a partial reason why  $x$  intends  $G$  is not that everyone in the collective we-intends  $G$ , but rather that she *takes* the other members of the collective to we-intend  $G$ . Note that if we interpret the definition like this, we have again taken the subjective path, since now it is only in the *definiendum* where the objective *fact* that someone has a we-intention can be found, whereas in the *definiens* we find only  $x$ 's *beliefs* about the others having we-intentions, and not a we-intention as such. We do not want to say that the reason why  $x$  holds the intention  $G$  is that the other members of the collective have the we-intention  $G$ , but rather the (partial) reason is that  $x$  believes that this is what the members of the relevant collective collectively intend. Hence, the formulation that avoids circularity, offers us a more realistic picture and allows us to capture what we intuitively want to say is as follows:

$$I_{we,x}^r(G) =_{df} I_x(G /_r B_x(I_{we,E}(G) \& MB(I_{we,E}(G))))).$$

Of course it could be argued that even if I do not assume we-intentionality actually to appear in the *definiens*, the circularity is nonetheless there, since the *concept* of we-intentionality is mentioned as part of the content of  $x$ 's beliefs that are required for the formation of a we-intention. My answer to this objection is that my analysis is a *factual* or *ontological analysis* and not *semantic*; supposedly  $x$  will have some sort of idea what it is to intend something together (see Tuomela 2000, 75 for similar reasoning), or what it is to be committed to a joint project. We do not have to presuppose that  $x$  actually possesses the concept of we-intentionality, and hence her belief needs not be stated in terms of it. What I am after with the belief about we-intentions is some sort of, possibly tacit, psychological framing of the situation in question as a collective task. This involves seeing also the others as intentional participants in the task (as willing to share in action, as the Gilbertian (H4) put this). It is only that from the perspective of the present

*analysis* of  $x$ 's we-intention it is more convenient to conceptualise the contents of  $x$ 's beliefs in terms of we-intentions. A semantic analysis would be a different matter. I conclude that my subjective version is not circular in any vicious sense.

Unfortunately, the ambiguities in Tuomela's discussion do not end here, since Tuomela goes on to explain that actually the facts about the others are "a partial reason for  $x$  to *we-intend*  $G$ " (Tuomela 2000, 52, my italics). It seems that even my subjective version of Tuomela's revised objective definition does not amount to this. However, once again Tuomela's informal description captures what we really *do* want to say better than the formalisations we have seen so far.  $X$ 's belief that  $G$  is what the relevant collective intends collectively may be a partial reason why  $x$  intends  $G$ , although  $x$  may have some completely different reasons for that too. Be that as it may, surely  $x$ 's belief that the relevant collective we-intends  $G$  is a partial reason why  $x$  adopts the we-intention  $G$ , regardless of her reasons for individually intending  $G$ .

The formalisation corresponding to Tuomela's informal descriptions of we-intentionality so far, should, I think, be the following:

$$I_{we,x}^r(G) =_{df} I_{we,x}(G /_r (I_{we,E}(G) \& MB(I_{we,E}(G)))).$$

This can be again transformed to the preferable subjective form as follows:

$$I_{we,x}^r(G) =_{df} I_{we,x}(G /_r B_x(I_{we,E}(G) \& MB(I_{we,E}(G)))).$$

Moreover, if we add – as I indeed think we should – the point I made about  $x$ 's simultaneous individual and collective intentions we get a definition of the following form:

$$I_{we,x}^r(G) =_{df} (I_x(G) \& I_{we,x}(G)) /_r B_x(I_{we,E}(G) \& MB(I_{we,E}(G))).$$

This definition says that  $x$ , a member of a collective  $c$ , we-intends  $G$  if and only if

- (i)  $x$  intends  $G$
- (ii)  $x$  we-intends  $G$
- (iii) (i) and (ii) at least partly for the reason that  $x$  believes that (a) everyone in  $c$  we-intends  $G$  and that (b) there is a mutual belief in  $c$  that (a).

However, by smuggling the factual notion of we-intentionality to reappear in the *definiens*, clause (ii) seems to force us to face the problem of circularity all over again.



Fortunately we can – for the problem is now beginning to be all too familiar – easily envision the cure to this problem: perhaps taking a more subjective path could once more save us from the problem of circularity. Undeniably, clause (ii) should not require that  $x$  literally speaking we-intends  $G$ , but rather that  $x$  sees herself as participating to what she thinks her fellow collective members are doing (in terminology of II.3.3,  $x$  identifies herself with the collective and their activities).

This, I think, amounts to saying that  $x$  is willing to share in action with her fellow collective members, or in other words, that  $x$  is willing to join her social partners to do  $G$  together. I will represent this formally by writing that  $x$  believes that she is we-intending  $G$  (i.e.,  $B_x(I_{we,x}(G))$ ), although what I have just said shows that this is not strictly speaking the ideal way of formalising the thought behind the new clause (ii). Firstly, the concept of we-intentionality appears again in the *definiens*, which is unpleasant although not fatal for a factual analysis (recall my comments above). Secondly, as was mentioned, what I take  $x$  to believe is not really that she is we-intending  $G$  but that there is a joint activity for her to participate in, and this contributes to the construction of her we-intention. (Both the first and the second point apply naturally also to  $x$ 's belief about everyone in group we-intending  $G$  that turns up in  $x$ 's reason for intending  $G$ .) But the definition is already complicated enough without introducing new symbols, so I will stick to this formalisation.

With these reservations, my penultimate definition of the reason-based notion of  $x$ 's we-intending  $G$  is as follows):

$$I'_{we,x}(G) =_{df} (I_x(G) \ \& \ B_x(I_{we,x}(G))) /_r B_x(I_{we,E}(G) \ \& \ MB(I_{we,E}(G))).$$

This definition should be read as saying that that  $x$ , a member of a collective  $c$ , we-intends  $G$  if and only if

- (i)  $x$  intends  $G$
- (ii)  $x$  is willing to share in action with other members of  $c$  in the situation in question
- (iii) (i) and (ii) at least partly for the reason that  $x$  believes that (a) everyone in  $c$  we-intends  $G$  and that (b) there is a mutual belief in  $c$  that (a).

Or to use Ruben's example,  $x$ 's we-intention to participate in the British custom to drink tea at breakfast amounts to the following:

- (i)  $x$  intends to drink tea at breakfast

- (ii)  $x$  sees herself as a member of the relevant collective (say the people who value cherishing traditional British customs) and  $x$  believes that her drinking tea at breakfast amounts to joining a prevailing custom of the collective in question
- (iii) (i) and (ii) at least partly for the reason that  $x$  believes that (a) everyone in the collective sees themselves as (ii) describes  $x$  to see herself and that (b) there is a mutual belief in the collective that (a).

This definition represents the situation from the perspective of a single member of the collective (*i.e.*,  $x$ ), whose beliefs about the situation could be radically mistaken. If this were the case,  $x$  would have a *we*-intention but, despite her sincere intention, there would not be a social project to which  $x$  could participate. However, when the requirements of the clauses (i)-(iii) are in fact fulfilled (even if the beliefs are not true), it is an objective fact that  $x$  *we*-intends to  $G$ . And, moreover, if this objective fact is true of all the members of  $c$  (if  $x$ 's beliefs are true), there is indeed a mutual belief about the intentions of the group-members, and moreover, the mutual belief is also a true mutual belief. Hence, in such a case the members of  $c$  indeed *we*-intend  $G$  jointly together, fully and objectively. Thus, I have developed Tuomela's account to show how the collective intentions of individuals can form a plural subject, *i.e.*, how we can get from the subjective perspectives of individual agents to an objectively existing, shared *we*-intention.

Hence, also in Tuomela's framework we can adopt the general perspective and say that agents  $x_1, \dots, x_n$  that form the collective  $c$  *we*-intend  $G$  together exactly when the criteria for  $I_{we,x}(G)$  are true for each  $x_1, \dots, x_n$ . As can easily be seen, this amounts *almost* to the same view we arrived at when the starting point was the general perspective rather than the perspective of a single group-member, namely to (H4) that I argued to be the best way to understand Gilbert's (1989) discussion. Although I do believe that the definition I have constructed starting from Tuomela's discussion is basically correct as it stands, the fact that I need to say that it almost justifies (H4) suggests that one more clarification must still be done.

What I have in mind is the distinction between individual and collective intentions. When discussing Gilbert's account I argued that it is an important merit of her theory that it allows people to adopt a collective intention even if individually speaking they do not consider it to be such a good idea in the first place. My analyses in this section seem to contradict this, since clause (i) in the informal analysis above requires that in order for  $x$  to collectively *we*-intend  $G$ , she must also individually intend  $G$ .

Thus, it seems that as it stands (i) fails to capture what we really want to say. Instead of requiring  $x$  to individually intend  $G$  we should say that when an agent  $x$  we-intends to treat the realisation of  $G$  as the collective goal of her group (when  $x$  we-intends  $G$ ),<sup>37</sup>  $x$  must also intend *to do her part* in the collective task. In the case of Gilbert and Searle this was expressed by saying that when agents  $A_1, \dots, A_n$  collectively accept  $X$  as their collective intention, their individual roles in the collective task of realising  $X$ , namely individual goals and intentions  $X^1, \dots, X^n$  are *derived* from the collective intention  $X$ . In light of this, it seems that I need to do one more final modification to the definition. Let  $g$  stand for an individual intention derived from  $G$  in the above sense, and  $x$ 's reason-based we-intention to  $G$  can be analysed as follows:

$$I_{we,x}^r(G) =_{df} (I_x(g) \ \& \ B_x(I_{we,x}(G))) /_r B_x(I_{we,E}(G) \ \& \ MB(I_{we,E}(G))).$$

As should be clear by now, this definition says that  $x$ , a member of a collective  $c$ , we-intends  $G$  if and only if

- (i)  $x$  intends to do her individual part in realising  $G$  (*i.e.*,  $x$  intends  $g$ )
- (ii)  $x$  is willing to share in action with other members of  $c$  in the situation in question
- (iii) (i) and (ii) at least partly for the reason that  $x$  believes that (a) everyone in  $c$  we-intends  $G$  and that (b) there is a mutual belief in  $c$  that (a).

When  $x$  has a we-intention her reasons for this must be largely social, although – as the tea example illustrates – she can of course have also some individual reasons for intending to perform the very same action, *i.e.*, she may have also an individual intention  $g$ .

The distinction between intentions that stem from collective-mode considerations and intentions that stem from individual-mode considerations (regardless of whether these intentions coincide or not) will be of utmost importance in Part II, and hence it is crucial to observe here that the present analysis can easily include this feature. Similarly, the final modification of the definition also eliminated the very last instance of an individual intention  $G$  (*i.e.*, the term  $I_x(G)$ ) from the definition. This highlights further the fact that individual intentionality and collective intentionality are not only distinct, but also fully independent of each other. In this chapter this independence has been used for establishing the irreducibility of collective intentions to individual in-

---

<sup>37</sup> For the present purposes we can use “goal” and “intention” interchangeably since we are talking about intentions to realise some goal.

tentions. In Part II this independence will play at least as crucial a role in my arguments on how agents capable for we-mode action can resolve social dilemma situations.

The final modification allows us also to hint at the sources of what Tuomela (2000, 80) calls the “quasi-moral” nature of social institutions, we-intentionality and we-mode action. In I.2.1 we saw how normativity is indeed crucial for the present theory of social reality, since social facts, entities and institutions were seen ontologically speaking to boil down to ontologically subjective but epistemically objective norms of appropriate action. The key, I think, is the fact that an agent who has adopted the we-mode sees individual tasks as *derived* from the collective-goal, which in turn is originally formed on the basis of what is the *rational* course of action for the collective to take. Hence the collective goal has a normative status in virtue of being (taken to be) collectively optimal, and the normativity flows down to individual roles in the collective task in virtue of the norms of rationality governing derivation.

I of course realise that at this point this sounds rather unsatisfactory. However, Part II addresses the collective optimality of collective goals in detail, and Part III the normativity of the rules of rationality. Thus the intrinsic and irreducible normativity of social reality and social practices cannot be properly seen until in the end of this dissertation. However, already at this point it is clear that social reality consists of essentially normative requirements and inferences.

What I have said here already suffices for seeing why Tuomela’s (and Gilbert’s) way to account for the quasi-moral right to expect that the others will do their part in the collective task in terms of the *communication of acceptance* (of the collective goal) condition and the *knowledge of other participants’ acceptance* condition (Tuomela 2000, 29), or the requirement that there must be mutual *knowledge* concerning the acceptance resulting from communication between the agents (Gilbert), is *not* the road I want to take. The conditions may be of great practical relevance when agents seek to sort out concrete problems (for example, in the case of a free-rider it is relevant to point out that the free-rider had openly communicated her intention to participate in the collective task). However, the relevance is due to the fact that, say, the communication of acceptance is almost universally assigned a normative status, and thus the conditions mentioned by Gilbert and Tuomela should not be built into the analysis, since ultimately (ontologically!) they simply beg the question. However, we must wait for Parts II and III before this can be seen properly.

Note also that the present analysis should be understood as compatible with Gilbert’s view (discussed above) that when a collection of agents actually forms a we-

intention, they form a plural subject of collective action, which, moreover, is equivalent to saying that the collection of agents in question forms a social group. On the first approximation it might seem that my informal explications of the formal definition of  $I_{we,x}^r(G)$  contradict this idea, since I have included in the *definiendum* the statement “*x, a member of a collective c*”. However, the expression should here be understood in a socially insignificant way, *i.e.*, as saying merely that *x* belongs to some set *c* of individuals. The set or the collective *c* becomes a socially significant group only when (or if) the members of the set *c* form the relevant we-intentions, just as was argued in the context of Gilbert’s plural subjects.

The reason why I have used the expression “collective” here instead of talking about individuals  $x_1, \dots, x_n$  (or something similar) is that the perspective here has all the time been that of a single agent. I have wanted to do justice both to the idea that the analysis should apply to cases where *x* joins an already-existing social group – even if she does not know all the members of the group (*e.g.*, the tea-drinking case) – and to cases where agents form a completely new group. Even in cases of the latter kind *x* needs to have an idea that there is a relevant collection (*i.e.*, *c*) of agents, although the collection will form a social group only insofar as they each adopt a we-intention and create a plural subject together. Hence the definition of a we-intention in the sense of  $I_{we,x}^r(G)$  does not undermine the role of collective intentionality as a building block of social reality, including social groups.

## CHAPTER 1.3: CONCLUSION

So what kind of picture of social entities and social facts emerges from the considerations in this Part of my study? To put it shortly, I have accepted Searle's view that the core of social reality is ultimately to be found, not in social entities or facts as such, but in collectively accepted and required patterns of behaviour. What is characteristically *social* in social entities is a collectively assigned status. As I have argued, to avoid circularity the status must be assumed ontologically speaking to consist in collectively accepted and required forms of behaviour, *i.e.*, a social practice. Ultimately, when we collectively construct social facts via collective acceptance, the acceptance does not construct its own object as a new solid entity, but rather a consensus that certain actions are to be performed in certain circumstances (cf. Collin 1997, 196-197).<sup>38</sup>

This, in effect, is one way of highlighting the difference between the methodologically holistic nature of the view I defend and the methodological individualism of, for example, mainstream rational choice account of social facts. In my view social facts are essentially normative practices based on collective we-mode attitudes that place normative requirements for those who wish to participate in the practices (be one of us). This holds both in the sense that social facts are ultimately status functions that consist of rules of appropriate behaviour and in the sense that full participation in social practices requires one to adopt the we-mode perspective such that one's individual action-intentions are derived from the collective task. Consequently, in truly social action individual performances are to be rationalised in terms of their role in the collective task and not in terms of individual-mode private goals (of course, however, individual-mode free riders can exploit the we-mode of others – and the practices they maintain – to their own advantage, as long as there are sufficiently many we-moders to maintain the practices).

Methodologically individualistic views, in contrast, tend to see social facts as non-normative aggregates of actual choices and expectations of individuals that essentially asocial individuals take into account when contemplating how to maximise their utility functions. This contrast is addressed in detail in Part II; here it suffices to note

---

<sup>38</sup> Searle (1995) is also very explicit concerning these issues: “it is tempting to think of *social objects* as independently existing entities on analogy with the objects studied by the natural sciences”, but in truth “[s]ocial objects are always [...] constituted by social acts; [...] *the object is just the continuous possibility of the activity*” (Searle 1995, 36) – “hence, our interest is not in the object but in the processes and events where the functions are manifested” (Searle 1995, 57). “What we think of as *social objects* [...] are in fact just placeholders for patterns of *activities*” (Searle 1995, 57).

that the emphasis my account puts on normativity and dynamic practices could form a sufficient reason for avoiding the terminology of facts altogether.<sup>39</sup>

As we saw in I.2, this applies fully also to the standard example of a social fact, namely that the pieces of paper in my pocket are money. It is easy to lose sight of this essentially dynamic feature of social facts for two reasons. First, in our everyday speech we rather misleadingly tend to classify social facts as independent facts ontologically on a par with brute natural facts, whereas in reality they are merely external and objective (I.1). Externality and objectivity is enough for social facts – and the practices they ultimately consist of – to be *things* in the technical sense of Durkheim, which is the second reason why the dynamic nature of social facts is easy to forget.

Although there are different ways of using the notion “social fact” (I.1.5), the crucial question must be the nature of social action that underlies social facts. In this part I have been mainly interested in approaches that challenge methodological individualism. Suppose there is a country where everybody simply drives on the left-hand side of roads. Is the prevalence of left-hand side traffic a social fact? In a sense it obviously is, for the fact involves a number of agents. However, in the Weberian tradition the left-hand side traffic is a social fact only if the actions the fact consists in are *social* actions: “Action is social in so far as, by virtue of the subjective meaning attached to it by the acting individual (or individuals), it takes account of the behaviour of others and is thereby oriented in its course” (Weber 1922/1947, 88). In other words, if the reasons why people in fact drive on the left do not mention the behaviour of others, the resulting practice does not constitute a social fact.

In this part, however, I have aimed to analyse social facts in even a stronger sense, according to which *strategic* individual action (which satisfies Weber’s criterion) does not suffice for truly social action and practices that social facts consist in. Following mainly Gilbert, Searle and Tuomela I have concentrated on social facts in the strong anti-individualistic sense that involves collective we-mode intentionality. To anticipate Part II, social facts in this sense require that agents act *together* instead of simply playing a coordination-requiring social game qua individual players.

This strong notion of togetherness that goes beyond methodological individualism is meant to capture social facts qua social practices in the core sense in which the reason for following the practice is not any individual-mode evaluation of costs and benefits, but rather the practice itself, or at least cost-benefit analyses at the collective

---

<sup>39</sup> This point plays a crucial role in III.3.3 and III.4.1, where I analyse Kripke’s notion of social practices that are not fact-based but rather normative practice-based. I argue that such normative practices can be fact-based in my sense of a social fact, although not in the sense of a methodological individualist.

level, *i.e.*, in the we-mode. This is what my development of Tuomela's notion of a reason-based we-intention was meant to capture. Social facts in the core sense consist of social practices in this strong sense which, in turn, requires collective we-mode intentionality such that it is participation in the practice that motivates individual action-intentions: In truly social action individual action-intentions are derived from collective-level we-mode considerations. This is the exact opposite of standard methodological individualism (such as the picture implied by rational choice approaches to social action), where compliance with a practice is motivated by individual-mode beliefs and desires. This is the well-known distinction between *Homo Sociologicus* and *Homo Economicus*.

Admittedly I have not provided ultimately satisfactory arguments as to why these strongly social notions are required for the core forms of social action. Thus far I have simply appealed to intuitive examples of social activities that are meant to show that an analysis that does not go beyond Weber's strategic interaction to full-blown collective intentionality (or plural subject) cannot suffice for many familiar social situations. Personally, I am satisfied with such appeal to intuitions. However, general, conceptual arguments in favour of the theory of collective intentionality are to be found in Part II.

Philosophers inclined to favour even more holistic approaches in the full *ontological* sense of rejecting (*OI*) (although more often than not such ontological anti-naturalism is obscured with an appeal to unanalysable "emergence") sometimes argue that the kind of constructivist view I am defending on the basis of Gilbert, Searle and Tuomela cannot capture the externality of social facts in a sufficiently strong sense (see, *e.g.*, Collin 1997 & Niiniluoto 1999). The basic form of this accusation is simply the claim that an account building on collective acceptance cannot capture social reality, for many of the most basic social facts are essentially more durable than the *de facto* existing attitudes of acceptance (Collin 1997, 206).

For example, it is conceivable that the social entity European Union outlasts all the individuals living at the moment. I think that this, simple as it may sound, is an important point – although not fatal for the present account. The point highlights nicely the importance of the strong *methodological* anti-individualism inherent in the present theory and the unacceptability of summative accounts. The social entity in question – here, the European Union – must be identified with the *practice* of normative requirements created by collective attitudes, not the *attitude-tokens* of individuals themselves. Ruben's example of the British institution of drinking tea at breakfast is also helpful



here. The institution consists of a practice constituted by collective we-attitudes, and my analyses of the formation of a reason-based we-intention in 1.2.4 explained explicitly how new individuals can join the practice, allowing thus other participants to leave the practice.

A stronger form of the accusation that mere externality (in the sense of intersubjectivity) cannot give sufficiently robust status to social facts can be made in terms of the threat of social relativism. As Collin puts this point, it is difficult to see how a Searlean analysis (including, I take it, the present analysis) could “provide for the distinction between a *correct* and an *incorrect* way to proceed within a convention [a practice]” (Collin 1997, 206).<sup>40</sup> A collective acceptance view can describe how a collective reproduces its practices, but how could it leave room for addressing the correctness of the practice, for it is the collective agreement concerning how to proceed that constitutes the very practice? It seems that the whatever the collective *takes* to be the correct way to reproduce the practice must *be* the correct way, in which case there is no distinction between correctness and incorrectness.

The paradigmatic example here, appealed to by both Collin and Niiniluoto, is the law. As Collin puts it, ontologically individualistic, constructivist views (known as *legal realism* in the philosophy of law) seem to lean towards the view that “the law is what Supreme Court judges would actually decide” (Collin 1997, 207), for in our legal practices Supreme Court judges have the final word on what forms of behaviour are accepted and required in certain situations. However, this is problematic:

In declaring that decisions of Supreme Court judges determine what the law is, legal realists overlook the fact that judges, including those of the Supreme Court, try to *conform* to the law and to *discover* what the rules actually dictate. The judges treat the rules as norms to be adhered to and do not consider themselves at liberty to *create* law in passing sentence [...]. Indeed, the realist construal renders it impossible to make sense of what Supreme Court judges are doing when they enter into a subtle arguments to decide a complicated case. Their deliberations are clearly not attempts to *predict* what they will themselves decide in the case in question, or else judges would be chasing their own coat tails forever.

A proper understanding of the judicial process must allow for a distinction between what the judges happen to decide and the true content of the law.  
(Collin 1997, 207.)

This is a noteworthy problem for all constructivist accounts of social facts. As I have argued (I.1.4 in particular), externality suffices for epistemic objectivity *from the point of any individual taken singly*. But if social facts are not independent facts (*i.e.*, if we

---

<sup>40</sup> Different aspects of Collin’s worry are addressed in detail in III.3 onwards.

subscribe to (*OI*), how can there be conceptual room for epistemic objectivity from the point of view of the whole community involved? It seems that a community cannot be collectively mistaken about social practices – there seems to be no distinction between “what the judges happen to decide and the true content of the law”.

This problem is so serious that both Collin and Niiniluoto see no other option than to revert to Karl Popper’s weak Platonism, where social facts are said to belong the Platonist, objective World 3, the inhabitants of which are in an important sense independent of human attitudes (that belong to the subjective World 2 of contentful psychology), although originally created by such attitudes. The view is supposed to be ontologically monistic, since World 2 is said to *emerge* from the physical World 1, and World 3 from Worlds 1 and 2. However, in my mind this cannot amount to a serious solution. The notion of emergence remains completely unaccounted for,<sup>41</sup> and therefore the view seems to hold confusingly that social facts are independent without being independent. This is hardly satisfactory.

Although I do not consider this kind of emergentism to be a serious alternative, I accept that my view must be able to answer Collin and Niiniluoto’s criticism. The first point to notice is that it is not at all obvious that social relativism – a collective being always right – is unacceptable when it comes to *social* facts. As Robert Brandom (1994, 53) points out, “[w]hatever the Kwakiutl treat as an appropriate greeting gesture for their tribe [...] is one; it makes no sense to suppose that they could be collectively wrong about this sort of thing”. Any individual member of the tribe can of course be mistaken about appropriate greeting gestures, but not the tribe collectively. Perhaps we could say that the situation described by Collin is simply a complicated form of this kind of situation.

The main difference between the Kwakiutl greeting practices and the law is that the law is not a practice unconnected to other practices. On the contrary, the law is explicitly tied to the rules of logic and rationality, prevailing norms and customs and so on, and the last word on these issues is not given to Supreme Court judges. Rather, our practices dictate that the Supreme Court is to have the last word on what shall be done in legal controversies (the Supreme Court is assigned a certain status function), but as explicitly our practices acknowledge that the law is connected to other social practices (many of which are governed by implicit, tacit norms) – ultimately, to the practices governing the meaning of words. Thus, *pace* Collin, I think the constructivist view does

---

<sup>41</sup> In particular, Collin’s (1997, 209) appeal to “construction by idealisation” as opposed to “construction by convention” does not help, for he gives no explanation what construction by idealisation behind Popper’s World 3 amounts to.

have conceptual room for the distinction between what the law really dictates and what the Supreme Court will decide.

The true content of the law is what really follows from our commitments to laws, norms of rationality, rules governing the use of words and so on – and the Supreme Court is not the last authority when it comes to, say, rules of logic. Moreover, since the Supreme Court cannot hope to make all these commitments explicit in any single case, we explicitly acknowledge the possibility that the Supreme Court may be mistaken. Our practices do not require that the Supreme Court is infallible. After all, the Supreme Court can be corrected afterwards – the practice of interpreting the law is open-ended (III.4 onwards). Rather, our practices say that for practical purposes we will let the Supreme Court to have the last word. However, because of their dependence on other practices, also legal practices are in principle open for criticism. This is one more reason why methodological holism is to be preferred to individualistic views. As I explained above, it is the methodologically individualistic conception of social facts that identifies the facts with actual choices of individuals, whereas a holistic picture leaves room for identifying social facts with open, normative practices (Appendix). Collin's problem is a problem for methodological individualism (including what I in III.3.3 call the naïve communitarian view), not for methodological holism.

I believe this answer suffices as an answer to Collin. The case of Niiniluoto (1999), however, is more complicated. Niiniluoto argues that to defend the kind of openness to criticism that is required to save the constructivist view, a constructivist must, just as I have done above, ultimately appeal to the rules governing forms of acceptable inferences and acceptable linguistic assertions. But, says Niiniluoto, then the constructivist cannot anymore build the argumentation on the point that social relativism regarding such rules is harmless. Surely we want to say that the collective can be collectively mistaken about, say, the principles of logic and mathematics or whether or not it is true to say that there is a cat on the mat. At least *these* practices must be objective in a more independent sense than intersubjective externality and hence, according to Niiniluoto, we must accept Popper's postulation of the emergent, independent World 3. Niiniluoto presents this line of thought as a criticism of social theories of meaning and language (especially the so-called *meaning finitism*): In Niiniluoto's mind the only way to make sense of the community being wrong about, for example, forms of deductive reasoning is to say that the true laws of logic exist independently in World 3, and our *representations* of them can be true or false.

This, I admit, is a serious criticism. Moreover, it takes us so deep into the philosophy of language that I cannot answer it satisfactorily here. However, Part III of this dissertation can be seen as an answer to Niiniluoto's criticism. For example, in III.3.3 I discuss – and reject – precisely the kind of social relativism (under the label “naïve communitarian view”) Niiniluoto has in mind and argue that my constructivism is not committed to anything like it. Moreover, I argue that actually it is Niiniluoto's alternative, the idea that rule-governed practices ought to be understood in terms of rules independent of the practices, that leads to grave problems. The details of this argumentation, however, must wait until Part III and Appendix.

PART II:  
THE EVOLUTION AND PSYCHOLOGY OF  
*HOMO SOCIOLOGICUS*

## INTRODUCTION TO PART II

In Summer Term 2001, Daniel C. Dennett gave a series of lectures and seminars at the London School of Economics on the topic of the evolution of human freedom (published as Dennett 2003). During those seminar meetings, Dennett once illustrated the basic dynamics of natural selection by telling the following story: Once upon a time two friends were camping in the wilderness of Alaska. One evening when the campers were relaxing by the campfire, they suddenly saw a huge bear approaching them. Naturally, they panicked. The first camper started very quickly to put his track shoes on. “Don’t be stupid”, cried the other camper, “surely you can’t outrun a bear even with your running shoes on!” The first camper, however, was not disturbed by the comment of his companion but just kept on tying his shoes: “But my dear friend, I don’t need to outrun the bear. It’s perfectly enough if I run faster than *you!*”<sup>42</sup>

Dennett’s story illuminates aptly a feature that seems to be essential part of evolutionary dynamics, namely that the tendency to take care of one’s own survival at the expense of helping others is a behavioural trait evolution never fails to favour. Natural selection appears to promote egoistic and self-centred behaviour so strongly that we should expect non-calculating solidarity to be selected away rather soon. Defection pays better than philanthropy in the struggle for the survival of the fittest. In short, evolutionary theory seems to dictate that all apparently altruistic or social behaviour must, ultimately, be based on mechanisms which guarantee that the behaviour will on average lead to the best possible outcome from the individual perspective of a single, egoistic agent, since otherwise such forms of behaviour would have been selected away. As Peter Singer sums this up, “for a Darwinian there is a problem in assuming that individuals behave altruistically for the sake of a larger group” (Singer 1999, 20).

Contrast this with the picture of social action painted by my analysis of collective we-mode action in Part I. I argued that the theory of collective intentionality holds that the essence of social action is such that the possible outcomes of the combined actions of individuals are assessed in collectivistic terms of considering what is optimal for the whole group, and individuals then derive their individual tasks from the collective level considerations. Consequently, the actions of individuals may well be sub-optimal if assessed from the individualistic perspective, albeit optimal in the task of re-

---

<sup>42</sup> John Dupré (2001, footnote 19 on p. 43 & 2002, 198) tells a version of the same anecdote. The moral Dupré draws from the story is, however, somewhat different from my treatment.

alising the collective goal, *i.e.*, precisely of the kind Singer says a Darwinian has a problem with.

Should it, then, be concluded that the argumentation in favour of irreducible collective attitudes and action as presented in Part I is based on wishful and idealistic thinking – on a refusal to take seriously the teachings of evolutionary biology? Does the fact that human beings form a species that has evolved via natural selection imply that Searle must be fundamentally mistaken when he states that “the selectional advantage of cooperative behavior [based on collective intentionality] is [...] obvious” (Searle 1995, 38)? Or, to put it in other words, is a theory that operates with plural subjects in Gilbert’s (1989) sense actually evolutionarily impossible? I think the answer must be “no”; we-mode co-operation is not only a possible outcome of evolutionary process, but also quite likely to have evolved (although, as I argue before, the selectional advantages of we-mode co-operation are far from obvious but require a rather lengthy defence).

This Part aims to show how evolutionary dynamics are perfectly compatible with the strong thesis of human sociality I am analysing and defending in this dissertation. I will start by following Elliott Sober and David Sloan Wilson’s discussion on the evolution of human altruism as presented in their book *Unto Others: The Evolution and Psychology of Unselfish Behavior* (Sober & Wilson 1998). My aim is to use Sober and Wilson’s theory to show, *pace* Singer, how respectable evolutionary biology can fit in the same picture with the strongly collectivistic views on human sociality I defend.

Moreover, I will also show how independent social ontology and philosophy of social action, building on collective we-mode attitudes – rather than altruism (which is an individual-mode attitude, albeit directed towards others) – as an ultimate motivational mechanism behind social behaviour, can also contribute to the development of Sober and Wilson’s theory. Thus, what is at stake in this Part of my study is not merely the question of social solidarity versus self-centred egoism. Rather, the underlying issue is again the debate between (methodological) individualism and holism: are uncontested evolutionary facts compatible with and supportive of the kind of holistic sociality the theory of collective intentionality subscribes to? The main argument of Part II is that not only is the answer to that question positive; moreover, the evolutionary considerations provide a very strong, general argument in favour of we-mode intentionality.

I start by examining whether evolution can favour forms of behaviour that are fitness-maximising from the perspective of the group an agent belongs to but sub-optimal from the perspective of the personal fitness of the agent. Such behaviour would be *evolutionarily altruistic*.

## CHAPTER II.1: EVOLUTIONARY EGOISM AND ALTRUISM

When we talk about egoism and altruism we usually have in mind psychological theories of motivation. An action – even if it is an intentional act of helping others – is *psychologically egoistic* if it is ultimately rationalised in terms of the expected welfare (or pleasure) of the agent performing the action. A psychological egoist may value the psychological satisfaction and social esteem she gets from helping others so highly that she is willing to pay the price (putting herself in danger, not maximising her financial benefits, or something similar) of helping others. However, in such a case the welfare of others is not the ultimate goal of the agent, and thus the action does not count as psychologically altruistic. In contrast, the *evolutionary* concepts of egoism and altruism do not involve intentional explanations of action at all. The evolutionary concepts concern solely the effects a given behaviour has on survival and reproduction (*i.e.*, the fitness of the agent). As Sober and Wilson put it, “[i]ndividuals who increase the fitness of others at the expense of their own fitness are (evolutionary) altruists, regardless of how, or even whether, they think or feel about the action” (Sober & Wilson 1998, 6). In this section I will concentrate on the evolutionary notions of egoism and altruism, for behaviour associated with collective intentionality are unquestionably evolutionarily altruistic. The psychological question is addressed in Chapter II.2 onwards.

### II.1.1 GROUP SELECTION

To highlight the evolutionary problem of altruism, let us look at the following example (adopted from Sober & Wilson 1998, 19-21) of a model that shows why we should not expect altruistic behaviour to evolve. To make the model simpler, let us first assume (i) asexual reproduction and (ii) that altruistic behaviour influences only reproduction. Thus, in what follows the number of offspring is a sufficient measure of fitness, although normally fitness is understood to include both the individual’s ability to survive and its ability to reproduce.

Consider a population of  $n$  individuals such that there are two genetically encoded traits, altruism ( $A$ ) that occurs with frequency  $p$  and selfishness ( $S$ ) occurring with frequency  $(1 - p)$ . Thus the number of altruists in the group is  $np$  and the number of non-altruists  $n(1 - p)$ . Moreover, let the average number of offspring of each individual in the absence of altruistic behaviour be  $X$ , and the influence of altruism be such that an



individual behaving altruistically will cause itself to have  $c$  fewer offspring and a single other member of the group to have  $b$  more offspring.

Hence, each altruist will experience the cost of its own altruistic behaviour ( $-c$ ), but each altruist may also receive the benefit  $b$  as the result of altruistic behaviour of some other altruistic member of the group. However, the number of possible benefactors for each altruist is  $np - 1$ , whereas a non-altruist can receive benefits from  $np$  individuals. Since the altruists are dispensing their benefits among  $(n - 1)$  individuals, we get the following average fitnesses of each altruist ( $W_A$ ) and non-altruist ( $W_S$ ):

$$W_A = X - c + [b(np - 1) / (n - 1)]$$

$$W_S = X + [bnp / (n - 1)]$$

Obviously,  $W_A < W_S$ . Hence, altruism will be selected against within this population.

To make the result even clearer, let us look at a particular example (also from Sober & Wilson 1998) where the parameters of the model get the following values:

$$n = 100$$

$$p = 0.5$$

$$X = 10$$

$$b = 5$$

$$c = 1$$

With these values the average fitnesses can be calculated as follows:

$$W_A = 10 - 1 + [5(100 \cdot 0.5 - 1) / (100 - 1)] = 9 + 245 / 99 \approx 11.47$$

$$W_S = 10 + [(5 \cdot 100 \cdot 0.5) / (100 - 1)] = 10 + 250 / 99 \approx 12.53$$

Hence, the presence of altruism increases the average fitnesses of all the members of the population. Nevertheless, non-altruists benefit more than altruists, since non-altruists do not have to pay the price  $c$ . The total number of offspring ( $n'$ ) is

$$n' = npW_A + n(1 - p)W_S = (100 \cdot 0.5 \cdot 11.47) + (100(1 - 0.5) \cdot 12.53) = 573.5 + 626.5 = 1200.$$

Thus, the frequency of altruists among the offspring ( $p'$ ) is

$$p' = npW_A / n' = (100 \cdot 0.5 \cdot 11.47) / 1200 \approx 0.478.$$

Obviously,  $p' < p$ , which is equivalent to saying that there is a selection force acting against altruism. Moreover, since the population cannot grow to infinity, let us assume that mortality returns the population to the size  $n = 100$ . Since it was assumed that altruism does not affect survival, reduction of the population does not alter the frequency  $p'$  of altruists and, hence, if the procedure is repeated over many generations, the altruists will continue to decline in frequency and ultimately non-altruism will go into fixation. In other words, altruism will be selected away.

This model seems to suggest that altruism indeed cannot evolve, because natural selection never fails to promote evolutionary egoism. Note that since at this point we are talking about evolutionary concepts of altruism and egoism, it is indeed *true by definition* that natural selection always favours egoism, since egoism was defined as behaviour that maximises the fitness of the actor. Still, Sober and Wilson insist, the conclusion does not follow: “On the contrary, it is easy to show that altruism can evolve when more than one group is present” (Sober & Wilson 1998, 23). To see how Sober and Wilson seek to justify their claim, let us look at the simple model of the evolution of altruism Sober and Wilson provide as an example (Sober & Wilson 1998, 23-26).

Using the symbols introduced above, let us look at a population where

$$n = 200$$

$$p = 0.5$$

$$X = 10$$

$$b = 5$$

$$c = 1$$

Let us this time assume that before reproduction takes place, the global population of size  $n$  divides into two separated sub-groups of sizes  $n_1$  and  $n_2$  such that  $n_1 = n_2 = 100$ . Further, let us also assume that in this division altruists and non-altruists tend to associate with other altruists and non-altruists, respectively, such that  $p_1 = 0.2$  and  $p_2 = 0.8$ . Thus, when the individuals in the sub-groups (let us call them Group 1 and Group 2) reproduce, they produce the following outcomes:

Group 1

$$n_1 = 100$$

$$p_1 = 0.2$$

$$W_{A1} = 10 - 1 + 5(100 \cdot 0.2 - 1) / (100 - 1) = 9 + 95 / 99 \approx 9.96$$

$$W_{S1} = 10 + (5 \cdot 100 \cdot 0.2) / (100 - 1) = 10 + 100 / 99 \approx 11.01$$

$$n_1' = (100 \cdot 0.2 \cdot 9.96) + [100(1 - 0.2) \cdot 11.01] = 199.2 + 880.8 = 1080$$

$$p_1' = (100 \cdot 0.2 \cdot 9.96) / 1080 \approx 0.184$$

Group 2

$$n_2 = 100$$

$$p_2 = 0.8$$

$$W_{A2} = 10 - 1 + 5(100 \cdot 0.8 - 1) / (100 - 1) = 9 + 395 / 99 \approx 12.99$$

$$W_{S2} = 10 + (5 \cdot 100 \cdot 0.8) / (100 - 1) = 10 + 400 / 99 \approx 14.04$$

$$n_2' = (100 \cdot 0.8 \cdot 12.99) + [100(1 - 0.8) \cdot 14.04] = 1039.2 + 280.8 = 1320$$

$$p_2' = (100 \cdot 0.8 \cdot 12.99) / 1320 \approx 0.787$$

And in the global population of both groups taken together ( $n = 200$ ),

$$n' = n_1' + n_2' = 1080 + 1320 = 2400$$

$$p' = (p_1' \cdot n_1' + p_2' \cdot n_2') / n' = (0.184 \cdot 1080 + 0.787 \cdot 1320) / 2400 \approx 0.516$$

Note that the sub-group with higher frequency of altruists grows larger than the other group ( $n_2' > n_1'$ ), resulting altruists to increase in frequency in the global population ( $p' > p$ ). This is not, however, sufficient for showing how altruism can evolve, since both sub-groups behave just like the group in the one-group example: within each group, there is a selection force acting against altruism ( $p_1' < p_1$  and  $p_2' < p_2$ ), and hence natural selection will eliminate the altruists in both groups, just as was the case in the first example. The growth in the global frequency of altruists will turn out to be a transient phenomenon.

Suppose, however, that after the first round of reproduction the two sub-groups merge again to one global population only to be divided into two *new* sub-groups before producing the next generation. If this new division is again done in such a way that altruists tend to concentrate in one group,  $p''$  will be higher than  $p'$ , just like  $p'$  was higher than  $p$ . Hence:

If this process is repeated over many generations, altruists will gradually replace the selfish types, just as the selfish types replaced the altruists in the one-group example. Of course, we still must explain how, generation after generation, altruists tend to find themselves living with altruists, and selfish individuals tend to associate with other selfish individuals. [...] *Altruism can evolve to the extent that altruists and nonaltruists become concentrated in different groups.* (Sober & Wilson 1998, 26.)

Thus, in Sober and Wilson's model the necessary conditions for altruism to evolve are that (i) there must be a population of groups, (ii) the groups vary in their proportion of altruists, (iii) the groups with more altruists will produce more offspring (or be more fit) than groups with low proportion of altruists and (iv) the population must be divided into relatively isolated groups (relatively isolated, since the groups must be mixed again for the formation of new groups) (Sober & Wilson 1998, 26). Of these the only problematic prerequisite is (ii) (the variation of groups in Sober and Wilson's terminology). However, it seems plausible to assume that organisms can develop devices for distinguishing between altruism and egoism, by learning from past experience or otherwise (see also Skyrms 1996 who discusses this under the heading of correlation). Be that as it may, models in which even random variation appears to be sufficient are discussed below (and thus the condition (ii) does not seem to be necessary after all).

The conditions (i)-(iv) do not of course undermine the fact that evolutionary altruism is inescapably "maladaptive with respect to individual selection", but, nevertheless, explicate when altruism can be "adaptive with respect to group selection. Altruism can evolve if the process of group selection is sufficiently strong." (Sober & Wilson 1998, 27.) Sober and Wilson mean that although the altruistic trait in question by definition indeed decreases the fitness of the individual possessing the trait, the group-level dynamics can nonetheless be such that the trait evolves. Although the trait is not beneficial for the individual, it is natural to say that it is nevertheless beneficial for the whole group, or that the trait increases the fitness of the group.

Sober and Wilson think that their group selection model reveals the group-level causal processes responsible for the evolution of altruism, although they admit that the process can be seen also as a standard individual-level selection process. But the individualistic perspective falls, according to Sober and Wilson, guilty of what they call the *averaging fallacy*, and fails to provide proper understanding of the process. The individualistic perspective involves calculating the average number of offspring of *A*-types and *S*-types across the sub-groups (thus looking only at the global population). In the example above, the result is that *A*-types have an average of 12.38 offspring and *S*-types an average of 11.62 offspring. This averaging procedure, which does not change any

facts of the original two-group model, seems to hint that *A*-types evolve by normal individual selection because on average, *A*-types are more fit than *S*-types. “In short, a single trait can appear to be altruistic *or* selfish, depending on whether fitnesses are compared within groups or averaged across groups and then compared” (Sober & Wilson 1998, 32).<sup>43</sup>

Sober and Wilson think that the averaging approach is not able to say more than that in the one-group model, which was an example of the evolution of evolutionary selfishness, *S*-types were on average more fit than *A*-types but, miraculously, in the two-group model the very same *A*-types are on average more fit, whatever the *causes* of this are. The averaging approach is silent about the (causal) dynamics of complex, multi-group evolutionary processes, including the tension between the within-group selection force on the one hand and the between-group selection force on the other.

Sober and Wilson call, correctly in my mind, the averaging approach a *fallacy* due to its concentration on the net outcome and the corresponding blindness concerning the component factors that in fact determine the net outcome. Although the averaging approach is able to describe the correct outcome and even to explain it in some sense, I think it is clear that the group selection model offers a better explanation, since it gives us illuminating and more detailed information about the causal processes producing the outcome.<sup>44</sup>

In their models of the evolution of altruism, mainstream evolutionary theorists tend to assume a population structure that fulfils the requirements of group selection.<sup>45</sup> First, there is a large global population that divides into smaller sub-populations of size *n* in which the interaction that determines fitness takes place. An individual’s fitness is assumed to depend on its own behaviour and the behaviour of its social partners (*i.e.*,

---

<sup>43</sup> A paradigmatic example of uncritical acceptance of the averaging fallacy is Matt Ridley’s popular book *The Origins of Virtue* (1996). Ridley (1996, 19) even claims that the so-called Price equation proves “indisputably” that apparently altruistic traits are, ultimately, evolutionarily egoistic. Thus Ridley (1996, 175) concludes that “biologists have thoroughly undermined the whole logic of group selection. It is now an edifice without foundation.” The present Chapter shows how Ridley’s conclusion is mistaken. Moreover, Sober and Wilson (1998, 71-79) demonstrate how the Price equation is very explicit about the contributions of both within-group and between-group selection forces to the outcome of natural selection. So instead of proving group selection impossible, the Price equation is, *pace* Ridley, one of its clearest expressions.

<sup>44</sup> Sober and Wilson think that also other models of the evolution of altruism that are commonly seen as alternatives to group selection, most notably kin selection (including the closely related selfish gene approach), are in a similar manner just different ways of viewing evolution in populations of several sub-groups. It is just that “the theories are formulated in a way that obscures the role of group selection” (Sober & Wilson 1998, 57). I will not enter into the discussion about the status of kin selection theory here. Sober and Wilson’s arguments how sibling groups form just the kind of population structure needed for group selection to be effective can be found on pages 62 – 67 of their book; see also Skyrms (1996, 61) for a similar line of thought.

<sup>45</sup> This applies fully also to Skyrms’ (1996) discussions.

the fellow members of the same sub-group). After the interaction has occurred, the sub-groups dissolve and the individuals blend back into the global population to form new sub-groups. (Sober & Wilson 1998, 79-80.)

Let us look at a simple model where  $n = 2$  (i.e., the division into sub-groups amounts simply to pairing up for the purposes of a round of play). The possible outcomes can then be presented in a simple  $2 \times 2$  payoff matrix. Suppose that altruists ( $A$ ) give their partner 4 fitness units at a cost of 1 unit to themselves. Thus, when playing with a selfish type ( $S$ ) an altruist suffers a loss of 1 fitness unit, and when paired with another altruist gets a net benefit of 3 units. The selfish type gets 4 units when playing with an altruist, otherwise nothing. When these possibilities are put into the payoff matrix, we get the following matrix that obviously corresponds to the well-known social dilemma situation known as the Prisoners' Dilemma:<sup>46</sup>

		Individual II	
		A	S
Individual I	A	Both get 3	I gets -1, II gets 4
	S	I gets 4, II gets -1	Both get 0

Obviously, altruism has a low relative fitness within sub-groups (an altruist will never do better than its partner). However, the presence of altruism increases the fitness of the sub-group taken together ( $3 + 3 = 6 > 4 + (-1) = 3 > 0$ ). Although this model is a standard group selection model for the evolution of altruism – the outcome depends on the amount of variation among groups – biologists applying evolutionary game theory, however, tend to use in their calculations the individual fitnesses calculated by averaging across groups as follows (using the notation of Sober & Wilson 1998, where  $p_{ij}$  is the proportion of  $i$  types that interact with  $j$  types):

$$W_A = p_{AA}(3) + p_{AS}(-1)$$

$$W_S = p_{SA}(4) + p_{SS}(0)$$

It is tempting to conclude, as many game theorists indeed do, that  $A$  evolves by individual selection when  $W_A > W_S$ . In the light of what has been said before it is, however, easy to see that such a conclusion would be a typical example of the averaging fal-

<sup>46</sup> The dilemma is usually called the *Prisoner's Dilemma*, but in line with my emphasis on the group-perspective, I follow Bruno Verbeek (2002, 35-36) and talk about the *Prisoners' Dilemma*, since from the perspective of a single prisoner there is no dilemma: defection ( $S$ ) is clearly the dominating strategy.

lacy. Actually, what evolves depends on the relative strengths of the within-group selection force for  $S$  and the between-group selection force for  $A$ . In the example the within-group selection force for  $S$  is so strong that if individuals are randomly distributed into groups,  $A$  will not evolve. Or, as mainstream evolutionary game theorists would presumably like to put this,  $S$ 's average fitness is higher and thus evolves by individual selection. However, if the amount of variation among groups is increased, between-group selection will eventually become stronger – causing  $A$  to have the higher average fitness – and altruism will evolve. If we concentrate on individual selection modelled by the average fitnesses, “self-interest is defined as whatever evolves in the model, and altruism and group selection are defined out of existence” (Sober & Wilson 1998, 84). The averaging approach *defines* “individual selection” to be a synonym for “natural selection”.

Sober and Wilson have shown that altruism can evolve by group selection when there is a sufficient amount of variation in fitness among groups, or to put it in other words, when there is a sufficient amount of correlation so that altruists tend to group with other altruists (since in Sober and Wilson’s models the proportion of altruists correlates with the variation in fitness). However, random groupings (or random variation) can also suffice for altruism to evolve, given that altruism entails only a modest disadvantage within sub-groups.

Consider one of the most celebrated strategies (see, for example, Axelrod 1984 and Singer 1994 & 1999) in evolutionary game theory, that of Tit-for-Tat. Let us look again at a model where the size of a sub-group  $n = 2$  but the members of the sub-groups interact with each other repeatedly before blending back into the global population. Tit-for-Tat is the strategy of behaving altruistically in the first interaction and imitating the partner’s behaviour thereafter. The payoff matrix of different pairings of Tit-for-Tatters ( $T$ ) and selfish types ( $S$ ) is as follows (assuming that an average of  $P$  interactions take place within each group before the groups dissolve).

		Individual II	
		T	S
Individual I	T	Both get $3P$	I gets $-1$ , II gets $4$
	S	I gets $4$ , II gets $-1$	Both get $0$

Sober and Wilson observe that “[a]s before, groups of altruists (TT) outperform mixed groups (TS), which in turn outperform groups of selfish individuals (SS)” (Sober &

Wilson 1998, 85). Group selection is still needed for *T* to evolve, since *T* can never do better than *S* when they are paired against each other. Nevertheless, random groupings (random variation in Sober and Wilson's terminology) appear to be sufficient for altruism to evolve.

For those who have become comfortable with the multilevel framework, it is child's play to see the groups in evolutionary game theory, calculate relative fitnesses within and between groups, and determine what evolves on the basis of the balance between levels of selection.

(Sober & Wilson 1998, 85.)

Of course it is once again possible to calculate the average fitnesses of competing strategies across groups and insist that individual selection causes *T* to evolve. The averaging process, however, falls again guilty of hiding the relevant dynamics of the model, and, hence, insofar as it guides one to think that Tit-for-Tat is a selfish strategy, it is a seriously misleading way of looking at the process. This can be easily seen from the payoff matrix, which, as mentioned, shows that *T* can never win a paired encounter (and *S* can never lose), but groups of *T* do so much better than other groups that *T* can, nevertheless, evolve.<sup>47</sup>

Sober and Wilson's conclusion is twofold. First, their model of group selection can resolve the puzzle of the evolution of (evolutionary) altruism. Second, the process of group selection does not include any features unacceptable for a naturalistic biologist. Group selection in Sober and Wilson's sense is a natural process that is most probably rather common in nature (see the empirical evidence discussed in Sober & Wilson 1998). Moreover, although sibling groups can offer exactly the kind of population structure required for group selection, the group selection theory does not restrict altruistic behaviour to relatives. Hence the theory appears tailor-made for resolving the evident tension between the observed co-operative behavioural tendencies among certain social animals (including humans) and the theoretical implications of the uncritical rejection of group selection theories in favour of kin selection in contemporary biology.

A good example of this is Anne E. Pusey's (2002) puzzlement regarding the social behaviour of our close relatives, chimpanzees. Pusey observes that intergroup hos-

---

<sup>47</sup> As a matter of historical interest, Sober and Wilson (1998, 85 ff.) provide evidence for the claim that Anatol Rapoport, the inventor of Tit-for-Tat, saw the strategy as an example of altruism that evolves by group selection in Sober and Wilson's sense. Moreover, Robert A. Axelrod concludes his report on his famous computer Prisoners' Dilemma tournaments won by Tit-for-Tat, followed by other strategies that are "nice" (*i.e.*, that are never the first to defect), as follows: "The nice rules did well in the tournament largely because they did so well with each other, and because there were enough of them to raise substantially each other's average score" (Axelrod 1984, 35). Obviously also this is highly compatible with the group selection interpretation favoured by Sober and Wilson – also Axelrod is forced to refer to the substructure of the global population in order to *explain* the average fitnesses.



tility and intragroup co-operation are characteristic features of chimpanzee groups (Pusey 2002, 17). Pusey (2002, 25-26 & 33) also points out explicitly that the constant within-group co-operation among unrelated male chimpanzees poses a problem for the kin selection orthodoxy of contemporary biology. For a group-selectionist, of course, there is no problem here – especially since, as Pusey herself explains in the very same article, chimpanzees typically live in a “fission-fusion society, in which the individuals of the community spend some time alone and frequently join and leave temporary sub-groups” (Pusey 2002, 14; see also Stanford 2002, 100). Moreover, Pusey observes that although such a social system “is unusual among primates and mammals in general” it “does occur in lions, hyenas, elephants, spider monkeys, and humans” (Pusey 2002, 14) which are all paradigmatic examples of a highly social species. In sum, these social species tend to (i) exhibit altruistic behaviours that are not explainable in terms of individual or kin selection and (ii) characteristically live in the kind of fission-fusion societies required for the group-level selection of altruistic sociality. As a good kin-selectionist Pusey sees here a problem. I do not.<sup>48</sup>

Thus, it seems that Sober and Wilson’s group selection theory is exactly what is needed for an evolutionary explanation of human sociality (including the biological basis of collective intentionality). Nevertheless, before I can go on to discuss the relevance of Sober and Wilson’s theorising for social ontology and action, let us look at Samir Okasha’s criticism of Sober and Wilson’s programme. Okasha provides, I believe, an important insight that helps to see more clearly the ontological status of group selection. The moral I draw from Okasha’s argumentation is, however, quite different from the objectives of his original discussion.

### II.1.2 THE ONTOLOGICAL STATUS OF GROUP SELECTION

Okasha (2001) addresses the question concerning the ontological status of the group selection processes in Sober and Wilson’s sense. Sober and Wilson argue that their theory describes a group-level process analogical to individual selection where groups play the

---

<sup>48</sup> Kitcher (1998) shares my suggestion that the known facts about the strong sociality of the Great Apes and about the typical structure of the groups they live in might support the importance of group selection in the evolutionary explanations of sociality. Of course the question concerning the role group selection in fact played in the evolution of the social traits of humans and other animals is a question I must leave to biologists. A very recent review of these issues concludes that although the question is still somewhat open, empirical estimates show that the genetic structure of early human groups were such that they could account for the evolution of human sociality via a group selection process (Bowles 2006). Be that as it may, this Part shows that evolutionary considerations do not speak against the we-mode view of this study but rather support it.

roles of individuals. However, what really matter in the group selection models are normal individual interactions in temporary groupings. All the actions that take place in the models are interactions between individuals. Moreover, the fact that Sober and Wilson's theory treats as any temporary sub-group within a larger population where individuals "influence each other's fitness with respect to a certain trait but not the fitness of those outside the group" (Sober & Wilson 1998, 93) shows that the term "group" in Sober and Wilson's theory is used to pick out a class very different from the class of groups in any socially significant sense. Groups in Sober and Wilson's sense may have no role whatsoever outside evolutionary models: they are not what we mean by social groups.<sup>49</sup>

Hence, I think it is reasonable to conclude that Sober and Wilson's model is not a model of group-level processes as such. It is not the behaviour of groups as independent entities that is at stake, but that of individuals, and the system of organising the relations between interacting individuals. Sober and Wilson's group selection does not operate at some higher ontological level. The operative processes are essentially – ontologically – individualistic; that is, they take place between individual organisms. Sober and Wilson simply show that sometimes the outcome of a series of such processes is influenced crucially by the groupings of these individual interactions, and that behaviour that benefits the group can be selected. Nonetheless, for a group selection process to be truly a group level process there should be real group reproduction and group heritability. These features, however, are missing from Sober and Wilson's theory (Okasha 2001, 35).

Sober and Wilson do not want to accept this interpretation but commit themselves to a serious attempt to show that it does indeed make sense to talk about group heritability and group reproduction within their model (Sober & Wilson 1998, 110 ff.). Be that as it may, I tend to agree with Okasha that even if it is *possible* to apply the concepts of group reproduction and group heritability in Sober and Wilson's model, this possibility is in fact *irrelevant* to the issue of group selection. After all, Sober and Wilson's fundamental motivation for the introduction of group selection models is the need

---

<sup>49</sup> Note that this is highly compatible with the present suggestion that group selection in Sober and Wilson's sense can explain the evolution of collective intentionality: Sober and Wilson's theory can justify Searle's seemingly circular (recall Part I) claim that although collective intentionality is a fundamental building block of social reality *including social groups* (Searle 1995), nonetheless "collective intentionality seems to presuppose some level of sense of community before it can ever function" (Searle 1990, 413). However, the "sense of community" assumed in Sober and Wilson's model is based on interactions quite independent of groups in any socially significant sense and hence do not presuppose the existence of social groups. Correspondingly, the fact that Sober and Wilson's theory does not operate with socially significant groups must in the present context be seen as a virtue of the theory.

to explain the evolution of altruism, which is a trait of an individual (albeit social in the sense of being directed towards others), not of a group. Moreover, since the concept of fitness applied in the model is the average individual fitness of the organisms in a group, and not the group's propensity to leave offspring groups, the notion of group heritability appears indeed to be irrelevant (Okasha 2001, 40-41).

The whole idea of "benefiting the group" that plays such a crucial role in Sober and Wilson's theorising is not measured by an individual's capacity to cause a "group to leave more offspring groups", but whether an individual will cause "the individuals in the group to leave more individual offspring" (Okasha 2001, 43). Hence, when Sober and Wilson conclude that "[a]t the behavioral level, it is likely that much of what people have evolved to do is *for the benefit of the group*" (Sober and Wilson 1998, 194), this means simply that understanding the dynamics of a group-level selection force allows us to see that acting for the common good of one's group, even at the expense of one's own fitness, is a strategy that can be, and most probably has been, selected during the course of evolutionary history.

Thus, I think that Sober and Wilson succeed in showing the importance of explicating the social dynamics in the models of the evolution of group-behaviours, but their theory should nonetheless be understood as supporting *ontological individualism (OI)* in social ontology. When Sober and Wilson insist that their group selection theory is a theory about the behaviour of groups, they are in a sense misrepresenting their own theory. In Sober and Wilson's models the only functional units are individuals, not some metaphysically questionable social objects at a new ontological level.

Although the group selection theory subscribes to ontological individualism, it does speak against *methodological individualism* in its standard form. The theory points towards exactly the kind of view reaching beyond the traditional individualism versus holism debate that also the collective intentionality theorists are envisaging. The group selection theory suggests that although there are no group agents, we should not understand individuals as disparate social atoms either. Rather, the theory supports the plausibility of seeing humans as *social animals*. By this I mean that we should expect humans to be inclined to perform actions the rationality of which cannot be understood if humans are seen simply as individuals acting strategically to maximise their own individual benefits. The group selection theory hints that some forms of behaviour can be understood as rational action only when we see the behaviours essentially as interconnected tasks that are derived from a collective goal, and cease to consider the behav-

iours as private strategies for private goals. Group selection favours social behaviour that follows collective-level rationality.

Thus, the group selection arguments for evolutionary altruism suggest that the picture of human sociality appropriate for explanations of human social behaviour is not the picture one finds in standard rational choice models. Some forms of behaviour can be understood only when the agent is seen *essentially* as a member of her group, *i.e.*, in the we-mode. To put this in Gilbert's terminology, some actions may be properly understood only when the individual performing the action is seen as a member and constituent of a plural subject.

The reason why this view is only suggested and not implied by the group selection arguments is that so far I have discussed only the *evolutionary* concepts of altruism and egoism. However, we explain human *action* in *psychological* terms of beliefs, desires and intentions. As emphasised above, the evolutionary arguments presented so far consider only the fitness consequences of different behavioural strategies regardless of the psychological states associated with those strategies. It is possible that even if human evolution has favoured (evolutionarily) altruistic behaviour, the psychological processes involved are nonetheless egoistic (or at least individualistic) in nature.

This idea is familiar from several traditions in the history of philosophy. The great Swedish natural philosopher Carl von Linné (Linnaeus) argued that the divine order of the world guarantees that individuals seeking their own benefit will unintentionally follow God's plan, realising thus also the collectively optimal state of affairs (see, for example, Larson 1971). Early liberalists, such as the Finnish economist/philosopher Anders Chydenius (1765/1994), accepted this idea, or, like the Scottish philosopher Adam Smith, aimed to give it a more naturalistic account in terms of the doctrine of the invisible hand.<sup>50</sup> Similar ideas can also be found in the context of classical German idealism, within which it was argued that sometimes individualistic actions must be understood as participating in the collectively rational action of the objective spirit.

However, ontological individualism compels us to abandon such ontologically dubious solutions – with the exception of the invisible hand doctrine. Thus the question is whether evolutionarily altruistic behaviour, selected by a group selection process, can be satisfactorily accounted for in terms of exclusively egoistic and individualistic psychology. We shall see Sober and Wilson arguing that this cannot be done; psychological

---

<sup>50</sup> Although Smith is routinely seen as a methodological individualist, Robert Sugden (2002, 82) argues convincingly that Smith's notion of fellow-feeling is after an idea similar to the collective we-mode. If Sugden is correct, then Smith might actually have agreed with the anti-individualistic view reaching beyond the rational choice framework that I defend in the next Chapter.

altruism as an ultimate human motivation must be assumed. However, as can be anticipated, I argue that even psychological altruism is not sufficient. A naturalised version of the Durkheimian collective consciousness is needed – we must reject Individualistic Accounts in favour of the collective intentionality view of Part I. These considerations form the topic of Chapter II.2.

## CHAPTER II.2

THE PSYCHOLOGY OF SOCIAL ACTION<sup>51</sup>

Evolution can select traits and patterns of behaviour that do not maximise the individual fitness of the agent. Thus, if the explanatory question we are interested in is why individuals perform behaviours that benefit the group the individuals in question belong to but decrease the personal fitness of the individual performing the behaviour, the answer may be that during the (biological and/or cultural) evolutionary history of humans group selection was such an effective force that such behaviour was indeed selected, despite the individualistic within-group selection force acting against it. However, sometimes this is not the kind of explanation we want. We may rather be interested in the nature of the psychological processes underlying the action in question. If this is the case, the answer should be stated in terms of the intentions, beliefs and desires people have. In short, we are not interested only in evolutionary explanations of patterns of behaviour. Sometimes we are after psychological explanations of actions.

Let us begin the quest for psychological explanations of evolutionarily altruistic actions with Sober and Wilson's question: Are human motives always ultimately self-directed or are we sometimes motivated by other-directed considerations? In Sober and Wilson's terminology a person's psychological motives are *self-directed* if the ultimate goal of the person's actions is to maximise her own benefits and well-being. Similarly, a person's motives are *other-directed* if she is concerned about the well-being of others as an end in itself. Thus:

*Psychological egoism* is the theory that all our ultimate desires are self-directed; (...) *psychological altruism* maintains that we sometimes care about others for their own sakes. The theories agree that people sometimes want others to do well; the debate concerns whether such desires are always instrumental or are sometimes ultimate. (Sober & Wilson 1998, 201.)

Note that egoism is by definition a strictly *monistic* theory. Egoism says that *all* the *ultimate* motivations are self-directed: we perform other-directed actions only as means to our more fundamental self-directed goals. Altruism, in contrast, is essentially a *pluralistic* theory of motivation: it says that although self-interest no doubt often explains our behaviour, we also have *other* ultimate reasons for action.

The pluralism of altruism is very important from the point of view of my dissertation. I argue that there is no *a priori* reason to restrict the ultimate ends an agent may

---

<sup>51</sup> This Chapter expands on Saaristo (2007).

have exclusively to egoistic ends on the one hand and altruistic ends on the other. Psychological pluralism can go further by accepting also some other ultimate ends, such as the benefit or well-being of a group. Moreover, I argue that we need a view that allows not only pluralism in terms of the *directedness* of intentional attitudes, but also in terms of the *mode* in which the attitudes are held. However, in order to demonstrate the importance of pluralism in terms of the modes of attitudes, it is important to see why Sober and Wilson's pluralism in terms of the directedness of attitudes is not sufficient. This is the aim of the next section.

### II.2.1 AN EVOLUTIONARY PERSPECTIVE

How can we tell whether intentional attitudes that explain altruistic actions are always ultimately self-directed or sometimes also other-directed? Since neuroscience does not answer this question (and arguably cannot even in principle, see Part III), the most straightforward way of finding out the answer appears to be to construct psychological experiments that provide direct evidence concerning motivational mechanisms. However, psychological work has not been able to achieve a consensus in this matter (see Chapter 8 of Sober and Wilson 1998 for a review of a number of psychological studies – my own analysis of the prospects of such empirical approaches is given in II.3 and III.5.3 below).

Moreover, it is practically impossible to distinguish between psychologically altruistic and psychologically egoistic desires insofar as we allow egoism to appeal to purely internal rewards such as feelings of satisfaction when behaving (seemingly) altruistically. Desires seem to be such that in most cases we can expect that the realisation of a desired state of affairs will bring feelings of satisfaction regardless of the directedness of the desire. Thus, the mechanism of internal rewards in terms of satisfaction allows an egoist to insist that an advocate of altruism cannot falsify the egoistic theory. An altruist, on the other hand, could say that the satisfaction derived from fulfilled desires is part of the nature of our intentional psychology in general, and has nothing to do with the question of the directedness of our motives.<sup>52</sup> In other words, it is not clear what solid evidence for or against psychological egoism would even look like.

---

<sup>52</sup> For a similar line of thought, see Rachels' (1986, Chapter 5) "Psychological Egoism". Rachels concludes that "if someone desires the welfare and happiness of other people, he will derive satisfaction from helping them; but this does not mean that those feelings are the object of his desire. They are not what he is after. Nor does it mean that he is in any way selfish on account of having those feelings." (Rachels 1986, 60.)

Since direct empirical evidence has not been able to solve the problem of psychological egoism and altruism, let us use as a point of departure what Sober and Wilson (1998, 298) call an *indirect* method of addressing the question and analyse evolutionary models of social action. The problem this method concentrates on is an evolutionary *design problem*. What kind of psychological mechanism is the most plausible candidate to be the device that has evolved to be responsible for evolutionarily altruistic action?

Note that all the competing views here, *i.e.*, self-directed (egoism), other-directed (altruism) and we-mode psychology (collective intentionality), differ from each other only in their representations of ultimate attitudes, not in the physiological devices they require from the agents the views apply to. Thus there should be no considerable differences in terms of availability or energetical efficiency either (cf. Sober and Wilson 1998, 320-327). Hence, when considering which of the competing views has the highest degree of evolutionary plausibility as a theory of the central psychological mechanism explaining evolutionarily altruistic actions, we can concentrate exclusively on *reliability*. The psychological mechanism that turns out to be the most reliable device for producing evolutionarily altruistic action is the one that we should think of as the most plausible candidate to be the mechanism that in fact has evolved for that purpose.

It is, of course, crucial for my argument that Sober and Wilson are correct in claiming that there are no differences in terms of availability and energetical efficiency between the competing mechanisms. Optimality arguments, such as my concentration on reliability, can explain only why we should expect the evolution of the most optimal of the traits that were *available* in a given population. Needless to say, such a trait may not be the most optimal trait we can *imagine*. As Sober (1993, 120) puts it, “[a]daptationists might expect zebras to evolve from *Slow* to *Fast* but will not expect them to evolve machine guns with which to counter lion attacks”. However, the competing psychological mechanisms, unlike the mechanisms needed for running and for producing machine guns, do not differ in terms of the equipment and devices they require. Hence the deciding factor is reliability.<sup>53</sup>

At this point I must add a word of warning. In what follows my arguments concentrate on evolutionary dynamics purely in the abstract. In this part I do not even take a stand on whether the evolutionary dynamics I analyse should be understood as biologi-

---

<sup>53</sup> If this conclusion turns out to be false, my conceptual reliability arguments need to be completed by biological studies concerning availability and energetical efficiency. Such a completion would not, of course, undermine the plausibility of the reliability arguments discussed in this study.



cal or cultural evolution.<sup>54</sup> Recall how Section I.1.5 identified two possible ways of understanding the nature of intentionality: the Wittgensteinian account that sees intentionality essentially as a social institution and the anti-Wittgensteinian picture of intentionality as a biological feature of the brain. This huge issue was left open in Part I. Similarly, in this Part I want my arguments to be relevant for all philosophers interested in the nature of rational social action.

Thus, if one thinks, as for example Searle does, that intentionality is an internal, biological feature of the brain, then my arguments can be interpreted as applicable to standard biological evolution of causally efficacious natural properties. If one is more inclined, in the manner of, for example, John Haugeland (1990), to consider intentionality as a social institution, then my evolutionary arguments should be interpreted in terms of the cultural evolution of social institutions and norms. The question of the ontological nature of intentionality, including its implications for the philosophy of mind and intentional explanation, is addressed in Part III.

## II.2.2 SOCIAL DILEMMAS: THE UTILITY TRANSFORMATION RULES

Although Sober and Wilson see the goal of their defence of the plausibility of both evolutionary and psychological altruism to be the creation of a naturalistic and robust theory of the foundations of human sociality, their argumentation in favour of psychological altruism is conducted in terms of human parental care, which is *not* a paradigmatic example of human sociality. After all, parental care is both strictly limited in its scope and, arguably, largely based on blind instincts instead of deliberations with conceptual content. Both features are clearly uncharacteristic of human sociality as we know it. When we turn our attention to paradigmatically social cases that I call social dilemma situations, Sober and Wilson's arguments in favour of other-directed altruism will be found wanting.

The first thing to notice is that although Sober and Wilson talk about other-directed altruism in the sense of preferring someone else's benefit in comparison with one's own benefits, also for them it is often *group-directed* sociality (in the sense of

---

<sup>54</sup> My discussion builds largely on the views of Sober and Wilson (1998) for whom the context is mainly biological evolution. However, they share the view that what matters is the basic structure of group selection, regardless of whether this structure is instantiated in a biological or cultural process. After all, Sober and Wilson base their theory largely on the Price equation which, as Knudsen (2004) argues, offers an account of selection without any explicit biological content, being thus in principle applicable also to non-biological evolutionary processes. A good example of this is Wilson's (2002) own later work, which is an application of the ideas of Sober and Wilson (1998) to purely cultural processes.

asking what is the best behaviour for maximising the benefit of one's group, *including* the actor) that matters. As opposed to an altruistic agent, a group-directed agent is not asking herself how to maximise someone else's benefit, but what to do to maximise *our* benefits, how to act *qua* a group-member: "Behaving as part of a coordinated group is sometimes a life-or-death matter in which the slightest error – or the slightest reluctance to participate – can result in disaster for all" (Sober & Wilson 1998, 335-336). Or elsewhere, even more clearly: "The 'I' is defined by relating it to a 'we.' Human beings don't simply *belong* to groups; they *identify* with them. This is an important fact about human experience."<sup>55</sup> (Sober & Wilson 1998, 233.) This is indeed the kind of sociality we are after, not parental care. It will turn out that Sober and Wilson's defence of psychological altruism fails to account for this.

Thus, the basic problem here concerns the most reliable candidate for the intentional mechanism rationalising the kind of social action that makes possible the rise of social reality with robust social institutions and social facts (and which is selected by group selection), not the explanation of parental care. Accordingly, instead of looking at parental care, I follow an influential tradition in social philosophy and accept that the problem boils down to the problem of achieving collectively rational solutions in so-called social dilemma situations.<sup>56</sup> I define a social dilemma situation by following a general description provided by Raimo Tuomela (Tuomela's "collective action dilemma" can here be understood as a synonym for my "social dilemma situation").<sup>57</sup>

The problem of collective action or a collective action dilemma is a dilemma or conflict between collectively and individually best action, where the action required for achieving the collectively best outcome or goal is different from (and in conflict with) the action required for achieving the individually best outcome. Or, as we may also put it, means-end rational action realizing what is collectively best is in conflict with means-end rational action realizing what is individual[ly] best.  
(Tuomela 2000, 258.)

In a social dilemma situation agents need to find a way to co-operate so that they can achieve the collectively (socially) best outcome even when from each agent's indi-

---

<sup>55</sup> Obviously, this supports not only group-directedness but also the plausibility of we-mode collectivism, *i.e.*, the possibility of having we-mode desires *qua* a member of a group, regardless of the content and directedness of those desires (see Part I and Section II.2.4 below). Moreover, Sober and Wilson's explicit appeal to the notion of *social identification* points directly to social identity theories in social psychology, which are, as I argue in II.3 below, best understood in the context of we-mode notions.

<sup>56</sup> For an extensive discussion and arguments in favour of this tradition, see, for example, Hardin (1982) and M. Taylor (1987).

<sup>57</sup> For Tuomela this description is just a general illustration of such a situation; much of Chapter 10 of Tuomela (2000) is devoted to defining in detail what is a collective action dilemma. For my purposes, however, the informal description suffices.

vidual point of view the best strategy for maximising the expected individual benefits is not to co-operate, *i.e.*, to defect. I call the co-operative strategy *C* and defection *D*. Moreover, I concentrate on the strongest (in terms of conflict) kind of social dilemma situation, famously exemplified by the one-shot Prisoners' Dilemma (PD) game. No doubt some weaker problems would also be of considerable interest to discuss in the present context, but since my problem is the reliability of psychological motivational mechanisms for producing mutual co-operation in social dilemma situations, I need to tackle the problem in its strongest form. And that is exemplified by one-shot PDs.<sup>58</sup>

So let us look at the payoff structure of a standard one-shot PD game:

(PD)

		Player 2	
		C	D
Player 1	C	3,3	1,4
	D	4,1	2,2

Let us assume that players 1 and 2 recognise the payoff structure PD, formed by the objective and external features of the situation. These objective features form the *given utilities* of the agents. In what follows these are represented such that  $u_{ijk}$  stands for player  $k$ 's given utilities of the outcome created by player 1 playing the strategy  $i$  and player 2 the strategy  $j$ .<sup>59</sup> In the present context of a two-person PD the parameters  $i$  and  $j$  must be either *C* or *D*, while  $k$  can be either 1 or 2. For example, player 1's utility in the outcome produced by both players playing *C* is  $u_{CC1} = 3$ . In the above PD  $u_{DC1} = 4$ ,  $u_{CC1} = 3$ ,  $u_{DD1} = 2$  and  $u_{CD1} = 1$  and, similarly,  $u_{CD2} = 4$ ,  $u_{CC2} = 3$ ,  $u_{DD2} = 2$  and  $u_{DC2} = 1$ . Thus, both player 1 and player 2 will reason that whatever the other player is going to play, she herself will maximise her expected utility by playing *D*, resulting to the collec-

---

<sup>58</sup> Sometimes it is argued (*e.g.*, M. Taylor 1987) that a one-shot PD cannot exemplify a social dilemma situation realistically, since in social life we quite probably will face similar dilemmas over and over again, and, hence, we should talk only about repeated PD games. However, we have already seen (II.1.1) that the repetition simply brings in the standard structure of group selection models, and the successful behaviour in a repeated PD game that appears to be egoistic is in fact evolutionarily altruistic, the selection of which requires group selection. For the present psychological discussion this implies that successful strategies in repeated PD games, because of their evolutionarily altruistic nature, already presuppose a psychological mechanism capable of explaining altruistic actions. Hence, to use such games in the present context would beg the question. Moreover, humans exhibit co-operative social behaviour also in cases where they are not likely to interact with the same partners in the future. I think that this is a sufficient answer to the worry concerning the justification of my concentration on one-shot games, but I will nonetheless return to this problem in II.2.4 both in the context of Tit-for-Tat strategies and in the context of the so-called argument from long-term considerations.

<sup>59</sup> This notation is a simplified version of the notation of Tuomela (2000, 219 ff. and, especially, pp. 281 ff.), whose line of thought my discussion partly follows.

tive outcome *DD*. In this manner individually rational deliberations will in a PD game produce a collectively sub-optimal result.

I read Sober and Wilson as arguing that the design problem of the nature of our ultimate psychological motivations can be formulated as the question of how the players transform the objective features of the situation (the given utilities) into *motivationally effective utilities*  $u'_{ijk}$  upon which the agents act. The crucial presuppositions<sup>60</sup> in Sober and Wilson's argumentation are that the given utilities  $u_{ijk}$  represent external rewards and, moreover, that the agents are not required to act on those utilities, but instead they are allowed to form their effective utilities  $u'_{ijk}$  in accordance with how they personally value different possible reward distributions.

Since we are here especially interested in to what extent the agents are disposed to let the considerations concerning the other agent's benefits enter into their own decision making, we can follow Tuomela (2000, 281) and represent player 1's effective utilities  $u'_{ij1}$  as the sum  $au_{ij1} + bu_{ij2}$ , where the parameters  $a$  and  $b$  can have the values from  $-1$  to  $1$ . As Tuomela puts it, although such a linear model can at best be "empirically approximate", it nonetheless "serves to give conceptual illumination" (Tuomela 2000, 281), which is what I am after in this dissertation. Moreover, let us look at the "pure" cases in which  $a$  and  $b$  can only have the values  $-1$ ,  $0$ , and  $1$ . Looking at the situation from player 1's perspective, we get the following 9 possible *utility transformation rules*:

	a	b		
(i)	-1	-1	$u'_{ij1} = -u_{ij1} - u_{ij2}$	(group-sacrifice)
(ii)	-1	0	$u'_{ij1} = -u_{ij1}$	(self-sacrifice)
(iii)	-1	1	$u'_{ij1} = -u_{ij1} + u_{ij2}$	(self-sacrificial altruism)
(iv)	0	-1	$u'_{ij1} = -u_{ij2}$	(aggression)
(v)	0	0	$u'_{ij1} = 0$	(apathy)
(vi)	0	1	$u'_{ij1} = u_{ij2}$	(altruism or other-directedness)
(vii)	1	-1	$u'_{ij1} = u_{ij1} - u_{ij2}$	(competition)
(viii)	1	0	$u'_{ij1} = u_{ij1}$	(egoism or self-directedness)
(ix)	1	1	$u'_{ij1} = u_{ij1} + u_{ij2}$	(group-directedness)

Of these 9 possible "pure" psychological mechanisms for adjusting oneself to a social interaction situation it is, obviously, cases (vi), (viii) and (ix) that have direct bearing on

<sup>60</sup> These presuppositions will be discussed – and rejected – below. Let us, however, play along for a while since it allows me to show where exactly Sober and Wilson go wrong in their defence of psychological altruism and why collective we-mode concepts are needed.

the present discussion. To resolve Sober and Wilson's design problem we must examine which of the three rules is the most reliable rule for transforming the given utilities that form a PD situation into such effective utilities that will guide the players to realise the collectively rational universal co-operation outcome *CC*.

Case (viii), egoism, recognises the given utilities also as the effective utilities. Thus, when two egoistic players face a social dilemma situation exemplified by PD, both players figure out that whatever the other player does, the agent herself will be better off by playing *D*. The result is, of course, universal defection and a socially inferior outcome, which, moreover, is the second worst also from the individual point of view of each player. Thus, egoistic motivations fail to provide collectively rational action in social dilemma situations such as PD.

Case (vi), altruism, is more interesting. If both players in a PD game are altruists, the payoff matrix PD' of their final, *i.e.*, transformed effective utilities, can be represented as follows:

(PD' – two altruists)<sup>61</sup>

		Player 2	
		C	D
Player 1	C	3,3	4,1
	D	1,4	2,2

Here player 1 will reason that whatever the other player does, she can better realise her altruistic goal by playing *C*, and similarly for player 2. Thus, two altruistic players should end up in the collectively optimal outcome *CC*. Are we to conclude, then, that psychological altruism is indeed a reliable mechanism for providing collectively rational outcomes? I do not think so. The peculiar feature of reciprocal altruism as a solution to social dilemma situations is that although it leads to *CC*, both players will also feel *unsatisfied* with the result. This is due to the fact that as altruists both players would prefer the situation where she herself plays *C* and the other player plays *D* to the universal co-operation outcome *CC*.<sup>62</sup> This is why psychological altruism cannot be a reliable

<sup>61</sup> The label PD' might be a bit misleading, since the game is not a PD game anymore. However, the label is simply meant to indicate that the new game is formed from a PD game by applying a utility transformation rule.

<sup>62</sup> Compare this to the somewhat perverse game between an altruist and an egoist. Whatever strategy the other player follows, the altruist will reach her ends better by playing *C* and the egoist by playing *D*. The resulting outcome is *CD* in which the egoist ruthlessly exploits the altruist and both players prefer this situation to any other possible outcome. They manage to solve the PD in a psychologically satisfactory way, but this combination of psychological mechanisms cannot produce a socially bearable – or indeed stable – outcome (or behaviour selectable in a group-selection process).

device for rational production of socially optimal outcomes in social dilemma situations.

Suppose that when two altruists transform the original PD into PD' in accordance with their altruistic inclinations they do not act immediately on their transformed utilities but are able to contemplate also the new payoff matrix PD'. Player 1 may now reason that if they both play *C* as the transformed utilities would imply, she cannot (for the reason stated above) be satisfied with the outcome. Moreover, player 1 may continue her reasoning and infer that player 2 will no doubt play *C*. If player 1 will indeed reach this conclusion about the expected behaviour of player 2, player 1 will also realise that the rule (vi) says that as an altruist she in fact must play *D*, since that will maximise the benefit of player 2. The twist here is of course that player 2 may engage in exactly similar reasoning, and hence they may end up in the mutual defection outcome *DD*. This is what Tuomela (2000, 288) calls "the Altruist's *first* dilemma".

As Tuomela's naming of the dilemma suggests, there is also another problem that altruists will have to face. It is also possible that after the players have transformed their utilities and formed PD', the players will not act on those utilities even in the sense of the Altruist's first dilemma. Rather, the agents may realise that after the transformations have taken place, what they actually are looking at is a completely new situation. And the altruistic theory of motivation tells us that when facing a new situation, an altruist will transform her utilities as dictated by the rule (vi). The problem is of course that mutual application of (vi) to PD' of two altruists will lead to the original Prisoners' Dilemma PD. As Tuomela puts it, "[i]f our altruists cannot stop their transformations, they are caught in a never-ending new dilemma, as they never get down to action. This indecision problem we can call the Altruist's *second* dilemma." (Tuomela 2000, 288.) I conclude that insofar as we understand altruism as a psychological mechanism corresponding to the utility transformation rule (vi), altruism is not a very probable candidate for solving the design problem of producing collectively rational action. The two altruist's dilemmas show that the altruistic solution is far too unstable for being the most reliable solution available.

Let us, then, look at the group-directed utility transformation that was captured by the rule (ix) above. If two agents that are disposed to group-directed behaviour face the PD, the rule (ix) will lead their effective utilities to form the following payoff structure:

(PD' – two group-directed agents)

		Player 2	
		C	D
Player 1	C	6,6	5,5
	D	5,5	4,4

As Tuomela puts it, the rule (ix) has transformed PD “into a harmless coordination game” (Tuomela 2000, 283); each player will reason that whatever the other will do, she herself will be better off by playing C. Hence, the result is the universal co-operation outcome CC, or in other words, the collectively rational and socially optimal result CC is also individually the number one choice of each agent. Thus, the crucial difference from altruism in the sense of (vi) is that group-directed players will also be psychologically satisfied with the outcome CC and will feel no need to re-modify their behaviour or utilities. Thus, it seems plausible to conclude that in a PD with a payoff matrix as the one presented above, group-directedness in the sense of (ix) is a more reliable psychological mechanism for producing socially optimal action than pure other-directed altruism in the sense of (vi).

However, this kind of criticism of Sober and Wilson’s emphasis on altruism overlooks Sober and Wilson’s insistence that altruism should not be understood as a monistic doctrine. The criticism of altruism as presented above has to assume that the pluralistic nature of altruism means simply that sometimes we base our actions on purely egoistic motives and sometimes on purely altruistic motives. This, however, is not the most charitable way of interpreting altruism as a pluralistic theory of motivation in general, and in particular it is *not* how Sober and Wilson interpret it.

On the contrary, Sober and Wilson are very explicit that the most interesting cases of motivational pluralism are those in which egoistic and altruistic motivations coexist *simultaneously in one individual* (see especially Sober & Wilson 1998, 242-250). For example, Sober and Wilson (1998, 354, Footnote 23) discuss the possibility of an action where both egoistic and altruistic motivations are effective and, moreover, weighed as *exactly equally important*. This would, of course, model a situation similar to (ix), since, as Tuomela (2000, 283) points out, the group-directedness of (ix) is essentially the *average* between purely egoistic and purely altruistic mechanisms. Perhaps group-directed psychology should not be seen as a third form of human motivational mechanisms but rather as a combination of simultaneous egoistic and altruistic motives.

However, the group-directed transformation rule (ix) does not always lead to a uniform solution in PD type games. Actually, the plausibility of (ix) is very sensitive to the given utilities. Tuomela (2000, 284-285) discusses a PD situation in which free riding is slightly more beneficial than in the original case discussed above (PD). Suppose, for example, that if an agent plays *D* while the other player plays *C*, the agent playing *D* would get 6 utiles instead of 4, *i.e.*, let us consider a situation with the following payoff matrix:<sup>63</sup>

(PD\*)

		Player 2	
		C	D
Player 1	C	3,3	1,6
	D	6,1	2,2

The rule (ix) would then transform the payoff matrix of the effective utilities to be the following:

(PD\*\* – two group-directed agents)

		Player 2	
		C	D
Player 1	C	6,6	7,7
	D	7,7	4,4

Clearly in this case *C* is no longer the dominating strategy; the players have created a new co-ordination problem for themselves. Although Tuomela admits that insofar as *CC* is the collectively preferred outcome, what the group-directed players indeed have here is a new dilemma to be solved, he unfortunately does not discuss the new dilemma too much. Tuomela is mainly interested in explaining how individuals can reach an outcome they both agree with and thus he is able to by-pass this new dilemma, since here “the participants can easily come to agree on one of the (7,7)-pairs, which are equilibria in this coordination game” (Tuomela 2000, 285).

Admittedly Tuomela is right that the new dilemma is not very challenging if the explanatory goal is *agreement*, since there is no serious clash between the utilities of the

<sup>63</sup> Note that PD\* does not satisfy some standard definitions (*e.g.*, Axelrod 1984, 10) of a Prisoners' Dilemma game, since in PD\* the players can get out of their dilemma by taking turns in exploiting each other, *i.e.*, a reciprocal pattern of *C-D-C-D-...* and *D-C-D-C-...* is for each agent better than mutual co-operation. I return to this point below.



players (although the players may of course have serious practical problems in choosing the same equilibrium). But if the goal is to explain the achievement of collectively optimal behaviour in a more objective sense, the situation is harder to deal with. Firstly, it is not that clear what is the collectively optimal combination of actions here. Remember that the context of discussion is a one-shot game<sup>64</sup> and the given utilities are assumed to represent external rewards. Hence, it seems that the realisation of *CC* is indeed objectively speaking a better outcome than *CD* or *DC* (see Skyrms 1996 for a discussion on the objective advantages of fairness), although both players would prefer them to *CC* because of their psychological mechanisms. If this is the case, *PD\**' represents a serious dilemma.

Fortunately Sober and Wilson, although they do not explicitly discuss social dilemma situations, are nevertheless able to offer a way of dealing with *PD\** and *PD\**'. Sober and Wilson appear to think that real agents are quite probably what they call 'E-over-A pluralists' (Sober & Wilson 1998, 245). E-over-A pluralists do want others to do well, but when the self-interest and the welfare of others are in too strong a conflict, an E-over-A pluralist will abandon the altruistic standpoint and secure her own gains. In a situation such as *PD\** an E-over-A pluralist might reason that she cannot ascribe the utility of 7 to the situation in which she plays *C* and the other player plays *D*.

However, this type of pluralism is available also from the point of view of a person who thinks that group-directed desires should be understood as psychologically ultimate motivations. Moreover, it could be argued that such a E-over-G pluralism is at least sometimes a more realistic model of what is in fact going on in the minds of ordinary agents than the E-over-A pluralism combining egoism and altruism. In some cases an agent may reason that even though a given form of behaviour would bring great benefits to the group, it nevertheless requires a too dramatic sacrifice from her own part. In such a situation the agent might conclude that it just is not fair to require such a behaviour from her, and she is therefore justified in performing the egoistic action *D*.

This kind of reasoning corresponds to free-riding situations in which an agent is considering whether or not she should participate in the production of a collective good. An example might be cleaning a park. From the egoistic point of view the agent would be better off by letting others to do the dirty work of cleaning the park, and just enjoy-

---

<sup>64</sup> If the context were that of a repeated game, the socially optimal solution would no doubt be to alternate between *CD* and *DC*, not to always go for *CC*. But a reliable realisation of such a rule would presuppose that the players indeed understand that following that rule is better for *them* collectively, and to follow that rule is what they should do *together* as their joint task. Below (in the context of *PD\**, Tit-for-Tat and the so-called argument from long-term considerations) I will argue that these solutions are, once again, best understood in the context of we-mode notions.

ing the outcome of the others' efforts. Compare this to the case of a group-directed agent who might reason that it is better for the group if all the members participate in the collective task, and hence she will participate too. However, if it is clear that in order to co-operate she would have to go through a considerable amount of trouble (maybe she is the only one who showed up in the meeting organised to plan the voluntary work of cleaning the park), the egoistic perspective might take over. No doubt there are cases in which a pluralism combining egoism and altruism would intuitively seem more plausible than a pluralism combining egoism and group-directedness, but the point I am making is that there are situations in which group-directed attitudes seem intuitively quite realistic.

The argument presented above as well as everyday experience seem to suggest that as long as the personal price to pay is not too high, our motivations are quite often group-directed in nature. In other words, the E-over-G pluralism is often psychologically more believable than the E-over-A pluralism. The twist here is that the E-over-A pluralism can be interpreted such that it does better in theoretical models.

I interpreted above Sober and Wilson's E-over-A pluralism such that an E-over-A pluralist may adopt an attitude towards one possible outcome within one game that is different from the attitudes she has towards the other outcomes. This reading would imply that Sober and Wilson's E-over-A pluralism is even more comprehensive pluralism than merely an application of the rule (ix). Sober and Wilson seem to think that agents can apply different utility transformation rules to different action combinations even within one game. Thus, when two E-over-A pluralist face a situation such as PD\*, they will both recognise the given utilities and apply the altruistic utility transformation rule (vi) to all the other possible outcomes, except the one in which the agent herself plays *C* and the other player plays *D*. As E-over-A pluralists the agents will realise that the outcome *CD* includes an unacceptably high personal price and, hence, apply the egoistic utility transform rule (viii) to that outcome. The result is the following payoff matrix:

(PD\* – two E-over-A pluralists)

		Player 2	
		C	D
Player 1	C	3,3	1,1
	D	1,1	2,2

E-over-A pluralists would transform PD\* into a very straightforward<sup>65</sup> co-ordination game in which both players prefer the universal co-operation outcome *CC* to all the other possible outcomes. Note also that since the E-over-A utility transformations eliminate the temptations of the *CD* and *DC* outcomes, the above payoff matrix is also the result of the utility transformations when two E-over-A pluralist face the situation corresponding to the original PD above.

Compare this to the case where E-over-G pluralists face a situation with the (given) utility structure PD\*. E-over-G pluralists apply the group-directed utility transformation rule (ix), except when their personal losses would be unbearable. The payoff matrix of their transformed utilities would be as follows:

(PD\* – two E-over-G pluralists)

		Player 2	
		C	D
Player 1	C	6,6	1,7
	D	7,1	4,4

In this transformed game defection is once again the dominating strategy. Similarly, the E-over-G transformation turns the original PD into a game in which achieving the universal co-operation is not as straightforward as it is with the E-over-A utility transformations:

(PD\* – two E-over-G pluralists)

		Player 2	
		C	D
Player 1	C	6,6	1,5
	D	5,1	4,4

In this (transformed) game, known in game theory as the Stag Hunt, the universal co-operation outcome *CC* is again preferred to all the other outcomes by both players. Nonetheless, the E-over-A utility transformation rule does not appear to be a very reliable motivational mechanism: if an E-over-A pluralist cannot be positive that her co-

<sup>65</sup> Straightforward from the perspective of common sense, that is. From the perspective of standard game theory, however, PD\* of E-over-A pluralists instantiates a Hi-Lo game, which is (in)famously irresolvable in standard game theory (since there are two Nash equilibria between which we cannot rationally choose in the context of standard game theory).

player is not an egoist, she may not find it rational to play *C*, since the possible loss is far greater than when playing *D*.

So what is the conclusion one should derive from all the above discussions? Is my thoroughly pluralistic interpretation of Sober and Wilson's E-over-A pluralism a reliable mechanism for producing collectively optimal outcomes? I do not think so, but this does not mean that I would think that any of the other options discussed so far would be a better choice. Firstly, if we follow Hardin and M. Taylor's line of thought faithfully, we should consider only pure PD situations, and not some modifications that are not strictly speaking PD games (cf. Axelrod 1984, 10), such as PD\*. In pure PD games genuine group-directedness in the sense of the transformation rule (ix) might be more reliable than E-over-A pluralism, since a group-directed player will play *C* regardless of what she believes the other player will do. An E-over-A pluralist, on the other hand, would play *D* if she were led to believe that the other player is going to play *D*. Furthermore, the E-over-G pluralism, which appears to be the most intuitive option in the light of the reflections of our own motivational mechanisms, does not do very well in theoretical models. It seems that we have only bad options available.

I hope that what I have said above shows that the present line of thought should in fact be seen as a kind of *reductio ad absurdum* treatment of the whole programme building on utility transformation rules. When the game theoretical models are employed in testing hypotheses about motivational mechanisms in the sense of utility transformation rules, it seems the even the slightest modification of the (given) utility structure turns the corroborating support from one hypothesis to another. Similarly, by playing with the different interpretations of pluralism the original game can be transformed such that the problem vanishes. One cannot but feel that the successful utility transformation rules are more or less bound to be tinkered to fit the particular model at hand, *i.e.*, to be quite *ad hoc* in nature.<sup>66</sup>

However, our task was to find a psychological mechanism that guides agents to collectively optimal outcomes across a range of social dilemma situations. This is something the context-specific (or indeed case-specific) transformation rules cannot deliver. Thus, I think that although Sober and Wilson's rhetoric as well as our general intuitions suggest that sometimes we have ultimately group-directed motives,<sup>67</sup> insofar as the dis-

<sup>66</sup> Compare, *e.g.*, Routledge (1998, 98-99) who demonstrates how "[a]ltruism may resolve the dilemma in some situations but not others [...] the amount of altruism required to prevent all PD's is extremely precise."

<sup>67</sup> In particular I think that any acceptable solution must be able to accommodate the E-over-A (or E-over-G) idea that although we are sometimes willing to act for the common good, egoism can override such solidarity if the individual cost is too high. Crucially, the we-mode solution I favour is not only much

cussion proceeds exclusively in terms of the directedness of our desires, the discussion cannot help with the fundamental problem I am addressing in this Chapter. Utility transformation rules cannot give us stable collectively rational behaviour in social dilemma situations such as PD and PD\*.<sup>68</sup>

### II.2.3 BEYOND PHILOSOPHICAL EGOISM

There is still a deeper reason why a fundamental change is needed in the way we approach the problem of the psychological motivations underlying co-operative, collectively rational action. Both Sober and Wilson's discussions on the directedness of our ultimate motivations and Tuomela's models of utility transformations rest on a pair of debatable assumptions. The first is that the given utilities represent some external rewards towards which an agent can, in accordance with her psychological inclinations (captured by the relevant transformation rule), form a new attitude. The second is that the formed attitude can transform the given utilities into effective utilities, *i.e.*, the agent is not compelled to act upon her given utilities. In the present context, however, the whole idea of transforming players' utilities simply begs the question.

Tuomela's utility transformation rules can be informative when used for explicating different possible psychological inclinations,<sup>69</sup> but when one uses them for substituting a new game for the original game, one falls guilty of refusing to face the important theoretical problem represented by the original game. What we need to tackle is a PD game of effective, transformed utilities. All relevant altruistic, group-directed etc. considerations must be included in the utility structure that forms the PD.<sup>70</sup>

This point is put well by Martin Hollis (1996), who argues that the reason why individually rational choices do not sum up to a collectively preferred outcome in social dilemma situations is the *philosophical egoism* inherent in individualistic rational

---

more general than the case-specific utility transformation rules but it also caters easily for the possibility of egoism taking over in some situations (this is discussed in the context of Tit-for-Tat strategies below).

<sup>68</sup> See Verbeek (2002, 86-98) for a similar result.

<sup>69</sup> See, for example, Kollock (1998), who uses utility transformation models to interpret empirical studies of collective action dilemmas.

<sup>70</sup> This, I think, is also the core of Binmore's (1994, 80 & 180 ff.) criticism of Gauthier's (1986) programme, according to which individuals can rationally accept principles that constrain their utility maximisation, resolving thereby PDs in a collectively optimal way. Binmore thinks that such considerations should be included in the payoff matrix that forms the PD. I think that this criticism applies, *mutatis mutandis*, also to Frank's (1988) programme insofar as Frank's commitment model is thought to solve the fundamental theoretical problem. As an explication of (an aspect of) the evolutionary role of emotions Frank's view may well succeed – especially if the group selection structure inherent in Frank's models is made explicit. However, I will not enter into these debates in this dissertation. See, however, Section II.2.4 below, where I analyse McClennen's argument from the long-term perspective, which is closely related to the views of Gauthier and Frank.

choice theory. By philosophical egoism Hollis means a view “that the only desires which can move us are our own” (Hollis 1996, 6). That is, my desires can be self-directed, other-directed, group-directed or what have you, but to have an effect on my behaviour they must be my personal desires. Hollis’ philosophical egoism is not a moral or psychological view, but is rather meant to express the ontological thesis that although individual agents may care for the welfare of others, individuals are still inescapably the only units of action there are in social life.

Hollis argues that even if we can come up with a sophisticated model of utility transformations that converts any PD situation agents may face into a harmless coordination game, this is not an answer to the real problem we should be confronting. The egoism that causes the problems in social dilemma situations is not psychological but philosophical. In other words, regardless of the psychological directedness of the motives of agents (regardless of what utility transformation rules they use), it is possible that their motives lead the agents to a situation representable as a PD game. “There are Prisoner’s Dilemmas for altruists as well as for egoists. Also a maxim commanding unselfishness can paralyse everyone, if everyone follows it. (Try cooking dinner for a group of relentless altruists with different tastes.)” (Hollis 1996, 76-77.) We may add that there are PD situations also for group-directed agents, at least insofar as they see the situation a bit differently from each other (*i.e.*, insofar as they do not possess perfect knowledge of the situation).

Some writers seem to miss this point. Singer (1981, 47), for example, appears to think that agents who care as much for the interests of others as they care for their own interests can resolve PD situations in the collectively optimal way. Singer’s view amounts simply to favouring the transformation rule (ix). The technical motivation for this view stems from the fact that if the players agree about the given utilities, then the rule (ix) eliminates the conflict between the players. In addition to the problems discussed in II.2.2, Singer’s view presupposes that the agents have perfect information about the situation and about the preferences of the other player (and that inter-agent utility comparisons are seen as unproblematic), and hence I doubt if this solution can be very reliable.<sup>71</sup> As Barry Barnes (2000, 56) emphasises, mere individualistic utility cal-

---

<sup>71</sup> Actually, it is not obvious that even perfect information would help here. After all, there are no numerical utilities out there in the world. Thus it is not obvious that there is *the* correct way of representing a given situation in terms of utilities (or of comparing those utilities), and hence it seems to be conceivable that even group-directed agents with perfect information about the situation they face could find themselves in a PD.

culations – no matter how benevolent they are – simply cannot be sufficiently reliable for the construction and maintenance of co-operative practices.<sup>72</sup>

First of all, the group-directed solution seeks to circumvent this problem by refusing to discuss the problems involved with perfect information and inter-agent utility comparisons. Unfortunately, for a naturalistic (and realistic) theory of rational co-operation it is not sufficient to refrain from appeals to collective minds; also Platonist appeals to shared pieces of perfect information must be resisted. Real agents do not have access to God's point of view, and hence the possibility of a PD situation remains. Moreover, even if for sake of argument we bypass the problems of perfect information, the group-directed solution, as Bruno Verbeek puts it, "helps only in a small class of collective good problems and is very unhelpful in another class of problems" (Verbeek 2002, 97) and hence cannot offer the kind of general solution I am after.<sup>73</sup>

Second, as Barnes (1995) emphasises, we are not primarily interested in artificial mathematical models, but real cases of collective action that instantiate free-rider problems. Barnes' example is "an individual deciding whether or not to purchase a catalytic converter to purify her car exhaust and contribute to the provision of unpolluted air" (Barnes 1995, 27). The collectively optimal solution is that everyone acquires a converter. However, while the cost of purchasing a converter may be relatively high, the impact a single converter has on air quality remains negligible. Thus, the directedness of our motives is irrelevant here. Purchasing a converter is irrational regardless of whether motivations are self-directed, other-directed or group-directed: "To make a negligible difference to the air benefits nobody" (Barnes 1995, 29), and therefore (individual-mode) considerations of benefits, no matter whose benefits they are, cannot rationalise collective action. Barnes conceptualises this by stating that "the problem is not that [...] individuals are *self-regarding* but rather that they operate *independently*" (Barnes 1995, 29). In other words, the fundamental problem of social action is not created by psychological egoism as such but by Hollis' philosophical egoism.

Hollis urges us in the name of intellectual honesty to take the original PD to represent the final, effective utilities of players 1 and 2 in a social dilemma situation, re-

---

<sup>72</sup> Barnes (2000) argues that in order to explain stable co-operation we need to assume the reality of irreducible *collective agency* (and thus overcome the philosophical egoism of Hollis). Below I argue that the theory of collective intentionality does precisely this. Barnes' views on collective agency are discussed in detail in Part III.

<sup>73</sup> I thank Damien Fennell, Raul Hakli and Govert den Hartogh for pushing me to explicate my reasoning concerning the problems with the group-directed utility transformation rule. They all pointed out that mathematically speaking my claim that there are PD situations also for group-directed agents is false. I trust that what I have said above shows that when we move our focus from the idealised world of abstract models onto the world of real agents with limited (individualistic!) perspectives, my claim still holds.

ardless of whether we think that human agents are egoists, altruists or group-directed agents. Hollis' philosophical egoism maintains that insofar as we are dealing with individual agents that make rational choices from their own individual perspectives, the possibility of facing an irresolvable PD situation will remain. There is no guarantee that there will always be present a benevolent invisible hand securing that agents making individually rational choices will arrive at a collectively rational outcome.

Since people nevertheless quite often manage to achieve collectively rational outcomes in social dilemma situations, the conclusion Hollis (1996, 5) feels compelled to draw is that often human action simply is not instrumentally rational. Hollis (1996, 6) appears to be saying, just as was the case with Tuomela's utility transformation rules, that the rational choice theory can no doubt give us models that are useful tools in some contexts, but as a general theory (or as *the* general theory, as its adherents all too often seem to think) of human sociality and the driving force of social reality it cannot succeed. According to Hollis, truly social action simply is essentially irrational. According to Hollis' irrationalism, co-operation is highly characteristic of human social action, but it cannot be rationalised in instrumentalist terms.

It seems to me that, for example, Verbeek (2002) agrees with Hollis here. The conclusion Verbeek draws from considerations largely analogous to my criticism of utility transformation rules is that to solve the design problem of agents capable of collectively optimal action in social dilemma situations we must assume the existence of a non-psychological disposition that simply *causes* us to behave altruistically in social dilemma situations *despite* our rational deliberations pulling to the opposite direction. Verbeek calls such behavioural dispositions co-operative virtues.

While I appreciate the motivation behind Verbeek's causalism and Hollis' irrationalism, I cannot accept their conclusions: If Hollis and Verbeek are right, truly social behaviour cannot be rationalised at all, and thus there can be no *intentional* (rationalising) explanations of social behaviour. This, in turn, implies that there are no truly social *actions*: social behaviour consists of meaningless bodily movements, not intentional (rationalisable) actions (cf. Part III).

However, by moving to discuss at the psychological level I have already expressed my commitment to the view that co-operation in social dilemma situations is the paradigmatic example of truly social *action* and thus must be explained and understood in terms of contentful, intentional psychology. I cannot follow Verbeek and explain individual action in terms of intentional psychology, and social action in terms of non-psychological dispositions. Crucially, we all *know* on the basis of first-hand experience



that sometimes co-operative action requires a considerable amount of rational deliberation. This is a feature Verbeek's dispositional solution cannot capture, for co-operative virtues are explicitly seen as opposed to rational deliberations. Indeed, this point implies also the unacceptability of Hollis' irrationalism, for, in Barnes' words, "[c]ollective action is scarcely well-described as irrational, since it may be exquisitely calculated and highly effective instrumental action, but it cannot be rationalised by reference either to altruistic or self-interested [or indeed group-directed] individual goals" (Barnes 2000, 57).

Thus we find ourselves in a difficult situation indeed. The utility transformation rules cannot save the solution basing on individual-level considerations of instrumental rationality, but we cannot follow Verbeek to straightforward dispositionalism or Hollis to irrationalism either. However, according to Hollis' philosophical egoism these are the only options we have.

Hence, I think we must make room for an acceptable solution by rejecting Hollis' philosophical egoism. Since social actions cannot be rationalised in individualistic terms, perhaps the rationality should not be attributed to individual agents, but to the collective of agents from whose perspective the social interaction appears as instrumentally rational. Maybe the true agent of social action is not the individual, but the group. Hollis does not even consider this kind of solution, since in his view the rejection of philosophical egoism amounts to a commitment to the existence of a group mind or something ontologically equally dubious.

However, it is a mistake to think that an *ontological* thesis denying the independent existence of group minds implies *methodological* individualism in the sense of the thesis that all actions of individuals must be individual-mode actions. The analysis of the ontological structure of social facts in Part I led us to construct a modern, ontologically acceptable naturalistic interpretation of the Durkheimian notion of a collective consciousness, and the discussion concerning the psychology of social action in this Part has pointed out exactly the same need. Not so surprisingly, my suggestion is that indeed the notion of collective we-mode intentionality of Part I provides just the kind of solution we need also in the present context.

To put it in Gilbert's terminology, I argue that in order to achieve the collectively preferable outcome in social dilemma situations the agents must overcome their individual perspectives and form a plural subject (or, as Barnes (2000) puts this, only agents capable for collective agency can perform truly social actions). This, I believe, is the key for making the required *methodologically holistic* move of overcoming Hollis'

philosophical egoism without rejecting the ontologically *naturalistic* underpinnings of his view. In what follows I argue that the collective we-mode of truly social actions and attitudes is not only the most plausible psychological mechanism for producing collectively rational outcomes in social dilemma situations, but also that such a mode is ontologically perfectly acceptable. The collective we-mode, therefore, is the most plausible and probable solution to the evolutionary design problem of reliable production of evolutionarily altruistic, group-benefiting behaviour.

#### II.2.4 COLLECTIVE INTENTIONALITY AND CO-OPERATION

The basic idea is very simple. Let us not concentrate so much on the directedness of the motives of agents in a social dilemma situation, but rather on the mode in which those motivations are held. Since we have seen that individualistic psychology cannot reliably account for collectively rational action, it is quite natural to think that the most reliable method for achieving that outcome must be to judge the rationality of different outcomes from the point of view of the whole collective involved, *i.e.*, adopting a collective perspective to the situation. This amounts to (i) choosing the collectively optimal outcome to form the collective goal of the group, (ii) figuring out what combination of individual actions realises that outcome and (iii) performing the derived individual actions. A process following the steps (i)-(iii) is, of course, suitable for dealing with Barnes' (1995) excellent example of purchasing expensive catalytic converters. Moreover, (i)-(iii) amount to the standard characterisation of collective we-mode intentionality as presented in Part I.<sup>74</sup>

Crucially, this requires us to postulate neither unnatural social entities nor group minds to form the collective-level plan. The only requirement is that ordinary individual agents have a tendency to overcome their individual perspectives and adopt a collective stance, and that this tendency is (sometimes) activated in social situations. If, that is to say, I adopt a collective we-mode attitude towards a social (dilemma) situation, I will consider myself primarily as a member of a we-group and reason first what *we* should do in this situation, and only then derive my own individual task *qua* a group-member from the group-level plan. This is not a form of irrationality, although it is not a form of

---

<sup>74</sup> In other words, the theory of collective intentionality in my sense agrees with Elizabeth Anderson's reconstruction of Amartya Sen's criticism of standard rational choice theory. Anderson argues that truly social action, exemplified paradigmatically by co-operation in Prisoners' Dilemmas, requires that agents *identify* themselves with their social groups and consequently ask, not "What should *I* do?", but rather "What should *we* do?" (Anderson 2001, 28). For the link between collective intentionality and the theory of social identification, see II.3.3 below.

individual-level instrumental rationality either. The rationality involved is *collective* rationality, although nothing ontologically suspect has been assumed. Collective we-mode intentionality is a question of a psychological framing of a social situation.<sup>75</sup>

If the players in a social dilemma situation adopt collective we-mode stance to the situation, it is relatively easy for them to infer that *CC*<sup>76</sup> is indeed the collectively rational outcome they collectively should aim to realise. Hence, they will both deduce that what each of them *should* individually do *qua* a group member is *C*. Clearly collective we-mode psychology is a more reliable psychological mechanism for delivering socially optimal behaviour than altruism or group-directedness. This claim agrees with Searle's (1990, 406) assertion that taking up the we-mode is a precondition of all true co-operation and with Gilbert's (1989) insistence that the willingness to create a plural subject is the starting point of all truly social behaviour. Moreover, as for example Michael Tomasello and Hannes Rakoczy (Tomasello & Rakoczy 2003, 127) argue forcefully on the basis of their empirical work, precisely this kind of essentially social "derived normativity" of individual actions is "a key characteristic" in the architecture of the human cognition.

Moreover, a mode-pluralist can alternate between the we-mode and individual-mode in a way that is not possible for a mode-monist. In particular, a mode-pluralist can act against her given individual utilities without transforming them. The given utilities are indeed final, but only final individual utilities, and hence *irrelevant* when the mode-pluralist has switched to the we-mode. In a sense the individual-mode considerations are present in the motivational basis of social action, but in suitable circumstances an agent

---

<sup>75</sup> Sugden (2000) gives a very similar description of collective reasoning. However, Sugden insists on building on the *directedness* of intentions, which is an approach I have rejected. However, Sugden may well be using "directedness" in a sense different from mine. In any case – due to his concentration on coordination games – Sugden does not explicate the important distinction between directedness and mode. Since I am interested mainly in games of conflict, the distinction is needed, for the central aspect of the proposed solution is that we give up Hollis' philosophical egoism – and social directedness remains insufficient for this. It seems to me that Hans Bernhard Schmid (2004, 2005) is developing a view similar to the one defended here when he argues that the theory of collective intentionality can be employed to make sense of Amartya Sen's (1977 & 1985 in particular) perplexing claim that sometimes a person's rational choices are not based on the pursuit of her own goals – in other words, Sen seems to agree that true co-operation presupposes a rejection of Hollis' philosophical egoism. A further closely related account is that of Carol Rovane (1998). Rovane argues that there is no *a priori* reason to restrict rationality considerations to the perspective of a disparate individual. Rovane uses this insight to argue for the possibility of group agents, whereas I opt for a more modest attempt to argue that individual agents may adopt a group perspective without becoming committed to the perplexing view of seeing groups as persons (cf. Part I). The we-mode group perspective view allows us of course to talk *as if* groups really were agents (Tuomela 2007).

<sup>76</sup> Actually this gets the conceptual order wrong. Strictly speaking we-mode agents do not start by characterising the collectively optimal outcome in terms of (a sum of) individual choices (as the term "*CC*" obviously does). Rather, figuring out that *CC* – the relevant combination of individual choices – allows them to realise their collective goal is a further problem social agents must solve in order to be able to derive their individual tasks in the collective project.

can switch to the we-mode and ignore her individual-mode utilities. This leaves open the possibility that if the we-mode considerations are in too strong a conflict with the individual-mode considerations, the agent will deviate from the we-mode and act on her individual-mode utilities and preferences. Thus, the we-mode view does justice to the core intuition behind Sober and Wilson's E-over-A pluralism.

Moreover, we-mode collective attitudes provide exactly the kind of psychological mechanism we need for a reliable production of, for example, the Tit-for-Tat strategy that was seen (II.1.1; cf., *e.g.*, Axelrod 1984 and Singer 1994 & 1999) to be such a successful strategy at the behavioural level. When an agent faces a social dilemma situation and realises that she is going to play it repeatedly with the same partner, it is quite natural for her to frame the situation as a collective task she and her partner face *together*. Hence, the agent may be led to adopt the we-mode approach and, consequently, to believe that she and her partner form a plural subject that will rationally choose the universal co-operation outcome *CC* as their joint goal. Thus the agent can deduce that her individual role in the collective task is to play *C*.

If the other player, however, defects and plays *D*, it is reasonable from the point of view of the co-operative agent to infer that she was wrong about the situation: her beliefs about the players forming a collective intention together – putting the *game* behind them – were not true. Thus, she thinks, there is no plural subject present after all, but the players face the situation qua individual agents. This amounts to saying that the agent's we-mode psychology collapses and she falls back to the perspective of individually rational strategic choices and, accordingly, plays *D* in the following round.<sup>77</sup> Had the other player answered the co-operative move by playing *C* as well, the stability of the first agent's we-mode attitudes would have been strengthened, and she would have had all the more reason to keep playing *C*. Her beliefs about the existence of a plural subject would have been confirmed.

Hence, a mode-pluralist can have the best of both worlds: she can avoid the problems created by Hollis' philosophical egoism without rejecting (instrumentally) rational considerations. It is just that the rationality in question is not always individual-level rationality. Sometimes the relevant rationality takes the form of collective rationality, contemplated in the we-mode. Crucially, this whole process does not involve any transformation of the original individual utilities. Thus the introduction of the collective we-mode perspective is not an *ad hoc* move of changing the game when drawn against a

---

<sup>77</sup> In real life a player's we-mode may of course be so strong that it endures a single play of defection from the part of the other player. This depends on the particular features of the situation at hand, personal histories of the players involved *etc.*

difficult social dilemma situation. Adopting a we-mode attitude means simply adopting a new collective perspective to the same old game. I conclude that the assumption that agents register their we-mode considerations as the ultimate considerations that rationalise action in social dilemma situations is the most reliable, and hence the most plausible, solution to the design problem of getting *rational* individuals to co-operate in social dilemma situations, for there are no intragroup Prisoners' Dilemmas for we-mode agents. Social action is rationalisable (and, hence, intentionally explainable – cf. Part III) only in collective we-mode terms.

Moreover, the present discussion allows me to pin down what, exactly, I take the core of the theory of collective intentionality to be. In short, in my view the crucial step is the rejection of methodological individualism in the sense of Hollis' philosophical egoism. Social life does not amount to the sum of individual-mode interactions. Ultimately, social life and truly social action is not presentable as a *game* of distinct (individual-mode) individuals, regardless of how benevolent one takes them to be. We must accept that truly social action is based on we-mode psychology.

This is the fundamental reason why in my discussions I have not included, for example, Michael Bratman's (*e.g.*, 1999) or Seumas Miller's (*e.g.*, 2001) work as an example of the collective intentionality approach, although Bratman, together with Gilbert, Searle and Tuomela, is widely considered as one of the major contributors in the field (see, for example, Deborah Tollefsen's (2004) influential review of the collective intentionality approach).<sup>78</sup> Contrary to Gilbert, Searle and Tuomela, Bratman and Miller seek to account for social action exclusively in terms of *individual-mode* attitudes that are collective only in terms of their content and directedness – and thus they operate firmly within Hollis' philosophical egoism. Hence, whatever virtues their accounts may have in other contexts, they cannot help with the problems I am addressing in this Part. Durkheimianism is required.

At this point I need to discuss one more way of arguing against the form of pluralism and methodological holism defended in this Chapter. This line of thought is motivated by the difference between short-term and long-term self-directedness. This idea is also sometimes (*e.g.*, Skyrms 1996) conceptualised as the problem of the modularity of rationality, and is defended, for example, by Edward F. McClennen and Scott Shapiro (McClennen 1997 and McClennen & Shapiro 1998).

---

<sup>78</sup> However, Tollefsen (2004) emphasises correctly that unlike those of Gilbert, Searle and Tuomela, Bratman's is essentially an account of *shared*, rather than *collective*, intentionality.

The standard example with which this approach can be illustrated – discussed also by Sober and Wilson – is the problem of quitting smoking. Imagine a person who is a regular smoker. She very much enjoys each cigarette she smokes, but knows also that in the long run her smoking can cause fatal health problems. The dilemma the smoker has is the following: Each time she considers lighting a new cigarette, she knows (i) that were she really to light the cigarette, she would gain a considerable short-term benefit and, moreover, (ii) that any single cigarette – including the one she is now about to light – will have only a negligible impact on her health (note the structural similarity to Barnes' converter example). If the smoker cares only about short-term benefits, she will light up the cigarette, whereas if she cares enough about the long-term quality of her life, she will stop smoking immediately and, hence, not light up. Put differently, she will voluntarily commit herself to a rule "do not smoke". The point to note here is that the two possible, explicitly opposite strategies of smoking and not smoking are both based on self-directed individual-mode motivations (Sober & Wilson 1998, 227).

The suggestion here is, of course, that "a strategic decision problem [one faces] with one's own future self [...] [is] a problem analogous to that faced by a person who interacts with other (rational) selves, who may have different preferences with respect to outcomes" (McClennen & Shapiro 1998, 365; see also McClennen 1997, 216). Hence, perhaps psychological egoism can deliver collectively optimal solutions to social dilemma situations after all. Perhaps a motivationally self-directed agent needs not to transform the utilities of the social dilemma game she faces, but just to realise that they are the utilities as judged from the standpoint of her short-term preferences, and the same situation can be presented as a different game when judged from the perspective of her long-term preferences. In a way, this would allow the agent to switch her perspective to the game in a manner similar to how I argued it is possible to alternate between individual mode and we-mode. Maybe the agent could reason that although she would maximise her immediate benefits by defecting, in the long run it is better for her to establish a reputation as a social and nice person or, alternatively, that in the long run mutual co-operation will be better for everyone.

However, this line of thought is based on a failure to see *philosophical* egoism in Hollis' sense as the source of problems in social dilemma situations. In some actual cases the long-term considerations may lead to co-operation but, as I have shown, this solution does not help us at all with the theoretical problem we ought to concentrate on. The final utilities that form the PD the agent faces are assumed to include the agent's evaluations of both the short-term considerations and the long-term considerations.

Moreover, I have argued (in the context of Tit-for-Tat strategies) that we-mode considerations are both intuitively more realistic and theoretically more plausible candidates for the psychological mechanisms behind collectively beneficial action even in the context of repeated games.

In short, the attraction of the long-term argument in favour of psychological egoism and individualistic instrumental rationality must be based either on considerations about reputation (or something similar), or on considerations that mutual co-operation will in fact lead to a result collectively preferable to universal defection. If the former, the argument may be relevant to some empirical scenarios, but remains inescapably powerless in front of Hollis' philosophical arguments. If the latter, it is an argument in favour of the collective perspective which is naturally combined with we-mode collective attitudes. McClennen is correct to argue that co-operation that overcomes individual preferences in a social dilemma situation can be instrumentally rational, but fails to appreciate that such rationality must be collective in nature.<sup>79</sup>

Indeed, McClennen's (1997, 243 ff.) discussions about the mutually beneficial co-operation in Assurance games and Prisoners' Dilemmas point directly towards the notion of the we-mode. McClennen just feels the need to deny all references to any sort of collective concepts, since he thinks that they inevitably include an appeal to some "notion of a 'communal' or 'collective' self" (McClennen 1997, 243), and he explains that he has no "metaphysical taste" for such notions. In my view this is nothing but the by now familiar mistake of thinking that (*OI*) implies full-blown methodological individualism (Part I) or that ontological considerations necessitate philosophical egoism (this Part). I would like to suggest that with the notion of the we-mode advocated in this study McClennen could have the results he needs while remaining true to his metaphysical taste, which is a taste I wholeheartedly share. Be that as it may, I feel justified to conclude that the argument from long-term considerations fails to challenge methodological holism as defended in this dissertation.

Finally, I want to emphasise that my argument for the success of we-mode collective psychology in social dilemma situations is based on the possibility of we-mode attitudes forming a *sui generis* form of psychology that is on a par with (qualitatively

---

<sup>79</sup> In fact, also McClennen and Shapiro drift to talk about how "each member of a group of interacting persons can often do better by adopting a rule which allows the members of the group to coordinate their actions" (McClennen & Shapiro 1998, 367). McClennen even explicitly builds upon notions such as what is "mutually advantageous to a set of persons who find themselves faced with a problem of interdependent choice" and suggests that his arguments support the rationality of thinking "more holistically about interactions with others" (McClennen 1997, 216). I think the lasting core of this reasoning in general and McClennen's plea for holism in particular are best understood in terms of the theory of collective we-mode intentionality.

different from and irreducible to) individual-mode psychology. In other words, I side with Sugden (2000), who argues that truly collective action simply cannot be accounted for in the individualistic terms of mainstream rational choice theory (henceforth, RCT). Social action does not amount to a *game* of disparate individuals, but to *acting together*.

Thus I am reluctant to adopt Tuomela's (2000, 315) notation, according to which the basis of an individual's decision-making act could be modelled with the following formula:  $EU_i(X) = w_i EU_i(X) + w_g EU_g(X)$ :

Here  $EU_i(X)$  means the total expected utility of a choice  $X$  for a participant in a situation of strategic interaction. The weights, *viz.*, the individual or  $i$ -parameter  $w_i$  and the collective or  $g$ -parameter  $w_g$ , add up to one:  $w_i + w_g = 1$ ,  $0 \leq w_i, w_g \leq 1$ . The factors in  $U_i$  are supposed to be factors viewed from an individual's perspective, whereas those involved in  $U_g$  are factors viewed from the group's perspective. [...] Some special cases and some dependencies between the parameters can be noted. Thus  $w_i = 0$  entails unconditional cooperation and  $w_g = 0$  entails acting on merely personal preferences (which of course may be other-regarding). If a person strictly accepts a goal  $G$  as his collective goal and is fully committed to it, this entails that  $w_g = 1$ . However, he can in a weaker sense take group factors into account while also respecting individual factors. [...] If a participant conforms to the standard game-theoretical dominance principle for  $i$ -preferences, then  $w_g = 0$  for him in a PD. (Tuomela 2000, 315.)

I fear that adopting this presentation translates the we-mode concepts back into the Individualistic Account of RCT, and thus loses the very essence of we-mode concepts (Sugden 2000), making we-mode arguments fall prey to the criticism of Hollis. Insofar as Tuomela's formula is interpreted to model the deliberative decision-making processes of an individual from the point of view of that individual, the appeal to we-mode concepts in the present context would simply beg the question. This is so because under this interpretation the agents facing a social dilemma situation would have to deliberately choose what values they will assign to the parameters  $w_i$  and  $w_g$ . If they both choose  $w_g = 1$ , they will reach a better outcome than if they both choose  $w_i = 1$ . However, the best outcome for each agent occurs when she herself chooses  $w_i = 1$  and the other player chooses  $w_g = 1$ . In fact, no matter what the other player does, each agent can maximise her own benefits by choosing  $w_i = 1$ . Thus, the situation the individuals would face in choosing which kind of rationality to apply would have the structure of a Prisoners' Dilemma. I conclude that a switch to we-mode psychology should not be modelled as a deliberative process.<sup>80</sup>

---

<sup>80</sup> This is not what Tuomela suggests either (personal communication). Hence the point I am making here is not a criticism of Tuomela – but an important clarification nonetheless.



In fact, this is exactly the core of my evolutionary arguments for understanding collective intentionality as a primitive feature: in social dilemma situations we-mode considerations *must* form our ultimate considerations. In particular, they *cannot* be based on more fundamental individual-mode considerations, since that would only activate the higher-order PD, and, hence, lose the reliability.

In this sense my view of collective intentionality as a primitive phenomenon amounts to the Dennettian (see especially Chapter 8 of Dennett 2003) idea of giving up the myth of the Cartesian Theatre. The Cartesian tradition understands the ultimate subject of human action to be the Cartesian self, the metaphysical ego, who sits in the command headquarters of the conscious processes (the Cartesian Theatre) of a human body and monitors all the possible self-directed, other-directed, we-mode etc. tendencies and then makes the decision about action. This, of course, is the core of the philosophical egoism thesis as well. My (and Dennett's) view rejects this picture categorically. In a sense, the self is nothing but those competing tendencies.<sup>81</sup> Sometimes, when the circumstances are suitable, I simply act on my we-mode considerations. Giving up Hollis' philosophical egoism amounts to giving up also the idea of all intentional actions being based on the choices of a metaphysical ego sitting in the Cartesian Theatre.

The conclusion I have arrived at allows us to take one more step on the road towards naturalised methodological holism. Part I argued that the anti-individualistic theory of collective we-mode intentionality is naturalistic in the sense of being ontologically perfectly acceptable. Now we can say that the theory is naturalistic also in the more specific sense of being supported by our understanding of evolutionary dynamics. Moreover, in addition to appeals to intuitive plausibility that dominate the collective intentionality literature, my evolutionary considerations form a strong, independent and general argument in favour of irreducible we-mode psychology.

---

<sup>81</sup> The nature of agency is analysed in detail in Part III. For an explication of what, exactly, my line of thought implies *vis-à-vis* the self and personhood, see Saaristo (2004a).

CHAPTER II.3:  
COLLECTIVE INTENTIONALITY AND  
EMPIRICAL SOCIAL SCIENCE<sup>82</sup>

The theory of collective intentionality in my sense is motivated by conceptual problems faced by attempts to provide individualistic analyses of (i) social practices constituting social facts in general and (ii) collectively rational action in social dilemmas in particular. The claim is that individual-mode intentions, even when combined with mutual beliefs or common knowledge concerning the intentions, will not add up to the strong sense of intending and acting *together* that is required for truly social action. Hence, irreducible collective intentionality is needed.

However, arguably the theoretical and philosophical arguments in favour of the theory of collective intentionality can only take us to the point where we can see that the theory is *plausible*, it makes *sense*, and that we have good *reasons* to think that collective intentionality most probably is an irreducible part of human intentionality. Hence, the theoretical and philosophical arguments in favour of the collective intentionality theory ought to be strengthened by empirical studies examining whether human social behaviour in fact shows evidence of collective intentionality. In what follows I shortly review certain empirical approaches to social action. I argue that even if such studies cannot settle the issue watertightly (cf. III.5.3), together with my philosophical arguments they make a strong case for the theory of collective intentionality.

### II.3.1 SOCIAL SANCTIONS

One reason for thinking that I am exaggerating the importance of we-mode social action might be the following. In most societies actions that have direct welfare consequences for others tend to be governed by more or less explicit norms and rules. Such norms typically reward socially beneficial actions and sanction egoism (often officially with formal laws when the behaviour is actually harmful to others and with informal social norms when just indifferent concerning the well-being of others). Such norms and corresponding punishments and rewards appear to turn seemingly altruistic action to one that is rationalisable from an egoistic/individualistic point of view. The objection is

---

<sup>82</sup> This Chapter builds largely on Saaristo (2006b).

(e.g., Fehr & Gintis 2007), then, that might not we-mode social behaviour, be uncalled-for in societies where individuals sanction free-riding behaviour of others?

The point to notice is that even if the secondary action of punishing and rewarding forms of primary action transforms the primary action so that it does not require non-egoistic motives, it is the *secondary action* which is now in need of explanation. And it clearly is not explainable from the egoistic standpoint, mainly because of the free-rider problem (see, for example, Barnes 1995, 79). From the point of view of each individual it is not rational to participate in the maintenance of a system of punishing and rewarding regardless of whether the others actually do it.

However, the individual costs of participating in the maintenance of the system of social norms may be relatively small (although even then the effect – reward or punishment – can be very effective). Thus it may be natural to consider the participation as individualistically rational. Evolutionarily speaking, such secondary actions are nevertheless group-level traits:

From the evolutionary standpoint [...] the fact that the cost is trivial does not alter the level at which the behavior evolves. Secondary behaviors evolve *more easily* by group selection than primary behaviors because they are less strongly opposed by within-group selection, but they still evolve by group selection. (Sober & Wilson 1998, 144.)

This means, first, that secondary action – promoting seemingly altruistic primary behaviour – might have evolved by group selection even where the within-group selection force is so strong that the resulting, seemingly non-egoistic primary behaviour could not have evolved by itself. Second, since a system of social norms is evolutionarily speaking a product of group selection, the arguments presented in this Part about the reliable psychological processes underlying evolutionarily altruistic behaviour apply fully also to such secondary actions.

I conclude that the existence of systems of sanctions and rewards that turn collectively beneficial action into individually rational action is not evidence against the pluralistic view of this essay. On the contrary, this line of thought just makes the we-mode view even more plausible. If it is only the secondary actions, the cost of which is rather small even from the egoistic perspective, that presupposes collective attitudes, it is all the more understandable that agents do not in general deviate from the collective standpoint to egoistic free riding. With this possible objection now cleared away, I can next turn to more complicated empirically orientated arguments.

### II.3.2 THE DISCOVERED PREFERENCE HYPOTHESIS

One allegedly empirically motivated line of thought that could be seen as an objection to the we-mode view of this essay comes from experimental economics. Economists studying human action sometimes subscribe to the so-called *discovered* preference hypothesis (the label is taken from Plott 1996). The idea can be seen (cf. Hausman 2000) as a reaction to the earlier doctrine of a *revealed* preference hypothesis. However, part of my argument will be that, rather surprisingly, *vis-à-vis* the present problem the discovered preference view actually fails to overcome the limitations of the revealed preference view. Therefore it is useful to start with the revealed preference view and then move on to the discovered preference thesis.

The motivation behind the revealed preference view is the behaviouristic conviction that a scientific perspective to human action must not engage in folk psychological speculations about what is going on in the minds of the agents. Rather, scientists must study actual choices as they become manifest in observable external behaviour. The preferences behind the behaviour are then reconstructed theoretically on the basis of observations. In this sense the behaviour of an agent is thought to *reveal* unambiguously the (individual-mode) preferences of the individual agent in question. An agent does what she prefers to do in that situation and, consequently, observations of behaviour are thought to provide direct empirical evidence concerning the preferences behind the behaviour.

Although in the spirit of Hollis' philosophical egoism it is no doubt quite tempting to interpret a choice made by an individual as an obviously individual-mode choice, the interpretation is based on a plain *stipulation* that whatever an individual chooses to do, the choice is based on the agent's individual-mode considerations. As it is usually set up, the framework of the revealed preference hypothesis simply defines *a priori* all action and intentionality to be in the individual mode. Thus, the word "revealed" is rather misleading here. The economists subscribing to this approach simply construct a theoretical model for explaining and predicting behaviour. The psychological-sounding concepts, such as preferences, desires and beliefs play a purely instrumental role in the model, for the whole idea is not to speculate about mental states. Obviously, however, this kind of instrumentalism cannot be used as an argument in the debate concerning the nature of real psychological mechanisms. In short, the revealed preference hypothesis does not form a relevant argument for the present discussion, which is about the psychology of real agents, not about the models of dogmatic economists.

However, some economists tend to drift away from the instrumentalist and rather simplistic standpoint of the revealed preference hypothesis and argue that results of economic experiments<sup>83</sup> do provide *a posteriori* evidence for the view that all deliberation is done in the individual mode when we acknowledge explicitly that the preferences of an agent are not constructed out of factual choice behaviour. This acknowledgement is conceptualised as a move from the revealed preference hypothesis to the discovered preference hypothesis. This is the direction in which, for example, Charles R. Plott (1996) and Ken Binmore (1999) want to take experimental economics.

In short, the discovered preference view holds that “each individual has coherent preferences, but these preferences are not necessarily revealed in decisions” (Cubitt, Starmer & Sugden 2001, 386). The individual must *discover* what her preferences in fact require her to do in a given situation. This may require a lot of time and effort (information gathering, deliberation, learning from experience and so on), but only when this process is complete will the behaviour of the agent reveal the true (discovered) preferences. It would be a crucial mistake to think that behaviour prior to the completion of the discovery process reveals real preferences. A characteristic example of the discovered preference approach is experimental work on the Ultimatum Game.

It is a well-known fact that real agents characteristically act irrationally (from the point of view of mainstream rational choice theory) in the Ultimatum Game. Instead of rationally seeking to maximise their own benefits, people all over the world tend to follow an implicit norm of distributive justice (Henrich, Boyd, Bowles, Camerer, Gintis, McElreath & Fehr 2001) that is rational only from the collective point of view. Binmore, however, argues – basing on the empirical experiments he has conducted (see, in particular, Binmore, Shaked & Sutton 1985) – that if the Ultimatum Game is repeated in an experimental setting such that agents are allowed enough time to think about the activity and learn from experience, it will be noticed that the agents will move closer to the behaviour predicted by the theoretical models built on the assumption of self-directed individual-mode motivations (Binmore 1999, F20).<sup>84</sup>

---

<sup>83</sup> I can of course concentrate only on a small aspect of experimental economics. For a general overview of the field, see, for example, Kagel & Roth (1995).

<sup>84</sup> Binmore’s (1999) interpretation of the results of Binmore, Shaked & Sutton (1985) is, however, controversial (see Binmore 1999, F20 where Binmore explains his interpretation paraphrased above). Many theorists think that unless the players are given explicit guidance (which Binmore *et al.* in effect did), the phenomenon of actual behaviour in the Ultimatum Game moving towards the theoretical predictions is never observed (see, *e.g.*, Henrich *et al.* 2001). Naturally, if Binmore’s description of the results is not warranted, the results do not challenge the theory of collective intentionality at all. In this section I aim to show that *even if* Binmore’s description is accurate, it does not automatically count against the theory of collective intentionality. I thank Joseph Henrich and Natalie Gold for discussions on this point.

It might be tempting to interpret Binmore's results as suggesting that our actions are indeed governed by individual-mode deliberations. Perhaps some actual situations just are so complex that real agents find it difficult to figure out what rationality, contemplated in the individual mode, requires them to do. Repeated games in the "pure" circumstances of Binmore's laboratory would then show that agents indeed gradually *discover* what self-directed individual-mode rationality calculations actually dictate in the situation at hand and behave accordingly, although their actions have all the time – even when manifestly different from theoretical predictions – been motivated by the same type of considerations. This would save the standard rational choice interpretation of instrumental rationality as individual-mode benefit maximisation from the apparent falsification by the empirical fact that real people tend to co-operate even when it is not rational from their individual point of view (not even from the perspective of their long-term considerations). This, of course, is precisely the goal of Binmore's argumentation.

However, I think that this is not a decisive case for the present problem of the nature of human psychology. It is quite possible to re-describe Binmore's results to fit the theory of collective intentionality. Note that the collective intentionality theory does not deny the existence of individual-mode rationality considerations. It is just that in social life agents are often able to overcome the individualistic perspective and act in accordance with collective rationality, *i.e.*, in the we-mode. Binmore's experimental settings could be interpreted as special circumstances that lead agents, contrary to "normal" social surroundings, to give up collectivistic considerations and to follow individual-mode rationality. Indeed, the empirical data (e.g., Henrich *et al.* 2001) suggests that people rather universally tend to approach social situations such as Ultimatum Games as the collective intentionality theory would predict, but with sufficient training they can be *taught*<sup>85</sup> to overcome their social tendencies and act in the individual mode.

Perhaps agents have natural tendencies for both collective we-mode action and individual-mode action. If so, no doubt it is possible to create circumstances that encourage or even require exclusively individual-mode strategic action. In such circumstances the individual-mode tendencies would be activated and we-mode tendencies suppressed. If this is the case, then Binmore's experiments do not tell us the whole truth of human psychology, even if it is true to say that in an environment encouraging individual-mode considerations agents are indeed capable of modifying their thinking and

---

<sup>85</sup> Thus, the crucial question is whether the process Binmore describes is a process where an agent *discovers* her true preferences or where she is being *taught* new preferences (cf. Footnote 80). A full answer to this question presupposes a general theory of intentional action, and hence it must wait until Section III.5.3.

behaviour accordingly, *i.e.*, capable of acting exclusively on their individual-mode considerations.

Note also that if this motivationally pluralistic interpretation of Binmore's results is correct, then the misinterpretation of the experiments that treats the emerging individual-mode picture as the complete picture of human psychology is not only unjustified, but it also has potential for rather unfortunate practical consequences. If the model of human agency we have in mind when designing our social institutions is that of individual-mode self-directed agents, we may end up creating circumstances where agents are required to act and think egoistically in order to be successful in their activities, although the agents would have had the capacity to collectively rational we-mode action as well.

The idea I am applying here is, to use Ian Hacking's terminology, that the categories of the human sciences are *interactive*, since "people [...] can become aware of how they are classified [or how rationality is characterised] and modify their behavior accordingly" (Hacking 1999, 32). A fitting example of this is Robert H. Frank, Thomas Gilovich and Dennis T. Regan's (1993) well-known study on how studying economics tends to transform the behaviour of students of economics to conform with the standard economic models that assume self-directed individual-mode motivational monism. Hacking calls this *the looping effect* of social scientific theorising. Social science does not simply describe mind-independent reality; rather, since social reality consists (partly) of the beliefs and attitudes of individuals, social scientific theories may transform their own object (C. Taylor 1985). Therefore, misunderstandings at the level of philosophical psychology may lead to unattractive consequences at the level of social reality by modifying human behaviour into an undesirable direction.

However, I shall not elaborate on these enormously important themes presently. Part III returns to these issues in more detail. For now it suffices to notice that the work on experimental economics cannot provide the kind of unambiguous empirical evidence against the collective intentionality theory that I am after. First, the instrumentalist framework of the revealed preference hypothesis is unsuitable for providing empirical arguments concerning the true nature of human psychology. Second, when experimental economists seek to move beyond instrumentalist model-building of the revealed preference view to the discovered preference hypothesis, their results can typically be interpreted as compatible with the pluralism of the theory of collective intentionality. Indeed, it seems to me that to defend the individualistic interpretation of his experimental results Binmore must return to the *a priorism* of the revealed preference view and sim-

ply *postulate* that the process of discovering the true preferences may be regarded as complete only when observable behaviour complies with the individualistic theory. But then, of course, the individualistic theory is not discovered empirically to be true. More straightforward empirical evidence is needed.

### II.3.3 THE SOCIAL IDENTITY APPROACH

In order to find empirical studies that target the precise problem at hand, the psychological processes present in social action, it is quite natural to turn to social psychology which, by definition, studies precisely this issue. However, all too often also social psychologists tend to simply assume the atomistic framework and regard social action as essentially similar to individual action (Hogg & Abrams 1988, 3), the only difference being that in the social case the context an agent must take into account is (partly) social. Such an account of sociality is fully graspable in terms of the Individualistic Account (I.1.5) and has no use for the notion of collective intentionality.

Nevertheless, within social psychology there are also research traditions that have reservations concerning the dogmatic acceptance of individual-mode social atomism. In this paper I concentrate on one such tradition, the so-called *social identity approach*.<sup>86</sup> Michael A. Hogg and Dominic Abrams describe it as follows:

The central tenet of this approach is that belonging to a group (of whatever size and distribution) is largely a *psychological* state which is quite distinct from that of being a unique and separate individual, and that it confers *social identity*, or a shared/collective representation of who one is and how one should behave. It follows that the psychological processes associated with social identity are also responsible for generating distinctly ‘groupy’ behaviours, such as solidarity within one’s group, conformity to group norms, and discrimination against out-groups.

(Hogg & Abrams 1988, 3.)

The three core elements of this theory are captured well by the title of John Drury and Steve Reicher’s article “Collective Action and Psychological Change: The Emergence of New Social Identities” (Drury & Reicher 2000). The social identity approach holds that truly collective action is such that it involves a psychological change in the agents performing the action. The change is taken to be that of *social identification*. Hogg and Abrams (1988, 7) emphasise that an individual’s identification with a social collective is a psychological state that is “very different from” or even

---

<sup>86</sup> For a concise history of the approach, see, *e.g.*, Brown (2000).



“quali[ta]tively distinct from” standard individual-mode psychology.<sup>87</sup> In social action agents identify themselves with a collective, forming thus a social group, and “[t]hese processes create identity and generate behaviours which have a characteristic and distinctive form, that of group behaviour” (Hogg & Abrams 1998, 17). Moreover, as Hogg and Abrams put it, “social identity theorists posit a *switch* of identity in the group (from personal to social)” (Hogg & Abrams 1988, 153). Drury and Reicher specify this by emphasising that “individuals in crowds do not lose their identity but rather shift from behaving in terms of disparate individual identities to behaving in terms of a contextually specified common social identity” (Drury & Reicher 2000, 581).

My point, of course, is that when the social identity theorists explain what their view amounts to, their descriptions are almost word by word identical with the standard descriptions of the collective intentionality theory. Accordingly, in my view the switch the social identity theorists conceptualise fashionably in terms of altering identities and identifications is nothing else than the capacity to act and deliberate both in the individual mode and in the collective we-mode.

Moreover, Hogg and Abrams (1988, 97-101) appear to share my scepticism concerning models that treat social action and social facts in the manner of mainstream rational choice theory, *i.e.*, merely as a result of a combination of individual-mode strategic choices (I.1.5). Hogg and Abrams argue that both empirical evidence and theoretical reasons “strongly suggest that a social identity analysis may be more profitable” (Hogg & Abrams 1988, 105). By a social identity analysis they mean an approach which interprets true social action as action taken *qua* a group-member (*i.e.*, the individual performing the action identifies herself with the group, and acts, accordingly, in the we-mode). The individualistic essence of the mainstream rational choice theory tends to lead rational choice theorists to interpret co-operative social agents as irrational and, moreover, to posit suppressed drives (or something similar – recall Verbeek’s co-operative virtues) to explain flights from rationality. Crucially, the social identity approach does not appeal to such *ad hoc* explanations.

Similarly, Hogg and Abrams accept the view that sometimes in social situations the rationality followed is collective rationality and thus unselfish, co-operative behaviour can be seen as (collectively) rational. “Rather than depicting collective behaviour as a manifestation of latent impulses, it is considered to result from altered self-conception. Rationality is not so much suspended as changed.” (Hogg & Abrams 1988,

---

<sup>87</sup> Thus, similarly to the theory of collective intentionality, also the social identity theory is essentially an heir of Durkheim’s sociology (Hogg & Abrams 1988, 15-16) – recall Durkheim’s almost identical characterisation of the difference between individual and collective emotions quoted in I.1.4.

136.) In sum, “apparently inconsistent social performances may result from switches between personal and social identity” (Hogg & Abrams 1988, 129). Translate Hogg and Abrams’ fashionable language of altering identities into the language of individual-mode and we-mode action and I could not agree more.<sup>88</sup> Correspondingly, if the social identity theory is well confirmed by empirical evidence, this evidence will also speak for the theory of collective intentionality.

A major empirical motivation for the social identity approach comes from empirical studies of social dilemma situations (Brewer & Schneider 1990, see also Klandermans 2000 and Kerr & Park 2001).<sup>89</sup> In coherence with the philosophical arguments given above, also Brewer and Schneider hold that the atomistic individual-mode dogmatism inherent in most contemporary social science “makes it difficult to account for the high level of apparently voluntary social co-operation evidenced in both field and laboratory studies of social dilemmas” (Brewer & Schneider 1990, 170). They conclude that empirical studies of social dilemma situations support the view that co-operation in collective action dilemmas is due to individuals identifying with their collective<sup>90</sup> – or, in my terminology, to adopting the we-mode. Correspondingly, if the individuals stick to their individualistic perspectives, the mutually beneficial co-operative outcome remains unreachable. In words of Brewer and Schneider:

When the interdependent group is seen as a collective of distinct individuals, individualistic motives are presumed to be modal and self-interest dominates collective welfare. When relevant social identities are activated, however, social motives are subject to transformation reflecting changes in the perceived nature of the interdependence among members of the collective. When social categorization corresponds to the collective as whole, co-operative interdependence is salient and decisions are motivated by a desire to maximize joint or collective outcomes.

(Brewer & Schneider 1990, 177-178.)

---

<sup>88</sup> There are even more similarities between the results of my theoretical arguments and the empirical theories of the adherents of the social identity approach. Firstly, Hogg and Abrams think that “a collection of individuals [...] becomes a group to the extent that it exhibits group behaviour” (Hogg & Abrams 1988, 106), *i.e.*, to the extent that the members experience the psychological change of switching from personal identity to social identity. Similarly, I argued in Part I that a collection of individuals forms a group insofar as the individuals lay aside their individual-mode considerations and adopt collective we-mode forming thus a plural subject.

<sup>89</sup> It might seem dubious that studies of social dilemmas provide both *conceptual* and *empirical* evidence, especially since we are searching for empirical approaches that would lend independent support to the conceptual arguments. The fact that the social identity theorists study social dilemma situations *empirically* rather than in terms of conceptual analysis should help to calm such worries. However, the fact that in the present context the conceptual and empirical aspects *are* hopelessly intertwined forms a major part of my philosophical argument in Part III, and thus I fully share the worry.

<sup>90</sup> Recall my references to Sen (via Anderson and Schmid) in II.2.4.

This phenomenon is well documented by the so-called minimal group studies characteristic of the empirical experiments of social identity theorists. For the present purposes the following rough summary of the results of minimal group studies suffices. If individuals agents are provided with conceptual tools – perhaps rather trivial<sup>91</sup> – with which to classify agents to “us” and “them”, this is sufficient for promoting cooperative attitudes towards the members of the we-group (and competitive attitudes towards the non-members) even in tasks completely unrelated to the underlying classifications:

Once group identification has been established, intragroup orientations are characterized by the best of human motivations: perceived mutuality, co-operation and willingness to sacrifice individual advantage for the sake of group goals. However, when in-group identity is achieved through differentiation from other groups at the same level of organization, intergroup orientations are characterized by just the opposite: perceived conflict of interest, social competition and willingness to sacrifice joint welfare for the sake of in-group advantage. (Brewer & Schneider 1990, 178.)<sup>92</sup>

Thus, Brewer and Schneider’s conclusions of the teachings of empirical studies of social dilemma situations seem to confirm Sober and Wilson’s conjectures concerning the implications of their multilevel selection theory: “Group selection favors within-group niceness and between-group nastiness” (Sober & Wilson 1998, 9). Similarly, the view of the social identity theory corresponds to the understanding of the relationship between an individual and her we-group raising from Sober and Wilson’s group selection theory: “The ‘I’ is defined by relating it to a ‘we.’ Human beings don’t simply *belong* to groups; they *identify* with them.” (Sober & Wilson 1998, 233). As Wilson indeed concludes, “[m]ultilevel selection theory is the perfect compliment to social identity theory” (Wilson 2002, 139; see also p. 144). Consequently, just as was the case with Sober and Wilson, also Brewer and Schneider’s findings are best interpreted as resulting from agents’ natural tendency to adopt we-mode attitudes.

---

<sup>91</sup> For example, the experimenter may claim, perhaps counterfactually, that the individuals can be divided into two groups according to their taste in music or something similar.

<sup>92</sup> Thus, a theory building on we-attitudes might provide a theoretical justification for Peter Singer’s view, according to which expanding the scope of ethically relevant we-groups (recall the title of Singer 1981) is the core problem of ethics. Indeed, the present marriage of the theory of collective intentionality and Sober and Wilson’s group selection theory is both better supported by arguments and ethically more ambitious than Singer’s own naturalism, which builds on a straightforward kin-selection theory (Singer 1981, 194). The present account can accommodate easily what Singer (1981, 111) sees as the essence of ethics, namely that the scope of ethics is not limited to family members or other individuals with whom the agent has had long-term immediate contacts. In contrast, Singer’s (1981, 194) kin selectionism presupposes that the human mind is able to break free from the natural order of the world and adopt “the point of view of the universe” (Singer 1994, 228-229) from which ethical judgements are made. I wonder whether this is compatible with Singer’s pronounced naturalism and kin-selectionism.

Moreover, the social identity theorists recognise the point emphasised in this study (in particular, Parts I and III), namely that when agents switch to we-mode their actions and intentions have (at least partly) purely social reasons. We-mode agents do something because that is what they think is appropriate for a member of their group to do. No further reason for action is required. This role of social identification is captured nicely by the title of a classical article in the social identity tradition, Nelson N. Foote's "Identification as the Basis for a Theory of Motivation" (Foote 1951). As Foote puts it, only by acknowledging such thoroughly social explanations of human action can social psychology truly deserve the name "social" (Foote 1951, 21). Similarly to my arguments in Parts I and III, also Foote's view is far from the Standard View (III.1), exemplified by mainstream rational choice theory, according to which actions must always be explained and rationalised in terms of an individual-mode belief/desire pair.

Notwithstanding the similarities between the views raising from the conceptual arguments of the present dissertation and the account motivated by the empirical studies within the social identity approach in social psychology, there are, however, also certain interesting differences. First, although I think the social identity approach can offer us a very realistic picture of the workings of human psychology in social settings, the psychologists do not always have suitable conceptual tools for expressing their views clearly and coherently. Brewer and Schneider, for example, repeatedly fail to be explicit about the precise nature of the distinction they are making between individual and social identity. In particular, they do not distinguish clearly whether sociality in their view amounts to the *mode* of intentional attitudes, to the *content* of such attitudes or perhaps to the *directedness* of such attitudes. These distinctions are absolutely crucial for the philosophical debates on rational co-operation and in this respect I think the present philosophical approach can assist the empirical tradition towards more advanced conceptual precision.

Similarly, Hogg and Abrams, who do not possess the concepts of we-mode and collective intentionality, are forced to use vague metaphors, such as "the group in the individual" (Hogg & Abrams 1988, 3 & 217) and obscure illustrations such as "people in groups, unlike atoms in molecules, can contain psychologically the whole within themselves" (Hogg & Abrams 1988, 101). Thus, I would like to suggest that the philosophical theory provided by the present study could be used to make the conceptual apparatus of the social identity theory more precise. The social identity theorists' main interests lie in empirical studies of group behaviour, not in conceptual clarifications.

The philosophical and the empirical approaches complement each other in an important manner.

Although it seems clear to me that the social identity approach in social psychology lends very strong empirical support to the philosophical theory of collective intentionality, it remains true – just as was the case with experimental economics too – that the social psychologists need to *interpret* the experiments they refer to quite radically. No doubt also some other kind of interpretation would be possible. In particular, mainstream economists would say that the categorisations of minimal group studies activate *individual-mode* pro-attitudes directed at one's group. The synthesis of the theory of collective intentionality and the social identity theory amounts to a possible conceptual framework within which social action can be analysed and understood. The economists offer another possible framework. Thus I can conclude neither that my theoretical results have been empirically proven beyond all doubt nor that the social identity approach in social psychology is unambiguously supported by empirical evidence.

This conclusion, even if not fully satisfactory,<sup>93</sup> is nonetheless largely *sufficient* for my purposes, for I have already presented strong *philosophical* arguments in favour of the collective intentionality theory, and in the social identity approach in social psychology we have at least one interesting empirical scientific research programme that could bring the philosophical theory of collective intentionality in touch with empirical social sciences in a fruitful way. Collaboration and co-operation of philosophers and empirical social scientists could lead to a mutually beneficial outcome: The philosophers could get much needed empirical support for their theories, and the empirical scientists could apply the conceptual clarifications of philosophers to get a better grasp of the cluster of problems they are approaching from the empirical point of view. What I have said concerning the social identity approach suffices to point one possible place where such a co-operation and collaboration could get off the ground.

---

<sup>93</sup> Actually, the examination of the nature of intentional action and its explanation in Part III demonstrates that we *cannot* require a stronger result here. There is no *fact* independent of our practices of accepting certain frameworks that would make one or the other framework unambiguously true (III.5.3).

## CHAPTER II.4: CONCLUSION

I opened this Part with Daniel Dennett's story that was meant to illustrate how evolutionary processes always promote (i) evolutionary egoistic behaviour and (ii) psychological motivations that operate strictly within Hollis' philosophical egoism. However, closer analyses have led me to reject both (i) and (ii). The question is, then, does my arguments show that we should reject Dennett's picture of evolutionary dynamics as fundamentally false? I do not think so. The picture given by Dennett's story is not so much *false* as it is *incomplete*. The story captures well the within-group selection force for egoistic behaviour, but remains silent about the between-group selection force favouring socially beneficial behaviour – and similarly for corresponding psychological mechanisms.

Thus I would like to replace Dennett's story as the recommendable illustration of the evolutionary dynamics with another story that better captures the complexity of human psychology. The story I have in mind is the famous opening scene of Ian McEwan's thought-provoking novel *Enduring Love*, which captures aptly the two modes of human psychology, as well as the tension between them. Somewhat tellingly, this story is not a comical story, but a tragedy.

In the novel the characters face an idyllic scene that turns into a nightmare. An enormous balloon, with a small child in its basket, is anchored to the ground. But it is a windy day: a violent blow lifts the balloon high into the air, tearing the anchor off the ground. Five men, previously unknown to each other, rush to help and manage to catch the ropes hanging from the basket. Within seconds the men find themselves hanging on the ropes several feet above the ground. Everything happens so quickly that the men have no time to form an explicit plan, or indeed to communicate a collective intention, about what they should do.

Here is McEwan's narrator:

I didn't know, nor have I ever discovered, who let go first. I'm not prepared to accept that it was me. But everyone claims not to have been the first. What is certain is that if we had not broken ranks, our collective weight would have brought the balloon to earth a quarter of the way down the slope a few seconds later as the gust subsided. But as I've said, there was no team, there was no plan, no agreement to be broken. No failure. So can we accept that it was right, every man for himself? Were we all happy afterwards that this was a reasonable course? We never had that comfort, for there was a deeper covenant, ancient and automatic, written in our nature. Co-operation – the basis of our earliest hunting

successes, the force behind our evolving capacity for language, the glue of our social cohesion. Our misery in the aftermath was proof that we knew we had failed ourselves. But letting go was in our nature too. Selfishness is also written on our hearts. This is our mammalian conflict – what to give to the others, and what to keep for yourself. Treading that line, keeping the others in check, and being kept in check by them, is what we call morality. Hanging a few feet above the Chilterns escarpment, our crew enacted morality's ancient, irresolvable dilemma: us, or me.<sup>94</sup>

Collective we-mode psychology can indeed collapse easily if the individual price to pay appears to be too high, or if there has not been enough time for communication to establish trust between the individuals. In McEwan's story there was no stable plural subject. Although, tragically, all the individuals could see that there could – and should – have been.

In addition to the themes discussed in this Part, McEwan suggests that the tension between collectively rational we-mode actions and individually rational individual-mode actions is a major source of our moral or ethical considerations. Although this is a theme I cannot discuss properly within the limits of this dissertation, let me say a couple of words about it – this may be taken as an indication for the need of further research. McEwan's story highlights the fact that there appears to be a collective requirement that in social life people should follow collective rationality and not their individual self-interest.<sup>95</sup> To put it in terms of traditional moral philosophy, the Kantian idea of contrasting mere individually instrumental action with truly moral action seems to capture our moral intuitions rather well. The extract from McEwan's novel suggests, however, that instead of attributing morality to those intentions that stem from abstract considerations of timeless and universal moral principles, we should, perhaps, compare individually instrumental considerations with considerations executed from the collective we-mode perspective.

Thus the universality required from ethics would be open to a naturalistic explanation without a reference to timeless and universal moral principles. So although according to this view the end products of ethical reasoning might be guidelines comparable to Kantian imperatives, the source of morality would not be unworldly contemplation of Platonist moral ideas. Rather, the starting point of morality would be the adop-

---

<sup>94</sup> Ian McEwan, *Enduring Love*, London, Vintage, 1997, 14-15.

<sup>95</sup> Correspondingly, social identity theorists have found in their experimental studies that people tend to conceptualise the tension between individual-mode rationality and collective considerations in the obviously normative terms of good and bad (see Kerr & Park 2001, 118). Similarly, for Singer (1994, 229) the requirement and starting point of ethics is "the possibility of detaching myself from my own perspective" – but, as I have explained above (Footnote 92), in Singer's view realising this possibility requires that we overcome our natural tendencies and adopt the view from nowhere.

tion of a collective we-mode perspective to instrumental rationality – which obviously corresponds in an interesting way to the Kantian maxim of never treating other agents as mere means relative to one’s (individual) ends (cf. Singer 1994, 231), *i.e.*, rejecting the rational choice picture of social action as *strategic* individual-mode action.

Since this dissertation is not a treatise in moral philosophy, I will not develop this line of thought further here. However, Part III defends a naturalised (in the sense of de-Platonisation), essentially Kantian theory of action, rationality and agency, which is highly analogous to the view of morality sketched above. Correspondingly, even such a brief excursion to moral philosophy brings once again forward the fundamental ontological view argued for in this essay. In short, whether we are interested in collective action and rationality, social institutions or morality, my argumentation points towards a view that is ontologically naturalistic in the sense of not appealing to unnatural entities, be they collective minds or timeless principles in Plato’s heaven. On the other hand, the arguments nonetheless support methodological anti-individualism.

Once the ontological picture is clear (*i.e.*, when *(OI)* has been accepted) and we are operating safely at the level of individuals, it is a mistake to think that the accepted ontological view forces us to treat individuals as isolated social atoms with exclusively individual-mode psychology – recall McClennen (II.2.4), who, mistakenly in my mind, thought that appealing to collective notions at the level of individuals would commit him to ontologically suspect notions. Similarly, Hollis’ mistake is to think that there are no ontologically acceptable alternatives to philosophical egoism.

However, there is a third way between atomistic individualism on the one hand and unnatural ontological holism on the other. Ontologically speaking, the social world consists exclusively of individuals, but they are essentially linked to each other. The link, I have argued, is we-mode psychology. Gilbert (1989), Pettit (1993), Searle (1995) and Tuomela (1995) all accept views that are, although different from one another, nonetheless essentially similar to mine in this respect. The view I defend is not methodological individualism, but it is not traditional ontological holism either. This is why Pettit calls such views *individualistic* (as opposed to collectivism in the sense of rejecting *(OI)*) *holism* (as opposed to atomism in the sense of disparate individual-mode agents), and Tuomela *interrelationism*. In this dissertation my goal is to explicate an acceptable version of such a view. I take this Part to have shown that the view is evolutionarily and psychologically plausible: It is well supported both theoretically (social dilemmas) and empirically (especially by studies within the context of the social identity approach in social psychology).



Note that the picture sketched in this Part, although drawing heavily from evolutionary considerations, is clearly different from approaches known as socio-biology or evolutionary psychology. Most importantly, my approach does not assume any kind of direct relation between genes and specific actions or mental modules and contents. This is something that mainstream evolutionary psychology is correctly accused of (Sterelny 2003). In fact, I consider my work to be highly compatible with such uncompromising criticisms of evolutionary psychology as Dupré (2001) (this connection is made explicit in III.5.3). Moreover, my account does not need to assume that humans are simple replicators that copy the most effective strategy they observe (cf. Skyrms 1996). My approach allows, *pace, e.g.*, Hollis and Verbeek, collectively rational action to be based on complicated deliberations of instrumental rationality (executed in the we-mode).

In short, the main theses of this Part are the following. First, contrary to what might appear to be the case, the claim that we-mode collective intentionality is a primitive feature of human psychology is compatible with our understanding of evolutionary dynamics. Second, we-mode collective intentionality is not only compatible with evolutionary dynamics; it is also a more plausible candidate than mere individualistic psychology (including the psychological altruism favoured by Sober and Wilson) to be the core psychological mechanism behind truly social action. Thus, in addition to accepting ontological individualism, the theory of collective intentionality is in this Part shown to be naturalistic also in the more specific sense of being compatible with and even supported by our understanding of evolutionary dynamics. The *Homo Sociologicus* does not contradict evolutionary theory. Third, the arguments of this paper continue to speak strongly against mainstream methodological individualism by showing that not only are purely social reasons possible reasons of human action, but evolutionary considerations should make us to expect it to be often the case in social interactions. Contrary to methodological individualism, rational human agency is, sometimes, irreducibly *collective* in its form and mode.

PART III:  
ACTION AND AGENCY

## INTRODUCTION TO PART III

“Anyone who cannot form a community with others, or who does not need to because he is self-sufficient [...] is either a beast or a god.”

- Aristotle, *Politics*, 1253a, 27.

I have already argued for two of the main theses of this dissertation. First, I have argued that, ontologically speaking, social institutions and the whole of social reality consist of collectively accepted and required patterns of behaviour (Part I). Despite their objectivity, social institutions are nonetheless mind-dependent; their construction requires there to be intentional agents who collectively assign and maintain normative functions. Second, I have argued that the theory of collective intentionality as a fundamental building block of social reality is supported by our understanding of the dynamics of the evolutionary theory and the evidence gathered by the empirical social sciences, and these arguments for a strong, general argument in favour of Durkheimianism in the sense of 1.1.5 (Part II).

The remaining question in the present ontological project concerns the justification of the broad strategy of building on the notion of intentional agency (and, ultimately, the status of *Wittgensteinianism* in the sense of 1.1.5). We must ask whether the general *framework of intentional agency*, which has been presupposed in the earlier Parts, is really warranted. More precisely, we must ask what is the nature and status of intentional explanations. This question is of course relevant to all human sciences that operate with the notion of intentional agency. Thus this part should be of interest even for those who do not accept the views defended in Parts I and II – although I shall argue that answering the question compels us to accept the views defended in the earlier parts of my dissertation. Moreover, also this question will guide us to accept naturalised methodological holism. Or so I argue.

When assessing the soundness of the framework of agency the animating questions are, once again, ontological in nature: What is intentional agency? How could there be intentionality, or *aboutness*, in the world that ultimately consists of blind physical processes? However, if agency and folk psychology, which are essentially conceived of in terms of contentful and intentional mental states, are mere illusions, or at least readily reducible to physicalist neuroscience that has no use for the notion of intentionality, then the whole programme established in the earlier parts of this dissertation seems to vanish into thin air. Hence the status of the views defended in this dissertation

cannot be established until the (ontological and explanatory) status of intentionality is addressed in detail.

The answer to the challenge of eliminativism in this Part takes the methodological holism defended in the earlier parts even further. I argue that the quotation above from Aristotle is literally correct: intentional agency and personhood as we know them are possible only within social practices. Intentional human agents are (partly) constituted by what I call social bedrock practices. Outside such practices there can be only non-intentional and non-rational beasts whose behaviour is determined by blind dispositions or gods who do not belong to the natural order of the world.

The starting point of my argumentation is the platitude that human activities can be studied from several different angles. What distinguishes the perspective of the moral, human or social sciences from that of the natural sciences is that the former study human *action* and not mere human *behaviour*. The distinction between action and mere behaviour is aptly captured by Max Weber's well-known depiction: "In 'action' is included all human behaviour when and in so far as the acting individual attaches a subjective meaning to it" (Weber 1922/1947, 88).<sup>96</sup> Thus, the defining feature of action appears to be its *meaningfulness*. In some sense action must, as it were, be understood from within – we do not grasp an action accurately if we do not understand the meaning it embodies.

The crucial question, then, is what does it mean to say that action must be understood from within. From within what? It seems that we have two competing intuitions on how to answer that question. It appears sensible to say that actions must be understood from within the mind of the acting individual. Acting is doing something for a reason: An agent intends to achieve something with her actions. Weber's definition seeks to capture this by emphasising the role of the *subjective* meaning an individual attaches to behaviour. In this *individualistic* view the intentionality of action is *derived* intentionality; its intentionality and meaning is derived from the intrinsic *original* intentionality of the mental states of the acting individual.

---

<sup>96</sup> Perhaps Weber would not approve of the way I group together the moral, human and social sciences. Weber famously argued that the social sciences study *social* action, and in his view "action is social in so far as, by virtue of the subjective meaning attached to it by the acting individual (or individuals), it takes account of the behaviour of others and is thereby oriented in its course." (Weber 1922/1947, 88). For the present purposes my grouping is, however, justified, since (i) even if one accepts Weber's distinction between individual and social action, the distinction between action and mere behaviour is still philosophically absolutely crucial and (ii) below (Chapter III.3 onwards) I defend a view similar to that of Peter Winch, according to which ultimately "all meaningful behaviour must be social, since it can be meaningful only if governed by rules, and rules presuppose a social setting" (Winch 1958, 116).

However, also a holistic intuition, according to which action must be understood from within the social rules that give human action its meaning, seems appealing. A physical behaviour of raising one's arm is the action of greeting a friend only if there exists a social convention that assigns the physical movement in question the status of a greeting (note the obvious connection between this view and the discussion on the assignment of functions and the construction of social statuses in Part I of this study). The very same physical behaviour could have different meanings – be a different action – in different social circumstances. Thus, similarly to the individualistic view, in a sense also the holistic view treats the intentionality of action as derived intentionality, but locates *original* intentionality in our social practices.

The holistic view of original intentionality seems to re-invoke the ever-present problem of this study (recall especially 1.2), namely the dilemma of choosing between the circularity of argumentation and the reification of society and social practices. According to holism, the intentionality of action presupposes social practices. However, general ontological individualism (*OI*) seems to dictate that to avoid reification, the practices themselves must be based on intentional actions of individuals. To avoid such circularity, the holistic view seems to require practices that exist *independently* of individual actions. But this, in turn, appears to imply an explicit rejection of the ontological principle (*OI*).

This dilemma motivates John R. Searle (1995) to think that the theories which see social facts as collectively assigned statuses, *e.g.*, the theory defended in Part I of this study, must in the end be *individualistic* in a stronger sense than the mere ontological sense of (*OI*). Searle thinks that holism is correct about all the other meaningful or functional things in the world except the minds of individuals, which are intrinsically intentional. Pieces of paper can be money, speech-acts can be communication, ink stains on a paper can have meaning and so on only if they are collectively assigned such statuses in our social practices of requiring and accepting certain behaviours. However, Searle reasons, to avoid ontological reification such practices must be based on individual acceptances (albeit in the we-mode) and, finally, to avoid circularity the mental states of individual acceptances must, according to Searle, be intrinsically meaningful and intentional. Hence, Searle concludes, we must accept individualism in the strong sense of thinking that individual agents are intrinsically capable of intentional thoughts and actions.

Daniel C. Dennett has ridiculed this line of thought on the grounds that it seems to postulate (to use Dennett's terms) the existence of some kind of magical *wonder tis-*

*sue* which can be intrinsically meaningful and intentional (about something) – and that brains are the only objects made of this wonder tissue. Dennett thinks that since argumentation along the lines of Part I of this study shows that there cannot be intrinsic functions and intrinsic intentionality in the physical world, one cannot go on to assume the existence of intrinsically intentional minds. This would surely be a case of treating individuals as gods outside the natural order of the world.

I think that Dennett is correct here: In Chapter III.3 onwards I argue that Searlean individualism cannot be accepted. We are not gods. However, I refuse to admit that there would be anything naïve or ridiculous in Searle's reasoning. The apparent alternatives are, after all, at least as unacceptable. Since we do not want to reify social practices it is no wonder that Dennett is often, despite his repeated protests, seen as an eliminativist who has no place for thoughts, meanings or intentional actions. If there is no original intentionality, there seems to be no intentionality at all. But this amounts to treating individuals as non-intentional and non-rational beasts.

Hence the fundamental task undertaken in this Part of my study is to salvage the intentionalist programme by providing an account of intentionality, meaning and action that avoids the four unacceptable solutions I have just sketched: (i) *individualism* (individuals as gods), (ii) *circularity* of practices and actions, (iii) *reification* (rejection of (OI)) and *eliminativism* (rejection of the intentionalist programme; individuals as beasts).

The solution, I argue, requires that we find a way to be holists without falling guilty of circularity or reification: we must naturalise holism. Such holism must maintain that our most fundamental social practices that make intentionality possible are based on *pre-intentional behaviour* (as opposed to meaningful action) at the sub-personal level (Chapter III.4). These fundamental practices then *bootstrap* intentionality into existence – and turn behaviour into action. Here intentionality is seen as a kind of self-validating *social performance*, and the intentionality of individual psychology, action and public language are all constructed at the same time. Original intentionality resides in social practices. This view is analysed and defended in detail in Chapter III.3 onwards. Ultimately, this examination of the nature and sources of intentionality will deliver a novel, constructivist theory of collective we-mode intentionality (III.5.2).

However, before we can look at the fundamental preconditions of intentional actions we must understand the nature of action. In particular, in addition to the problem of individualism versus holism (is original intentionality to be found in individual brains or social practices), the nature of meaningful action presents another central problem

relevant to the philosophical foundations of the views defended in this study. This is the long-lasting debate between causalist<sup>97</sup> *naturalism* and non-causalist *humanism*. Only when we have cleared this cluster of problems we can return to the problem of formulating an acceptable version of holism. So let us look at the nature of intentional action and its explanation in terms of the naturalism versus humanism debate.

---

<sup>97</sup> In this dissertation I cannot of course develop and defend a philosophical theory of causation. Thus I have done my best to write in such a way that my arguments are not tied to any particular view of causation. However, in what follows I am committed to a realist, mind-independent notion of causation, which sees causal relations as *nomological* relations (in terms of Part I, such law-governed relations obtain *independently*). I contrast such mind-independence with normative notions that are *rule-based* and depend on what we *take* the rules and norms to be (such norms may nonetheless obtain *externally* and *objectively*). Thus, if one holds a view according to which causation is in fact projected onto the world by our psychological or linguistic categories that are based on our notion of agency (or something similar), such that causal notions are ultimately psychological or normative notions, then my talk of causation should be translated into talk about whatever one takes to be governing the mind-independent physical world - cf. Stoutland (2005, 130), whose views the present study tends to agree with, not only in this respect, but also in general (see, in particular, Stoutland 1986 & 1988), although in accepting that also the “environment” of observable behaviour is “intentional, and not merely physical phenomena” (Stoutland 1988, 44), Stoutland comes, at least apparently, dangerously close to McDowell’s (1994) idealism, which is criticised and rejected in A.2.

## CHAPTER III.1: INTENTIONAL ACTION AND ITS EXPLANATION<sup>98</sup>

### III.1.1 THE STANDARD VIEW

When we aim to describe the sense in which behaviours have meanings and, thereby, are actions, we are led to statements such as that actions are behaviours performed for a reason or that actions are behaviours that *we do* as opposed to what *happens to us*. In A. I. Melden's words, there "is a difference between my arm raising and my raising my arm, my muscles moving and my moving my muscles – in short, between a bodily movement or happening and an action" (Melden 1960, 70). Hence, actions may be said to be behaviours that are under our control. The obvious question then, of course, concerns the nature of the control.

The standard way of approaching this question, which is one of the most fundamental questions in the philosophy of mind, action theory and the philosophy of the human sciences, is to rephrase the question as an inquiry into the fundamental nature of the following principle (*L*) that plays a central role in *intentional explanations* (the following formulations are adopted from Rosenberg 1995):

(*L*) If any agent, *X*, desires *D*, and believes that doing *A* is the best means to attain *D* under the circumstances, then *X* does *A*.<sup>99</sup>

(*L*) allows us to construct the standard form of intentional explanation as follows:

1. *X* desires *D*.

2. *X* believes that *A* is the best means to attain *D* under the circumstances.

(*L*) If any agent, *X*, desires *D*, and believes that doing *A* is the best means to attain *D* under the circumstances, then *X* does *A*.

---

*Ergo*: *X* does *A*.

---

<sup>98</sup> This Chapter expands on Saaristo (2006a).

<sup>99</sup> Obviously, (*L*) requires a *ceteris paribus* clause: (*L*) holds only if *X* does not have any other desires overriding *D*, if *X* knows how to do *A*, if *X* is able to do *A* etc. Arguably the *ceteris paribus* clause cannot be removed from (*L*). Important as this issue may be, it is not something I want to address presently, for in III.5.1 I argue that the *ceteris paribus* problem is not a problem for the interpretation of intentional explanations I defend in this study.



Here a behaviour  $A$  is (or counts as) an intentional action of  $X$  iff doing  $A$  is under  $X$ 's control in the manner described by  $(L)$ . Hence the question concerning the nature of the control can now be expressed in the context of intentional explanations as the question concerning the status of  $(L)$  and indeed the status of the practical inference that the explanation consists of.

Broadly speaking, we have two options. First, we could hold that Premises 1 and 2 pick out the mental states of  $X$  which in fact *caused*  $X$  to do  $A$ . Here  $(L)$  is regarded as expressing a universal, causal law of nature describing the causal functioning of characteristically rational agents such as humans. According to this view, our actions are under our control in the very natural sense of being caused by our intentional states (in particular, our intentions, beliefs and desires). This conception of  $(L)$  and intentional explanation seems to agree very well with general scientific naturalism. Rationality is seen as a natural property of beings such as humans, and  $(L)$  is a descriptive claim. Consequently, if this conception is accepted, intentional explanations can be regarded as naturalistic, *causal explanations* of action similar in nature to the explanations of the natural sciences.

Alternatively, we could maintain that reasons and actions are not causal issues and thus intentional explanations are not causal explanations at all. Rather, the gist of an intentional explanation would be to render an action *intelligible*. According to this conception, human actions, as the humanistic slogan has it, are not explained but *understood*. Here  $(L)$  is not seen as a causal law of nature but as a *normative* principle expressing how reasons and actions ought to be connected for the agent to count as rational (or indeed for the behaviour to count as action and the individual to count as an agent).

Adherents of this humanistic *Verstehen* conception will have to face at least two further, albeit closely connected questions concerning the nature of  $(L)$ . First, where does  $(L)$  come from? If one has Kantian or Platonist inclinations, one could argue that  $(L)$  expresses a universal rational principle graspable by pure reason. Those humanists who find Platonism in all of its forms metaphysically dubious tend to favour a social account, according to which  $(L)$  is socially constructed and resides in our practices rather than in the realm of universal ideas.<sup>100</sup>

---

<sup>100</sup> Or as Robert B. Brandom (1994, 2000, 2002) – who does not hesitate to spice up his detailed argumentation with very broad generalisations from the history of philosophy – sees this, the move from the causalist to the normative view amounts to a move from Cartesianism to Kantianism, and the de-Platonisation of  $(L)$  in terms of social constructivism amounts to a move from Kantianism to Hegel's (and later Wittgenstein's) philosophy. I think that this kind of anti-Platonist naturalism is also the core of, for example, Barnes and Bloor's (e.g., 1982) much-misunderstood *relativism*, which is simply the naturalistic

The second question concerns the role of (*L*). Does (*L*) merely *regulate* rational behaviour, telling how actions ought to be connected to mental states? Or should we see (*L*) as a *constitutive* rule and say that (*L*) interdefines beliefs, desires and actions, such that (*L*) partly constitutes rationality and intentional states?<sup>101</sup>

If (*L*) is seen as a *regulative rule*, it should be formulated as follows:

(*L*\*) If any agent, *X*, desires *D*, and believes that doing *A* is the best means to attain *D* under the circumstances, then rationality requires *X* to do *A*.

With this interpretation, if Premises 1 and 2 are true – and (*L*\*) holds – it still does not follow that *X* will indeed do *A*. Rather, *A* is what *X* rationally ought to do under the circumstances. If *X* does not perform *A*, no laws of nature have been broken. Rather, *X* has done something *wrong*. *X* has failed to comply with the norm of rationality expressed by (*L*\*).

If, in contrast, (*L*) is seen primarily as a *constitutive rule*, we can keep the original formulation of the principle. According to this reading, however, (*L*) does not connect distinct entities or events but rather explicates a normative (inferential) framework within which there can be beliefs, desires and actions. This notion leads directly to the *holism* of the intentional realm (III.1.2 & III.1.3). According to this conception, what it is for *X* to do *A* is that *X* has desires and beliefs as described by Premises 1 and 2, for they are what constitute *X*'s behaviour qua intentional action *A*. Similarly, what it is to have a belief described by Premise 1 is to be committed to performing *A* if the belief is accompanied by the desire described by Premise 2. Finally, to have the desire described by Premise 2 is to be committed to moving from Premise 1 to the action-conclusion *A*. Consequently, the humanist scheme states, we cannot say that the desire and belief cause the action, for they do not exist independently of the action (III.1.2).

Below I defend the humanistic *Verstehen* position in the rather extreme sense that accounts for (*L*) in terms of social practices (Chapter III.3 onwards) and sees (*L*) largely as a constitutive rule (III.1.2 and Chapter III.3 onwards). To motivate such an admittedly heretical view in contemporary analytic philosophy of mind and action<sup>102</sup> we need first to understand what the alternatives are and why they cannot be accepted.

---

claim that norms such as (*L*) must be brought from the Platonic heights of Kant's *noumenal* world into our natural world by showing how they are constructed and maintained in (how they are *relative* to) our social practices.

<sup>101</sup> Recall the distinction between regulative and constitutive rules in I.2.

<sup>102</sup> A view very similar to mine is defended by Stoutland (*e.g.*, 2005) but not by many others.

All the four combinations one can construct in terms of the two fundamental distinction I have introduced – individualism versus holism and naturalism versus humanism (or explanation versus understanding) – are possible positions and, arguably, ought to be discussed separately.<sup>103</sup> However, I will be explicitly concerned with only two of the four possibilities. I set aside the combination of causal explanation and (ontological) holism, for I have already rejected such a view in Part I of this study by accepting that the only causal factors in social life must be individuals and their physical environment. I also will not explicitly address the combination of individualism and humanism, because the considerations in Chapter III.3 clearly write off the acceptability of this combination. The view I think we ultimately should accept, the combination of ontologically naturalistic methodological holism and anti-causalist humanism, will be argued for and analysed in detail in Chapter III.3 onwards. Thus, for the time being I want to concentrate on what I call the *Standard View*, namely the customary way of combining individualism and causalism.

Among social scientists and philosophers of social science the Standard View is routinely seen as the view famously defended by Donald Davidson (1963). However, as I argue in III.1.3 and III.2 below, it is not obvious that Davidson in fact held the Standard View in any straightforward sense. Hence, rather than Davidson's own writings, a typical example of the Standard View in the present sense is, for instance, Jon Elster's interpretation of Davidson in the article "The Nature and Scope of Rational-Choice Explanations" (Elster 1985).

The Standard View maintains that the mental states of an agent – in particular, her beliefs and desires – are the causes of the actions of the agent. However, the Standard View sees an intentional explanation as something more than a mere causal *explanation* of an action. The explanation also renders the action *intelligible* by showing how the action was (instrumentally) *rational* in light of the beliefs and desires of the agent. This aspect of *understanding* the action is combined with causalism about intentional explanation in the sense of explaining the action in terms of a causal structure distinctive of rational agents. The *explanatory power* of intentional explanations comes from seeing the reasons for an action as the causes of the action.

The causal, explanatory aspect of intentional explanations requires us to focus on mental states *qua* relata in causal relations, *i.e.*, *qua* concrete *particular things or events* (depending on one's metaphysical categories) out there in the world. Modern ma-

---

<sup>103</sup> For example, Martin Hollis (1994) builds his entire examination of philosophy of the social sciences around the four possibilities.

terialists are inclined to think that this amounts to approaching mental states in terms of the physical brain states that are believed to realise the mental states (these issues are discussed and analysed in detail in III.1.3 and III.2). The rationalising aspect of intentional explanations, on the other hand, compels us to concentrate on mental states essentially *qua* mental states, *i.e.*, in terms of the *propositional contents* of or the *meanings embodied* by the mental states in question.

This being the case, it is worth noting that we appear to be dealing with two fundamentally different categories here. The causal conception seems to guide our attention towards the physical states picked out by the descriptions “*X*’s belief that *P*” and “*X*’s desire *Q*”, regardless of what propositions *P* and *Q* stand for. The relation that matters for causal explanation is a (causal) connection between concrete, physical particulars. The rationalising conception, in contrast, is pulling into the opposite direction. What matters for rationalisation is the propositional contents expressed by *P* and *Q*, regardless of how those contents are instantiated. The relations in which *P* and *Q* stand to one another that matter for rationalisation are normative, logical and conceptual relations. The fundamental problem addressed again and again in the present Part of my study is the compatibility of these two explanatory functions. I argue that in the end they are indeed so different that one ought not to imagine that a single form of explanation could capture the two.

This, however, is precisely the bold aim of the Standard View. What the advocates of the Standard View wish to say is that somehow the content of a mental state *is* its physical realisation (recall the wonder tissue theory), and hence it is the *content* that has the causal strength or efficacy to cause actions. As my chosen representative of the Standard View puts this point, intentional explanations of human actions must satisfy three conditions (Elster 1985, 311-312):

- (i) Beliefs and desires *rationalise* actions in the sense of (*L*).
- (ii) Beliefs and desires *cause* actions.
- (iii) Beliefs and desires *cause* actions *qua reasons*.

This is a very ambitious view, and a great deal must be said before we can see why it must be rejected.

### III.1.2 THE LOGICAL CONNECTION ARGUMENT

The Logical Connection Argument<sup>104</sup> against the Standard View is motivated precisely by the distinction between a non-normative, causal relation between physical entities and a normative, conceptual (or logical) relation between propositional contents. The Logical Connection Argument holds that the principle (*L*), which the Standard View regards as a law of nature describing the causes of actions, is rather a defining mark of rationality and rational agency. Moreover, (*L*) and other such rules governing theoretical and practical reasoning, are *constitutive* of rationality. Thus, if a person violates (*L*), she is not breaking any laws of nature. However, the Logical Connection Argument is not simply saying that in such a situation the person fails to act rationally. Rather, the argument goes, if the person violates all or most of such rules, she fails to display what we call beliefs, desires and performances of intentional actions. According to the Logical Connection Argument, what it *is* to have contentful mental states and to perform actions is simply that one's activities are coherently describable in terms of (*L*) and other similar principles.<sup>105</sup>

---

<sup>104</sup> It is difficult to pick one authoritative formulation of the Logical Connection Argument. The label refers to a cluster of views that were inspired by Wittgenstein's later philosophy and which had their heyday in the 1940s, 1950s and 1960s – Davidson, for example, says simply that his 1963 paper, which is generally seen as the decisive attack on the Logical Connection Argument (see III.1.3 below), joined Hempel in “swimming against a very strong neo-Wittgensteinian current of small red books” (Davidson 1976, 261). In the classical form the view is stated in Winch (1958), but I concentrate mainly on the elegant formulation of the argument given in another small red book, namely von Wright (1971). However, my motivations are systematic rather than historical: I am concerned with what I take to be the strongest form of the argument, *i.e.*, my own reconstruction rather than von Wright's actual version, although I do agree with Frederick Stoutland (personal communication) that my version may at least in spirit be relatively close to what von Wright had in mind all the time (cf. Footnote 182 below). However, my aim is *not* to trace the historical development of von Wright's thought (for this, see Schilpp & Hahn 1989, Kusch 2003 and indeed von Wright 1989a & 2001). For example, in 1971 von Wright was still reluctant to talk explicitly about rationality, not to mention the explicit principle (*L*) (for discussion, see Black 1989). I also talk in terms of reasons for action – a convention von Wright adhered to only in his later works (cf. Kusch 2003, 338). Stoutland, whose position I largely agree with, says that he defends a view that he attributes also to von Wright “in terms that he himself might not have used but that bring it into more explicit contact with current work in the philosophy of action” (Stoutland 2005, 128). The same applies to the present study. In addition to Von Wright, the spirit of the Logical Connection Argument, as we shall see, continues to live, for example, in the works of Brandom (1994 & 2000), Esfeld (2001) and McDowell (1998a & 1998b) and the present study.

<sup>105</sup> Thus, it is somewhat unfortunate that this view is known as the *Logical* Connection Argument, because “it is a mistake [...] to understand the intentionalist view [the Logical Connection Argument] to mean that there is a relation of logical entailment between the premisses and the conclusion of a practical argument” (von Wright 1976b, 422, cf. Kusch 2003, 339, which argues that von Wright actually changed his mind here and, finally, Malcolm 1989 for a discussion concerning this problem). The point is, rather, that the premisses and the conclusion do not name events or states conceptually independent of one another. This is what von Wright (1971, 117; 1972) means with his perplexing claim that although *ex post actu* the connection between the premisses and the conclusion can be seen as that of conceptual necessity, we cannot before the action happened predict the action on the basis of the mental states of the agent. Of course this is but a reformulation of the view that intentional explanations are rationalising expositions operating with constitutive rules rather than causal explanations operating with causal laws. Consequently, I will talk somewhat loosely of a logical, conceptual or normative connection.

Thus, according to the Logical Connection Argument, (*L*) serves to *interdefine* beliefs, desires and actions. In short, the Logical Connection Argument says, as G. H. von Wright puts it, “that a distinguishing feature of the causal relation is that cause and effect are *logically independent* of one another” (von Wright 1971, 93) and since beliefs, desires and actions are logically *dependent* on one another, they cannot stand in causal relations.

To drive home this point von Wright invites us to consider the problem of *verifying*<sup>106</sup> whether an agent has a certain intention:

Let it be asked how, in a given case, one ascertains (verifies) whether an agent has a certain intention, “wills” a certain thing – and also how one finds out whether his behavior is of a kind which his intention or will is supposed to cause. Should it turn out that one cannot answer the one question without also answering the other, then the intention or will cannot be a (humean) cause of his behavior. The facts which one tries to establish would not be logically independent of one another. I shall try to show that an investigation of the problem of verification must lead to this result.

(Von Wright 1971, 94-95.)

The starting point of von Wright’s argument is the observation that Premises 1 and 2 as well as the conclusion in our paradigmatic model of an intentional explanation above “are contingent, *i.e.*, empirically and not logically true or false, propositions. It must therefore be possible to verify and to falsify – or at least confirm or disconfirm – them on the basis of empirical observations and tests” (von Wright 1971, 107).

---

<sup>106</sup> For those uncomfortable with von Wright’s empiricist emphasis on verification I should add that in my view Tuomela is correct in claiming that “von Wright could as well [and indeed should] have made his point without employing the (methodological) notion of verification. For his general idea is the conceptual one that, both in discussing the premises and the conclusion of the practical syllogism [an intentional explanation], we have to assume the ‘teleological framework’ or the ‘standpoint of agency’. [...] Thus it seems we can say that action (as opposed to ‘mere’ movement) and the (overall) intention connected with and expressed by it are intrinsically connected both in an ontic and in a semantic sense (and not only methodologically). To describe something as action *means* [...] that *there is* a ‘conduct plan’ (e.g. a practical syllogism) which matches the action” (Tuomela 1976, 195). Indeed, Tuomela (1977, 185) points out that the mere methodological point about verification is clearly insufficient for establishing a conceptual or ontological dependence; it is quite possible that two logically independent scientific propositions are nonetheless never independently testable (cf. Kusch 2003, 332). My treatment of the Logical Connection Argument agrees with Tuomela’s “wide” reading of von Wright’s central ideas, where what matters is the ontology of actions and not methods of verifying propositions (cf. von Wright 1989b, 813-814). Kusch (2003) appears to read Tuomela’s point about the insufficiency of verification as Tuomela’s *criticism* of von Wright’s Logical Connection Argument, whereas I think that at least in Tuomela (1976), on which I concentrate, the aim is rather to explain what von Wright’s real position behind the unfortunate verificationist terminology is – and with which Tuomela (1976, 195-196) explicitly largely agrees (also von Wright (1976a, 402) writes that in this respect “Tuomela’s comments are not at odds with my own opinions”). Tuomela (personal communication) agrees with my interpretation of the 1976 paper; his 1976 criticism of the Logical Connection Argument is not based on von Wright’s alleged verificationism but on von Wright’s reluctance to include causal factors in intentional explanations: although Tuomela (1976) accepts that the conceptual connection constitutes behaviour qua action, for him the *explanatory* work is still done by causal connections.

Actually, however, the issue is trickier than it first appears. Assume, for example, that the action we are explaining is  $X$ 's greeting of her friend by raising her arm. In normal cases it is easy enough to verify what physical movements  $X$ 's arm went through. But in order to identify the movement as  $X$ 's *action* it is not enough to record the movements. The task is to verify  $X$ 's *action of greeting her friend*. To do that, we must show that the physical movement was not something  $X$  brought about only accidentally or by mistake – in other words, we need to show that what took place is something  $X$  did and not something that only happened to her. In short, to verify that  $X$  is greeting her friend we must show that  $X$ 's behaviour, the movement her body is going through, is *intentional under the description "greeting her friend"* (von Wright 1971, 108).

So to verify that someone intentionally *did* something, we must show that the person was *aiming at* doing that very performance. Here problems arise for the Standard View:

But to establish that a certain item of behavior aims at a certain achievement [...] is to establish the presence in the agent of a certain intention [desire, pro-attitude] and (maybe) cognitive attitude concerning means to ends. And this means that the burden of verification is shifted from the verification of the conclusion to that of the premises of a practical inference [an intentional explanation in the present terminology].  
(Von Wright 1971, 109.)

Hence, an attempt to verify the conclusion without verifying the premises of the intentional explanation of the action fails. This leads von Wright to conclude that "the intentionality is *in* the behavior" in the sense that "intentionality is not anything 'behind' or 'outside' the behavior. It is not a mental act or characteristic experience accompanying it." (Von Wright 1971, 115.) To say that a behaviour is an action simply is to say that the behaviour embodies intentionality. It is not possible to identify an action and then go on to discover the beliefs and desires that cause it, for what *makes* a behaviour an action in the first place is the conceptual connection to the reasons for it (in this case, to a certain belief-desire pair). According to the Logical Connection Argument, this amounts to saying that ( $L$ ) expresses a logical or conceptual connection, not a causal law.

In short, von Wright's line of thought seems to be the following. A bodily behaviour is an action only if the behaviour embodies a meaning. The behaviour embodies a meaning only if there is an agent who intends to do something by the behaviour. The agent intends to do something by her behaviour only if there is a description of the behaviour that can be rationalised in terms of the reasons the agent had (and ( $L$ )). The pos-

sibility of this description gives the behaviour the content in virtue of which it is an intentional action. Thus, the connection between the reasons and the action is a constitutive conceptual connection, not an empirical connection between self-sufficient events. Therefore, the connection cannot amount to a causal relation.

Another way of approaching the Logical Connection Argument is to look at the status of (*L*) directly. Von Wright's point that intentionality is in the behaviour and not "outside" it appears to amount to saying that (*L*) is a conceptual definition we use to *re-describe* the very event we seek to explain in terms of beliefs and desires. (*L*) allows us to do that in virtue of interdefining the meanings of "belief", "desire" and "action", not in virtue of picking out a causal connection. The Logical Connection Argument treats intentional explanations as largely analogous to explications of conceptual connections in a language rather than to empirical explanations of the natural sciences.<sup>107</sup> An intentional explanation is a clarification comparable to, say, a situation in which a person learning English finds the statement "John is a bachelor" intelligible only when it is explained that John is an unmarried man and all unmarried men are bachelors. Whatever this kind of explanation may explain, it surely does not amount to a causal explanation of why John is a bachelor (the example is from Rosenberg 1995, 43).

Thus, the Logical Connection Argument takes very seriously the Weberian insight that action is essentially *meaningful* behaviour. This is taken to mean that actions indeed have a *conceptual* content in virtue of which they are meaningful. John McDowell captures this nicely: "Kant says 'Thoughts without content are empty, intuitions without concepts are blind'. Similarly, intentions without overt activity are idle, and movements of limbs without concepts are mere happenings, not expressions of agency." (McDowell 1994, 89).<sup>108</sup> The way to understand McDowell's declaration is to follow G. E. M. Anscombe's (1959) central insight that actions are essentially actions "under a description". To use Anscombe's example, when we observe a sequence of physical movements of a man and ask "What was he doing?", there may well be more than one true answer. The man was moving a lever up and down. He was pumping water into the cistern of a house. He was pumping poisonous water into the house where evil men meet. He was poisoning the men in the house.

However, for a behaviour to be an action there must be a description, under which the action is intentional, that describes what one intends to *do* with one's per-

---

<sup>107</sup> Hence the often misunderstood hermeneutic metaphor of the human sciences treating social phenomena as *texts*.

<sup>108</sup> It is important to note the close similarity between thoughts and actions on the one hand and the essential connection of both to conceptual contents on the other. Both broadly Kantian themes will resurface repeatedly in this Part.



formance. Hence one may act intentionally under description *A* so that one performs also an action under description *B*, even though one does not intend to act under description *B*. The central role of descriptions in (Anscombean) action theory highlights the conceptual element inherent in all intentional actions. Conceptual descriptions are (partly) constitutive of intentional actions qua actions.

The Logical Connection Argument sees an intentional explanation precisely as capturing the process of forming a description under which an action may be performed. An intentional explanation redescribes the action in terms of beliefs and desires such that the action is intentional under that description. In other words, an intentional explanation captures the process of giving a conceptual content to behaviour (and thus turning it into an action) by describing the behaviour in intentional terms that place the action into a normative pattern constituted by rational rules such as (*L*). In von Wright's mind a rationalising description of behaviour – an intentional explanation of action – is what gives meaning and content to behaviour and thus constitutes the behaviour qua intentional action. Stoutland (1976) explains this well when he emphasises that von Wright does not seek to decompose intentional action into its more basic elements. For von Wright, the concepts of intentionality and intentional agency are irreducible.

One does not understand the concept of intentional action by first understanding a concept like (mere) behavior and then adding to it other concepts like causality or desire. To understand the concept is not to eliminate it in favor of other concepts but to see its place in a larger conceptual structure, by possessing which we are agents who can act and see others acting, and who can explain our action and the action of others.  
(Stoutland 1976, 279.)

Stoutland's description captures very aptly the essence of the non-causal humanism behind the Logical Connection Argument. Stoutland also points out the reasons why this kind of humanism is often seen as anti-scientific and anti-naturalistic. After all, the view appears to be that action, agency and other such phenomena (i) are real phenomena, (ii) do not belong to the causal order of the natural world and (iii) cannot be reduced to phenomena belonging to the causal order. To put it in Sellars' (1963, cf. Stoutland 2005, 131) terms, the notions of agency and action belong to the logical space of reasons and normative rationality and not to the logical space of nature (the space of non-normative causal explanations within which the natural sciences operate). This certainly appears to challenge the naturalism I have promoted throughout this study.

Since I nonetheless want to defend a revival of the Logical Connection Argument, one of the main tasks undertaken from Chapters III.4 onwards is to give an ac-

count of how there can be an irreducible normative space of reasons in the natural world. In other words, where many humanists advocate straightforward anti-naturalism, I wish to defend irreducible humanism within the framework of general naturalism. Before I can take up that task, however, let me continue to trace the fate of the Logical Connection Argument in contemporary philosophical debate.

The received wisdom in the contemporary philosophy of action is that the Logical Connection Argument is invalid. The conclusion does not follow. But *which* conclusion, exactly, is not thought to follow from the considerations that are clustered under the label “the Logical Connection Argument”? That intentional explanations as such are not causal explanations? That (*L*) is a normative, constitutive rule and not a causal law of nature? Or perhaps that terms such as “belief” and “desire” do not name (ultimately physical) states that could have as their effects events that are the behavioural components of intentional actions? In the course of this Part of my study I argue that precise answers to these questions – or in other words, what, exactly, we take the core of the Logical Connection Argument (and the standard Davidsonian criticism of it) to be – are absolutely crucial for adequate understanding of the nature of action, intentional explanation and the nature of human agency. The issue becomes clear when we turn to analyse the arguments of Davidson, which are generally supposed to rebut the Logical Connection Argument.<sup>109</sup>

### III.1.3 DAVIDSON’S ALLEGED REFUTATION OF THE LOGICAL CONNECTION ARGUMENT

Donald Davidson opens his attack on the Logical Connection Argument in his famous article “Actions, Reasons, and Causes” (1963) by pointing out that even if we grant that by giving an intentional explanation of an action we redescribe the action in terms of beliefs and desires (a point emphasised by von Wright), it does not follow that beliefs and desires cannot be the causes of the action. On the contrary, we often do redescribe events precisely in terms of their causes. For example, when someone is injured we may redescribe the event by saying that he was burned (Davidson 1963, 10). In this case the man’s injury is redescribed in terms of the man being burned, but surely the possibility

---

<sup>109</sup> As Alfred Mele puts this, “Davidson’s greatest contribution to the philosophy of action is his resurrection of causal theories of action and action explanation. As long as Davidson’s challenge to noncausalists remains unmet [...] causalism will be the biggest game in town, if not the only one.” (Mele 2003, 82.) In the next section I argue, contrary to the view accepted almost universally in contemporary action theory, that – *pace* Davidson and Mele’s claims to the contrary – Davidson does not even present a serious challenge for noncausalists to meet.

of such a redescription does not prevent us from seeing burning as the cause of the man's injury. The possibility of giving different descriptions of the very same event lies at the heart of Davidson's argument against the Logical Connection Argument.

However, an adherent of the Logical Connection Argument might think that this example is essentially different from a redescription of an action in terms of the reasons (beliefs and desires) for the action. After all, as von Wright emphasises, the main concern is the logical or conceptual connection between reasons and actions. We cannot make sense of the one without appealing to the other. This does not hold for burnings and injuries. The connection between them is empirical, not conceptual or logical.

Davidson has an answer readily available. Suppose "*A* caused *B*" is true. If so, we can redescribe "*A*" as "the cause of *B*". By substituting this redescription into the original claim, we have "the cause of *B* caused *B*", which, as Davidson (1963, 14) puts it, is an analytic and not synthetic statement. In other words, we know the sentence to be true in virtue of a conceptual connection. The lesson Davidson wants to teach with his example is that only the *descriptions* "the cause of *B*" and "*B*" are conceptually connected, whereas the *events* they pick out are causally connected. After all, the original sentence, "*A* caused *B*", is not an analytic statement. "The truth of a causal statement depends on *what* events are described; its status as analytic or synthetic [whether or not the connection is conceptual] depends on *how* the events are described" (Davidson 1963, 14).

Thus Davidson appears to be saying that even if the Logical Connection Argument is correct about the conceptual connection between our attributions of reasons and actions, it does not follow that the terms cannot pick out events that stand in a causal relation to one another. An intentional explanation talks about a certain process in *intentional* terms. These intentional *terms* or *descriptions* (or the attributions of reasons and actions) may be conceptually connected, much as the Logical Connection Argument argues. Since we are not Cartesian dualists, we shall nonetheless assume that the concrete particulars the intentional explanation refers to can be given also *physicalist* descriptions. To the extent that these new, physicalist descriptions (physical event-types) are covered by a causal law of nature, the connection between the particulars falling under such physical types is a causal connection no matter how we in fact happen to describe the particulars.

Note that Davidson is making an ontological point. For the argument to work we have to know neither the physicalist descriptions nor the law covering them; it suffices that the particulars could in principle be redescribed in law-governed physicalist terms.

If the law-covered redescriptions exist at least potentially, then, since the original intentional explanation talks in terms of mental types *about* the very same particulars as the alternative description does in terms of physical types falling under a causal law, the connection the intentional explanation refers to is in fact a causal connection, even if the intentional explanation fails to reveal that (see Davidson 1967, 155 & 159-160). In this case, Davidson argues, also the intentional explanation can be said to be a causal explanation.

However, it seems to me that the sense in which an intentional explanation can be a causal explanation in Davidson's theory is obviously a *derived* or even *contingent* sense. Let me explain what I mean by this.

For Davidson an adequate intentional explanation of action consists of a rationalisation *and* an identification of a causal connection (Davidson 1963, 4; 1974a, 233), and consequently the explanation must acknowledge the causal nature of the process the explanation picks out by including a premise asserting that the mental states cited as the reasons for action (here, a belief and desire) are also the causes of the action. It is important to understand that for Davidson's argument to work, such a clause must be seen as a singular causal claim stating that a certain desire and a certain belief caused a certain behaviour. This singular causal claim is about concrete particulars; it does not function as a general law connecting certain *types of propositional attitudes* with actions of certain kind (Stoutland 1976, 283-284). For Davidson, beliefs and desires are causally efficacious *qua concrete particulars* out there in the world.

Hence, we can model Davidson's view of intentional explanation as follows:

1. *X* desires *D*.
2. *X* believes that *A* is the best means to attain *D* under the circumstances.
3. *This concrete particular* instantiating *X*'s desire and *this concrete particular* instantiating *X*'s belief cause *this concrete particular* instantiating *X*'s action *A*.

(*L*) If any agent, *X*, desires *D*, and believes that doing *A* is the best means to attain *D* under the circumstances, then *X* does *A*.

---

*Ergo: X* does *A*.

Hence, in Davidson's view an adequate explanation of an action must, on the one hand, rationalise the action by placing it into the normative web of rational relations between the intentional *types* involved by describing the particulars the explanation talks about as tokens of mental types – just as the Logical Connection Argument says – *and*, on the

other, acknowledge the causal relations between the particulars in question by explicitly postulating that the relation between the *particulars* is a causal relation.

This reading of Davidson amounts to attributing more or less the same view of intentional explanation to Davidson as the one defended by Raimo Tuomela in his 1976 paper.<sup>110</sup> Both Davidson and Tuomela accept the core idea of the Logical Connection Argument (and Tuomela indeed puts this in terms of an explicit acceptance of the Logical Connection Argument<sup>111</sup>), namely that (*L*) is not a causal law of nature at all but serves to render an action intelligible by pointing to the conceptual or logical relation in which the action stands to beliefs and desires.

In fact, although Davidson does not explicitly accept the Logical Connection Argument, in his well-known writings on the nature of the mental (especially Davidson 1970 & 1974a) Davidson explicitly explains that his famous thesis of the anomalism of the mental (that mental types are not covered by strict laws) is based on the *holistic*<sup>112</sup> nature of the mental, *i.e.*, the view “that the content of a propositional attitude derives from its place in the pattern” (Davidson 1970, 221, see also p. 217 & 1987, 114 and, especially, 1995, 130) and therefore “[t]he meaning of a sentence, the content of a belief or desire [or indeed action], is not an item that can be attached to it in isolation from its fellows” (Davidson 1982, 183). Moreover, “the satisfaction conditions of consistency and rational coherence may be viewed as *constitutive* of the range of applications of such concepts as those of belief, desire, intention and action” (Davidson 1974a, 237; my italics – see also Davidson 1982, 184). As Jaegwon Kim (2003) puts this, for Davidson

[b]eliefs, intentions, and the rest are possible only as elements of an integrated, “holistic” system, and what give the system intelligible structure are the principles of rationality, consistency, and coherence. For Davidson, the norms of rationality and coherence [such as (*L*)], which underlie mental holism, are the “constitutive principles” of mentality; they give intentional mentality their distinctive identity as an autonomous domain.  
(Kim 2003, 119.)

<sup>110</sup> Tuomela (1998), on the other hand, defends a version of the Standard View – and, accordingly, I think the 1976 paper is actually more successful (see III.2 below).

<sup>111</sup> In his later writings also Davidson (*e.g.*, 1982, 173 & 174) talks explicitly about a logical connection.

<sup>112</sup> “Holistic” in the Quinean sense, not in the sense of social holism as the word is mainly used in this dissertation. Similarly, Dagfinn Føllesdal (1982) argues that Davidson’s theory of action in a sense widens Quine’s *meaning holism* to include actions and not only mental states such as beliefs and desires. Quine’s holism holds that the meaning of a belief is constituted by its place in a larger conceptual system (a web of beliefs). Since actions are routinely defined, following Weber, as *meaningful* behaviours, it is natural to think of actions as including propositional contents that are *constituted* by their place in a conceptual system that includes, *e.g.*, reasons. This view is of course highly compatible with von Wright’s Logical Connection Argument. I address the relation between Quinean meaning holism and social holism in III.3 onwards (including Appendix).

But this is of course precisely the claim of von Wright's Logical Connection Argument. Principles such as (L) constitute the conceptual framework within which we can talk about reasons and actions.

Moreover, the place in the conceptual framework that constitutes the contents of reasons and actions is defined in terms of normative, conceptual connections, and hence when we approach behaviour in terms of its reasons (*i.e.*, qua meaningful action) "we necessarily impose conditions of coherence, rationality, and consistency. *These conditions have no echo in physical theory*, which is why we can look for no more than rough correlations between psychological and physical phenomena." (Davidson 1974a, 231; my italics.) In other words, normative constraints and constitutive principles such as (L) are not descriptive causal claims.

The claim that the constitutive principles of the mental domain have no echo in the constitutive principles of the space of nature (roughly, causal laws of nature governing the physical domain) is a cornerstone of Davidson's philosophy of mind. A proper understanding of this thesis is absolutely vital also for the present study. The thesis is examined and defended in detail in III.3 and III.4 below, but let me shortly sketch the main line of thought also here (my sketch follows partly the wonderfully clear discussion of Kim (1985, 200 ff. & 2003, 119)).

Suppose one believes, as my first English textbook at school used to declare *ad nauseam*, that Spot is a dog. Davidson's holism of the mental – or (Quinean) *meaning holism* for short – states that the propositional content of this belief (in virtue of which the belief is intentional) is constituted by the rules of rationality that require one to hold also the beliefs that Spot is a mammal, Spot is not a mere machine, Spot is a living creature and so on. Since these requirements are constitutive of the belief qua contentful, if one does not believe these other things (or a large number of such beliefs) one does not hold a contentful belief at all.

Call the belief that Spot is a dog *M*. Suppose we are now wondering whether we should attribute the person who believes *M*, that Spot is a dog, also the belief *M\**, that Spot is a living creature, or the belief *M'*, that Spot is a furry, inanimate machine. Suppose further that available perceptual evidence does not discriminate between *M\** and *M'*. However, in Davidson's view the rules of rationality require the person who accepts *M* to accept *M\** and to reject *M'*, for *M'* is not rationally compatible with *M*. Similarly, as interpreters we rationally ought to attribute the person holding *M* also the belief *M\** and not *M'* (recall Davidson's (1973b) famous principle of charity). It is absolutely crucial to understand that in Davidson's picture the rules connecting *M* to the acceptance of

$M^*$  and to the rejection of  $M'$  are *normative* rules partly *constitutive* of  $M$ ,  $M^*$  and  $M'$  qua contentful judgements.

Suppose now that the claim that the rules of rationality have no echo in the physical domain is false. Presumably this would mean that there are purely physical, neural events  $N$  and  $N^*$  that are sufficient (although not necessary, if we want to make room for the multiple realisability of mental states – see III.2 below) for  $M$  and  $M^*$  to occur, respectively. Since by assumption the connection between  $M$  and  $M^*$  is reflected in the law-like constitutive principles of the physical domain, it is a law of nature that whenever  $N$  (which is sufficient for  $M$  to occur) occurs,  $N^*$  (sufficient for  $M^*$ ) occurs as well. This, however, is highly implausible, for it – and therefore the assumption that Davidson's no echo thesis is false – has quite unacceptable implications.

First, the *law* connecting  $N$  and  $N^*$ , and thereby  $M$  and  $M^*$  as well, would imply that the connection between  $M$  and  $M^*$  is not really a *normative* connection after all, but merely a projection of a non-rational (non-normative) causal law governing physical events. "In consequence, the intentional mental domain would be threatened with a loss of its distinctive identity, which is defined by norms of rationality and coherence" (Kim 2003, 120). Therefore, insofar as Davidson is correct in holding that the mental realm is *essentially* normative (in the sense of the normative rules of rationality being constitutive of mental states),<sup>113</sup> the rejection of Davidson's no echo thesis would in effect lead to the elimination (or reduction) of the mental realm.

Second, in terms of the example above, in Davidson's picture it is essential that the choice between  $M^*$  and  $M'$  is based on irreducibly *rational* (conceptual, normative) constraints. These rational constraints give, according to Davidson,  $M^*$  and  $M'$  their contents, *i.e.*, constitute  $M^*$  and  $M'$  qua mental states. Now if we reject the no echo thesis and assume that the occurrence of the neural state  $N^*$  is sufficient for  $M^*$  to occur, the rational constraints pointing to  $M^*$  instead of  $M'$  lose their essentiality. To recapitulate, for Davidson a constitutive feature of  $M^*$  is that the conditions of its attribution are essentially normative. However, the neural state  $N^*$  presumably has purely naturalistic (non-normative) attribution conditions. Thus, if  $N^*$  is sufficient for  $M^*$ ,  $M^*$  has sufficient, non-normative attribution conditions and, therefore, the normative attribution conditions of  $M^*$  lose their essentiality and, as before, this implies that the mental domain is either eliminated or reduced away.

---

<sup>113</sup> The argument that normative rationality is indeed necessary for mental qua mental and qua contentful (but that this does not challenge naturalism) is defended in detail in Chapter III.3 onwards.

These two ways of explaining the implicit content of Davidson's no echo thesis are based on the assumption that there is a neural event identical<sup>114</sup> with the mental event. However, rather than thinking about particular neural events, an anti-Cartesian monist could argue that the instantiation of a mental state depends on a wide system of causal processes, including, say, the agent's physical and social environment. Such a causalist could maintain that the physical materialiser that gives rise to the mental state is a system of a number of causal capacities, *each essentially different* from the capacity ascribed to the mental state. Thus, the causalist attribution conditions of any of the physical causal capacities would not be identical with the attribution conditions of the causal capacity of the mental state, although nothing Cartesian has been assumed. In such a case the mental state could be argued to have a causal capacity different from the capacities of any physical event (I take this to be close to what Dupré (1993) is saying – see III.2.5).

However, it seems to me that Davidson's no echo thesis is meant to cover also this line of thought. Post 1963, Davidson *avoids* talking about mind-brain supervenience, and this seems to suggest that he does not want his view to be restricted to views that see a neural event as sufficient for a mental event. The core of his no echo thesis appears rather to be the argument that if we try identify (or indeed talk about) the mental in causal terms, we have changed the subject, for the mental is *constituted* by addressing it in normative terms, whereas causal talk is always non-normative.

Thus, the philosophical core of Davidson's no echo thesis (and therefore his Anomalous Monism) is the mutual incompatibility of the following three theses: (i) the mental realm is essentially normative (it belongs to the logical space of normative rationality), (ii) the physical (causal) realm is essentially non-normative (it belongs to the logical space of non-normative nature) and (iii) there are systematic connections between the mental realm and the physical realm. Davidson argues that one of the three must be rejected, and he rejects (iii). However, this is compatible with Davidson's anti-Cartesian monism, since for Davidson both the physical and the mental realm are essentially conceptual categories, *i.e.*, incommensurable ways of conceptualising the world of concrete particulars either as tokens of mental types (the mental realm) or as tokens of physical types (the physical realm). The two conceptual frameworks, although about the same world, are not and cannot be systematically connected.

Contrary to what is often claimed, the no echo thesis implies that *intentional explanations of action cannot be causal explanations*. Intentional explanations operate

---

<sup>114</sup> At least *token*-identical; see III.2 below.



within the mental realm, whereas causation belongs to the physical realm. And since there are no systematic connections between the two realms, one cannot smuggle causal notions from the physical realm to the mental realm.<sup>115</sup>

Thus, it seems obvious that the notion of the normative holism of the mental, which, according to Davidson, implies anomalism, amounts in fact to the acceptance of the Logical Connection Argument in von Wright's sense. The *explanans* of the standard form of intentional explanation that consists of Premises 1 and 2 and the principle (*L*) has *nothing causal in it* (since the premises "have no echo in physical theory"). However, unlike von Wright, Davidson still wants to stick to the view that a successful intentional explanation must in some sense be a causal explanation, and hence he adds (implicitly) Premise 3 that contains all the causal import there is in the *explanans*.

The main reason for Davidson's insistence on the causal nature of intentional explanation is, as Tuomela (1976, 202) explains, that a mere rationalisation will not in Davidson and Tuomela's view allow us to pick out *X's operative reason* out of the class of *possible rationalisations* of *X's* action. For example, we could rationalise *X's* going to the theatre by pointing out that *X* desired to see the play and believed that the way to do that is to go to the theatre. But we could also point out that as a collector of autographs *X* desired to get the autograph of his favourite actor and believed that this can be done by going to the theatre. Now the question is, which intentional explanation is the correct explanation, *i.e.*, which explanation captures the operative reason for *X's* action. Tuomela's answer (1976, 203), which agrees with Davidson's original theory, is that the correct explanation is the one referring to those reasons that *caused X's* action. The operative reason is the cause of the action.

However, this means simply that we have to add a causal element (as expressed by Premise 3 above) to the non-causal, standard form of intentional explanation. Here is the point where it matters that we see the singular nature of the causal claim correctly. Premise 3 talks about the particular tokens of belief and desire *qua* concrete particulars. As Tuomela (1976, 203) emphasises, as it stands the singular causal claim of Premise 3 is indeed just a claim and hence in need of justification. In the view of Tuomela and

---

<sup>115</sup> It is important to notice that Davidson's no echo thesis does not appeal to the exact form of the causal relations governing the physical realm (or even to mental states being also physical states) but only to the essential difference between normative and non-normative relations. In III.2 I discuss views that reject the essential normativity of the mental realm and seek to account for the irreducible and ineliminable nature of the mental in terms of irreducible mental causation. I argue that most of these attempts fail in their own terms. However, even those views that survive my criticism in III.2 (notably views that do not suppose that the physical domain is causally closed) require that one denies either the essential normativity of the mental or the non-normativity of the causal. In III.3 onwards I argue that neither can be done. The problems created by my unbridgeable gap between the normative mental realm and the non-normative physical (causal) realm are discussed in Appendix.

Davidson, scientific causal claims must be backed by means of some laws, since “where there is causality, there must be a law: events related as cause and effect fall under strict deterministic<sup>116</sup> laws” (Davidson 1970, 208). However, since psychological concepts and connections (such as *(L)*) “have no echo in physical theory”,<sup>117</sup> the backing laws must be physical. This, as Davidson explains, amounts to redescribing the concrete particulars Premise 3 talks about as tokens of physical types and subsuming these descriptions under a causal law.

Crucially, such a law does not connect reasons with actions, *i.e.*, mental or intentional types (see Stoutland 1976, 285), but physical types with other physical types.

The Laws whose existence is required if reasons are causes of actions do not, we may be sure, deal in the concepts in which rationalizations must deal. If the causes of a class of events (actions) fall in a certain class (reasons) and there is no law to back each singular causal statement, it does not follow that there is any law connecting events classified as reasons with events classified as actions – the classifications may even be neurological, chemical, or physical. (Davidson 1963, 17.)

Suppose *m*, a mental event, caused *p*, a physical event; then, under some description *m* and *p* instantiate a strict law. This law can only be physical [...]. But if *m* falls under a physical law, it has a physical description; which is to say it is a physical event. (Davidson 1970, 224.)

Justifying the insistence on causation requires that we think of the particulars in question qua physical events – and in particular not qua reasons (since reason-talk has “no echo in physical theory”).

Hence, according to the Davidson-Tuomela view, intentional actions include two components that must, because of the no echo thesis, remain essentially separated: “Two ideas are built into the concept of acting on a reason (and hence, the concept of behaviour generally): the idea of cause and the idea of rationality” (Davidson 1974a, 233). This, I think, implies that Davidson and Tuomela are committed to the view that in fact there is no one form of explanation that would be suitable for intentional action. First, we need the traditional model of intentional explanation, which rationalises the action in a non-causal way. Second, we need a causal explanation of the behavioural component of the action in question, and this is given by a non-intentional causal expla-

<sup>116</sup> For the role of determinism, see III.2.5.

<sup>117</sup> In Tuomela’s (1976) view this anomalism of the mental (the lack of causal laws in psychology) is an *empirical* claim. As we saw, Davidson is prepared – and I agree – to make a stronger claim: he thinks that the holism of the mental shows that causal psychological laws are *a priori* impossible.

nation that does not invoke norms of rationality such as (*L*) but causal laws connecting physical types.

This, however, simply means that the Davidsonian model above tries to merge two completely different explanations (different *explananda* and different *explananses*) into one inference. One is a causal, non-intentional explanation of behaviour, and the other is a non-causal, intentional explanation of action. The obsession to formally include both these aspects in a single inference does not undermine the fact that there are two different issues at stake.

But this does not speak against the Logical Connection Argument anymore; in fact, it *is* the point of the Logical Connection Argument. As von Wright puts it, the *explanandum* of a causal explanation of behaviour is “why *parts of his body move*, under the causal influence of stimulations of his nervous system, and not why *he moves parts of his body*” (von Wright 1971, 119), which is an action and consequently must be explained in terms of a non-causal intentional explanation. The Logical Connection Argument is not aimed to show that both explanations cannot be important and interesting, but only that one ought not to confuse the two, as they are genuinely different explanations. This is something Davidson’s view does not – and cannot – challenge, for Davidson is simply saying that an intentional explanation can pick out a connection which is (also) a causal connection, but nonetheless the intentional explanation cannot talk about the connection qua causal (because of the no echo thesis).

As von Wright (1971, 119) emphasises, when we explain *actions*, we give non-causal, intentional explanations where mental types stand in rational relations. When we explain bodily behaviour, we give causal, non-intentional explanations where physical types stand in causal relations. Rather surprisingly, the same result follows from a careful reading of Davidson’s argumentation, although Davidson is generally regarded as the philosopher who refuted the Logical Connection Argument. If Davidson still wants to insist that what he says is sufficient for calling intentional explanations causal explanations, we can see that he must mean this in a very weak, derived sense, since the causal nature is not due to the intentional explanation as such, but to the possibility of redescribing the components of the intentional explanation in terms of a causal explanation of behaviour – which, as Davidson himself admits, amounts simply to “changing the subject” (Davidson 1974a, 230). Von Wright would not disagree with this.

Now we can finally return to the questions I asked in the end of Section III.1.2. It is often simply taken for granted that Davidson’s defence of the causal interpretation of intentional explanations shows that the Logical Connection Argument is invalid, full

stop. I suggested that such an unqualified claim is not likely to be acceptable. Indeed, if the core of the Logical Connection Argument is thought to be the claim that intentional explanations as such are not predominantly causal explanations, I conclude that Davidson's alleged refutation fails, since even in Davidson's picture the causal nature does not come from intentional explanations *per se* but is derived from the possibility of physicalist redescriptions that can be subsumed under causal laws. Similarly, if the core of the Logical Connection Argument is thought to be the idea that (*L*) does not express a causal law of nature but rather a conceptual or normative connection (that has "no echo in physical theory"), I conclude that Davidson's alleged refutation also fails.

However, if one regards the Logical Connection Argument primarily as an attempt to show that mental terms such "belief" and "desire", as they are used in intentional explanations, cannot talk about processes of which also a non-mental description may be given and which in this sense can feature in causal explanations, I think Davidson's refutation is successful.<sup>118</sup> But this should not be seen as a refutation of the Logical Connection Argument. Rather, this should be seen as an *acceptance* of the Logical Connection Argument with an explicit emphasis on the importance of the rejection of Cartesian dualism and the role of causal knowledge in modern descriptive (natural) science. Again, I doubt if a sophisticated advocate of the Logical Connection Argument, such as Georg Henrik von Wright, would call into question any of this (see, *e.g.*, von Wright 1989b, 806 & 1988).

Thus, in fact Davidson's argumentation implies a direct denial of the Standard View. Indeed, the Davidsonian arguments, if correct, render the Standard View *impossible*. This is rather surprising, since the Standard View is often explicitly supposed to be an essentially Davidsonian position. This is, for example, the claim of Jon Elster (1985), whose argumentation I will next turn to.

Elster starts with the customary assertion that he wants to analyse rational or intentional explanations of *actions* as opposed to mere behaviour. In his view this "amounts to demonstrating a three-place relation between the behavior (B), a set of cognitions (C) entertained by the individual, and a set of desires (D) that can also be imputed to him." (Elster 1985, 311).<sup>119</sup> Elster thinks that the relevant relation between

---

<sup>118</sup> However, it does not follow that mental states necessarily supervene on brain states; the view leaves open the possibility of *externalism*. Perhaps mental states supervene on a system significantly larger than mere internal brain states, *i.e.*, on a system including the agent's (physical and social) environment.

<sup>119</sup> Already here it is clear that this cannot be a Davidsonian view, for Davidson (1970, 1974a) was forced to admit that von Wright was right to think that reasons and actions (qua reasons and actions) are not distinct events but dependent on each other (the holism of the mental) and, moreover, exist only relative to the normative framework constituted by norms such as (*L*). Elster's very starting points show that by treating actions, cognitions and desires as distinct categories he fails to appreciate the mental holism char-

beliefs (or cognitions) and desires on the one hand and behaviour on the other can be defined in terms of three conditions. If the conditions are satisfied, we have reached a successful intentional explanation, and the behaviour in question has been shown to instantiate an action. I have paraphrased the conditions in the end of Section III.1.1. Now I wish to add a critical discussion concerning the conditions. My aim is to show that if the background theory really is Davidson's theory, as Elster says, then the three conditions necessary for the Standard View *cannot* be satisfied.

In Elster's view a successful intentional explanation must be a causal explanation, but since we are talking about causal explanations applicable to rational beings, Elster's first condition requires that "the desires and beliefs are *reasons* for the behaviour" (Elster 1985, 311), *i.e.*,

(i) Given C, B is the best means to realize D. (Elster 1985, 311; cf. Davidson 1963, 5.)

In the spirit of the standard causalist approach (recall Davidson and Tuomela's arguments above), Elster thinks that (i) cannot alone be sufficient. Besides the problem of multiple rationalisations, an agent may have the right belief and desire, and perform the required behaviour without the behaviour being his intentional action. For example, an actor may be asked to shudder as a part of a scene. He may have the appropriate beliefs and desires, but the real explanation of his behaviour may still be that he saw a snake and this caused him to shudder. In such a case (i) would be satisfied, but the actor nonetheless did not shudder intentionally. Shuddering was not what he did, it was not his action. Rather, it happened to him.

Elster thinks that this shows that we need to add another condition "ensuring that his behavior was actually caused by his intention to behave in that way" (Elster 1985, 311). Elster formulates this condition as follows:

(ii) C and D caused B. (Elster 1985, 311; cf. Davidson 1963, 12.)

The combination of (i) and (ii) is of course the core of Davidson's causal theory of intentional explanation. In the original version of Davidson's 1963 article, Davidson states (Footnote 5) that in addition to considering (i) and (ii) to be necessary conditions for intentional action, he also believes that (ii) "can be strengthened" to make (i) and (ii)

---

acteristic of Davidson's philosophy of mind (recall also the quotation from Stoutland 1976, 279 in III.1.2).

jointly sufficient for intentional action. In the 1963 article Davidson nonetheless concentrates exclusively on the necessity of the conditions.

However, in “Freedom to Act” (Davidson 1973a) Davidson returns to sufficiency. His view is formulated in relation to the following example:

A climber might want to rid himself of the weight and danger of holding another man on a rope, and he might know that by loosening his hold on the rope he could rid himself of the weight and danger. This belief and want might so unnerve him as to cause him to loosen his hold, and yet it might be the case that he never chose to loosen his hold, nor did he do it intentionally.  
(Davidson 1973a, 79.)

In this case both (i) and (ii) are satisfied, but as Davidson explicitly admits, the climber does not loosen his hold intentionally. It is not an action of his. Conditions (i) and (ii) are, Davidson thinks, necessary, but the example shows that they cannot be sufficient:

If the agent does *x* intentionally, then his doing *x* is caused by his attitudes that rationalize *x*. But since there may be wayward causal chains, we cannot say that if attitudes that would rationalize *x* cause an agent to do *x*, then he does *x* intentionally.  
(Davidson 1973a, 79.)

Hence, if the causalist interpretation of intentional explanation is to be defended, something needs to be added to (i) and (ii) – or, as Davidson put this in his 1963 footnote, (ii) must be strengthened. The required further condition is that the beliefs and desires must cause the action “in the *right* way” (Davidson 1973a, 79). We must require that the “causal chain must follow the right sort of route” (Davidson 1979, 78).

But this is hardly satisfactory. We want to know what “the right way” is. This is where Davidson’s confident argumentation starts to shake. Instead of giving a clear answer, he suggests vaguely that “we might try saying” that the belief and desire must cause the action “through a course of practical reasoning” (Davidson 1973a, 79) so that the action is, and here Davidson is paraphrasing David Armstrong in a non-committal way, “produced by a causal chain that answers, at least roughly, to the pattern of practical reasoning” (Davidson 1973a, 78). All this is terribly imprecise, which Davidson is well aware of: “What I despair of spelling out is the way in which attitudes must cause actions if they are to rationalize the action” (Davidson 1973a, 79).<sup>120</sup>

Elster, on the other hand, does not lose hope. He thinks that the problems in the example are due to the fact that it is “a mere accident that in the case of the climber they

---

<sup>120</sup> Note also that Davidson’s problem is precisely that he seems not to be able to get rid of *normative* notions, just as the Logical Connection Argument implies.

[the belief and the desire] happen to cause the very same behavior for which they are reasons” (Elster 1985, 312). As soon as we see this, Elster thinks, the solution is obvious. To rule out situations such as the climber example, we must add a third condition stating that in order for a belief and a desire to explain intentionally an action, the belief and the desire must have produced the causal effect they did in virtue of being reasons. Or in Elster’s notation:

(iii) C and D caused B qua reasons. (Elster 1985, 312.)

Elster thinks that (iii) offers the kind of strengthening Davidson’s original footnote in the 1963 article envisaged and the 1973a climber example requires. And as I explained in III.1.1, together (i), (ii) and (iii) constitute the Standard View.

Admittedly Elster’s reasoning is quite appealing. To save the essential intuitions behind the Standard View, namely the idea that human reasoning must matter (causally) to our actions, something like (iii) must indeed be added. However, for Davidson it is clear that we *cannot* add anything like (iii). As I have explained in detail above, a crucial part of Davidson’s reasoning is that we can treat intentional explanations as causal explanations precisely because the concrete particulars (or tokens) an explanation talks about do not have to be treated as reasons. Causation enters the picture precisely to the extent that the tokens are considered qua physical entities in general and not qua reasons in particular.

In sum, the whole Davidsonian possibility of treating intentional explanations as causal explanations is based on the explicit denial of Elster’s condition (iii). The reason why Davidson does not in fact support the Standard View and why he does not introduce a third condition similar to that of Elster is that his argumentation is incompatible with the Standard View as defended by Elster. If Davidson is correct, then Elster’s position is inherently and inescapably incoherent. In the Davidsonian context, Elster’s inquiry into the necessary and sufficient conditions of the Standard View has turned into a strong argument against the Standard View by exposing its inbuilt impossibility.

Rather revealingly, Davidson *omits* the suggestion that the conditions (i) and (ii) could be strengthened to make them jointly sufficient to define the relation between reasons and the actions they explain from the 1980 reprint of his 1963 article. In the 1980 version the message of Footnote 5 (p. 12 of the reprint) has been completely inverted.<sup>121</sup> The new version seeks “to cancel any suggestion” that the conditions could be suffi-

---

<sup>121</sup> I thank Damien Fennell for drawing my attention to this interesting detail.

cient, whereas the original version expressed very explicitly Davidson's suggestion that precisely such a strengthening can be done. I think Elster has shown that for such a strengthening to deliver the required conclusion (resulting in the Standard View) we would need to add something that cannot be added, *i.e.*, Elster's condition (iii). Thus, in his introduction to the *Essays on Actions and Events* (Davidson 1980) Davidson explains that the reasoning behind his 1973a article "Freedom to Act" forced him to conclude that intentional action "cannot be analysed or defined" (Davidson 1980, xvii) in purely causal terms. A causal theory cannot give "sufficient conditions of intentional (free) action" (Davidson 1980, xvii; see also Davidson 1974a, 232).

Indeed, the conclusion of Davidson's 1973a article, that "we cannot see how to complete the statement of the causal conditions of intentional action" (Davidson 1973a, 80), is a direct denial of both Davidson's own optimistic original Footnote 5 in Davidson 1963 and Elster's Standard View. As Davidson admitted in 1980, the causalist interpretation of intentional explanation cannot succeed "at least without appeal to the notion of intention" (Davidson 1980, xvii). The climber's loosening of his hold, thus relieving himself from the weight and danger, was his intentional action only if by his behaviour he intends to relieve himself from the danger and weight. Stoutland gets this exactly right:

The causation of the behavior is not relevant; what is relevant is that the agent intends a result by it, so that the behavior is understood in terms of the intention. In this way the intention relates to the behavior in that direct way required to rule out the aberrant case [Davidson's climber example].  
(Stoutland 1976, 291.)

As Davidson says, there is no hope of success without an irreducible appeal to intentionality.

Hence we are back with the Logical Connection Argument. Davidson's revised view amounts, of course, to what von Wright had been saying all the time. We cannot explain and understand *actions* by simply looking at their causes. What makes a behaviour an action is precisely that it is intentional. Moreover, "the intentionality is *in* the behavior [...], intentionality is not anything 'behind' or 'outside' the behavior." (Von Wright 1971, 115.) The intentionality of actions cannot be explained or reduced away. It is a conceptually or logically necessary component of actions, just as the Logical Connection Argument says – and as Davidson admitted explicitly in 1980.<sup>122</sup>

---

<sup>122</sup> Thus my reading of Davidson – unusual as it may be – agrees largely with at least one well-known interpretation of Davidson: Jaegwon Kim sums up his reading of Davidson as follows. "The view of psychology that emerges from Davidson is one of a broad interpretative endeavor directed at human action,



The upshot of Davidson's alleged refutation of the Logical Connection Argument is that the Logical Connection Argument is alive and well; what must be rejected is the Standard View as exemplified by Elster (1985).

---

to understand its 'meaning' rather than search for law-based causal explanations that are readily convertible into predictions; psychology is portrayed as a hermeneutic inquiry rather than a predictive science" (Kim 1985, 211). This could as well be a description of von Wright's position, and consequently I find it surprising that Kim does not consider von Wright's view to be very sophisticated (Kim 1991, 294) and that, moreover, Kim uncritically accepts the (mistaken) received view, according to which Davidson (1963) refuted the Logical Connection Argument (Kim 1998, 63; see, however, also Kim 1985, 195 where he seems to group Davidson together with the Wittgensteinians).

## CHAPTER III.2:

## THE MULTIPLE REALISABILITY ARGUMENT FOR MENTAL CAUSATION

## III.2.1 INTRODUCTION

The Davidsonian approach cannot be used to defend the Standard View, which amounts to seeing the reasons for action as the causes of the action in question. Davidson's anomalous monism implies that the assessment of reasons for actions is – in virtue of the holism of the mental, or the Logical Connection Argument – done in a framework characterised by rationality considerations and other *normative* issues, whereas talk of causation belongs to the framework of physical laws, brain states, behaviour and other non-normative *descriptive* matters of fact. If we insist, as the Standard View does, that agency is first and foremost a causal affair, then the Davidsonian line of thought is disastrous, for the contents of mental states and the meanings embodied in bodily behaviour cannot have anything to do with agency. Thus, *to the extent that* causation is indeed essential for agency, Davidson's view “implies that *what* we believe, intend, and desire has no bearing on what we do. It implies that *what a person thinks* has as much relevance to what he does as *what a sound means* has to the amount of pressure it exerts on a glass.” (Dretske 1989, 3.)<sup>123</sup>

Hence, if the Standard View is to be defended, we need a theory that allows mental states to be causally efficacious qua mental. In other words, the causalist interpretation of intentional explanations can be saved if there is a way of allowing mental types to feature in causal laws in a manner that is not simply derived from or parasitic on (and thus reducible to) standard physical causation in the sense of nomological connections between physical types. The task of demonstrating that there can be cases of irreducible mental causation in our thoroughly physical world is the bold aim of the doctrine of mental causation and non-reductive materialism.

I have already shown that both Davidson and von Wright think that this task is doomed to fail, since the connections between mental types are normative, rational connections and as such “have no echo in physical [causal] theory” (Davidson 1974a, 231). These normative relations are essential for and constitutive of the mental domain (the

---

<sup>123</sup> It may be a little misleading to appeal to Dretske here, since although I do think the arguments of this chapter apply fully also to Dretske's views, in what follows I concentrate on the problem of mental causation in general without paying explicit attention to Dretske's position. This is because I doubt if I can add anything original to Kim (1991), which I take to be a decisive criticism of Dretske's naturalism.

holism of the mental), and hence “[s]hort of changing the subject, we cannot escape this feature of the psychological; but this feature has no counterpart in the world of physics [realm of causation]” (Davidson 1974a, 230), and hence mental types cannot feature in causal relations.

I believe Davidson and von Wright are correct here. They both agree with Peter Winch’s observation, according to which when we try to speak of reasons and actions (qua reasons and action) in causal terms, “we literally do not understand what we are saying. We cannot understand it, because it has no sense.” (Winch 1958, 94.) As far as actions are considered, the focus must be on irreducibly normative relations, and to bring in causal connections amounts to changing the subject from actions to bodily behaviour, since what an intentional state (including actions) consists in is precisely its position in a web of normative connections.

Thus, in my view the attempt to salvage the Standard View on the basis of a theory of mental causation is deeply misguided. However, since the dual doctrine of mental causation and non-reductive physicalism tends to dominate contemporary philosophy of mind and action theory, I believe it is important to demonstrate its unacceptability in its own terms.

### III.2.2 THE MULTIPLE REALISABILITY ARGUMENT

We have already seen that Elster’s attempt to defend the causalist Standard View on the basis of Davidson’s philosophy of mind fails. However, most contemporary philosophers who accept the Standard View do not base their views primarily on Davidson, but on the so-called Multiple Realisability Argument.<sup>124</sup> As Elliott Sober (1999, 542) observes, the received view in contemporary philosophy of mind is that the Multiple Realisability Argument allows us to establish irreducible mental causation against eliminativism and reductionism without compromising the idea that, ontologically speaking, everything is fixed at the level of fundamental physics (cf., however, III.2.5).

<sup>124</sup> The argument is due to Putnam (1967) and it is discussed virtually in every serious paper on the mind-body problem published since that. In what follows I will concentrate mainly on the versions of the argument defended by Fodor (1974) and Tuomela (1998). On Fodor, because his article is clear and influential and on Tuomela partly because adopting his terminology makes it easier to connect the present discussion to the Logical Connection Argument of Chapter III.1 and ultimately to Parts I and II of this study, and partly because I find it interesting that both my criticism of the Multiple Realisability Argument (including Tuomela 1998) and the anti-causalist alternative I defend are largely inspired by my reading of Tuomela himself (1976 and 2002). My criticisms of Tuomela (1998) lead also directly to the further developments in III.2.5 and III.3 onwards. Excellent criticisms of other influential defences of the Multiple Realisability Argument and mental causation can be found in Kim (1998), which discusses, *e.g.*, Lynne Rudder Baker, Ned Block, Tyler Burge, Terence Horgan, Frank Jackson and Philip Pettit’s programme explanations and John Searle.

The problem is, in short, that if everything is ultimately physical, it seems to follow that everything that happens, including actions, are dictated by the laws of physics and thus there seems to be no room for responsible actions (things that we do as opposed to things that happen to us).<sup>125</sup> Von Wright aimed to rescue the relevance of agency by arguing that causation is irrelevant for actions, for actions are essentially normative issues that have no echo in causal relations. Views building on the Multiple Realisability Argument take a different route. They reject the essential normativity and non-causality of the mental and seek to show that mental states are causally efficacious qua mental, and therefore cannot be eliminated or reduced away. In this Chapter I seek to show that, despite its popularity, also this attempt to defend the Standard View is unsuccessful.

To recapitulate, the challenge of mental causation is to show that mental states can be causally efficacious qua mental states or, in other words, that mental states can feature in causal laws in virtue of being instances of a mental type. To use Tuomela's (1998, 7)<sup>126</sup> notation, we may use a four-place causal relation  $Q(m, M, b, F)$  to express a relation which holds just in case  $m$  qua  $M$  causes  $b$  qua  $F$ , where  $m$  and  $b$  stand for event-tokens and  $M$  and  $F$  denote predicates corresponding to mental or intentional types (including action-types). A typical case would be one where  $M$  expresses beliefs and desires rationalising an action of the type  $F$ , and  $m$  is an instance of  $M$  and  $b$  a behaviour instantiating  $F$ . General naturalism and ontological monism commits us to the view that the truth of " $Q(m, M, b, F)$ " entails the truth of " $m$  causes  $b$ " or, to stick to Tuomela's notation, " $C(m, b)$ " (mere singular causation), which in turn is a true causal relation (and not a mere accidental connection) only if there is another four-place relation  $R(m, P_1, b, P_2)$  such that  $P_i$  are physical types connected by a strict causal law (or a set of causal laws). In other words,  $R$  subsumes event-tokens  $m$  and  $b$  (qua instances of  $P_1$  and  $P_2$ ) under a physical law in the sense of the D-N model of explanation.

Now the problem of mental causation concerns the nature of the rationalising relation ( $L$ ) between  $M$  and  $F$ . The Standard View holds that  $Q$  expresses a *sui generis* causal relation and, consequently, that  $M$  and  $F$  are connected by a *sui generis* causal law or, in other words, that ( $L$ ) expresses a causal law of nature after all. Connections

---

<sup>125</sup> Thus, the questions addressed in this Part could easily be translated into the language of the so-called problem of free will. The view I ultimately defend is an updated version of Strawson's (1962) famous view, according to which even if causal determinism is true it is not a threat to human freedom, for human agency and intentional actions are not causal matters. Actions are, as Kant taught us, what we are *responsible for*, and hence belong to the normative space of reasons and not to the domain of causal connections and the natural sciences. As should become clear in Chapter III.3 onwards, I see my approach as related also to Dennett's (2003) recent compatibilist defence of free will.

<sup>126</sup> I have re-named the variables to be consistent with my other discussions.

such as the one between  $M$  and  $F$  (*i.e.*, ( $L$ )) are what the laws of the special sciences, such as psychology, consist in (Fodor 1974). Thus, the Standard View can be defended only if we can argue that  $Q$  is neither epiphenomenal, parasitic on nor reducible to  $R$ , while being nonetheless indispensable (so that we cannot be eliminativists about  $Q$ ), and this requires that there exists an independent causal law to connect  $M$  and  $F$  (Tuomela 1998, 7-8).

Tuomela (1998), for one, accepts these requirements. Thus, Tuomela's core problem is that although he accepts that "[a]ll concrete particulars are material" (Tuomela 1998, 4), to defend the Standard View he must be able to defend the thesis "that there are genuine, irreducible cases of mental causation" (Tuomela 1998, 1). For the Standard View to be accepted, mental causation must be an objective, irreducible feature of the world (Tuomela 1998, 4). In particular, mental causation cannot be a mere pragmatic notion or a statistical generalisation we use for practical purposes when we for one reason or another cannot be bothered to find out the real, fundamental causal factors in a given situation and the relevant laws of physics (III.2.4). This is what I call the *strong* notion of mental causation. My aim is to show that the Multiple Realisability Argument cannot deliver this, for the strong notion requires the mental types  $M$  and  $F$  to have causal powers *independent of* the physical properties  $P_i$ , and this is beyond the scope of the Multiple Realisability Argument.

Despite its Cartesian tone, mental causation in this strong sense is exactly what the causalist view on intentional explanation and human agency requires. In short, my intention to do something must matter causally to the movements of my body – and, more precisely, my intention must matter *qua* a mental event and not only *qua* a physical event. It is the *content* of my intention – *what* I set out to do – that must be causally efficacious if the causalist interpretation of intentional explanations and human agency is to be accepted (Stoutland 1988, 44). The Standard View stands or falls together with this extremely strong notion of mental causation. In Tuomela's terminology this amounts to the truth of the following thesis T\*:

T\*) (Em)(Eb)(EM)(EF)(M is mental & Q(m,M,b,F) & ~(Ep)(EP)(Q(p,P,b,F) & P is material)).

In words, there are singular events  $m$  and  $b$  and predicates  $M$  and  $F$  such that  $M$  is a mental predicate and  $m$  *qua*  $M$  caused  $b$  *qua*  $F$ , while there is no material event  $p$  and no material predicate  $P$  (especially no such predicate in the base supporting  $M$ ) such that  $p$  *qua*  $P$  caused  $b$  *qua*  $F$ .

(Tuomela 1998, 12.)

In standard cases we can require that  $m$  and  $p$  are identical in the causal sense of having the same causes and effects, *i.e.*,  $m =_c p$ .

Tuomela's goal is to argue that although  $C(m,b)$  and  $m =_c p$  together entail  $C(p,b)$ , "the causal identity of the events  $m$  and  $p$  does not give the analogous entailment for *qua*-causation [*i.e.*, the relation  $Q$ ]" (Tuomela 1998, 13), speaking thus for the independence of mental causation. It is not quite clear to me what Tuomela means by this. Presumably his thesis is that if we pick out  $m$  qua an instance of  $M$ , and see it as causing  $b$  qua  $F$ , we cannot pick out a causally identical event  $p$  qua an instance of a physical type  $P$  and see it as causing  $b$  qua  $F$ , even if by definition  $p$  does cause  $b$  (qua an instance of some physical type  $P_2$ ). In other words, from  $Q(m,M,b,F)$  and  $m =_c p$  one cannot conclude  $Q(p,P,b,F)$ . This holds, according to Tuomela (1998, 8 & 16), because

(i) the relation  $Q$  is backed by a causal law of nature connecting  $M$ -events with  $F$ -events

and thus

(ii) if  $m$  had not been  $M$  it would not have caused  $b$  qua  $F$ , even if  $m$  had caused  $b$ .

Tuomela explains that (ii) is simply a "more complex reading" (Tuomela 1998, 8) of (i), and hence he sees no need to discuss the transition from (i) to (ii) any further. I, on the other hand, think that this is precisely the point where the weakness of Tuomela's argument lies.

On the basis of III.1.3 we can see that (i) and (ii) are claims fundamentally different from one another. Clause (i) expresses a non-normative connection between  $M$  and  $F$ , whereas (ii) expresses a conceptual connection. After all, it was precisely the point of the Logical Connection Argument that when we talk in terms of intentional types (such as  $M$  and  $F$ ), to understand some event  $b$  as an intentional action  $F$ , is "to see its place in a larger conceptual structure" (Stoutland 1976, 279), and this conceptual structure – and hence the fact that  $b$  is  $F$  – is constituted partly by seeing some other events as reasons, *i.e.*, by seeing  $m$  qua  $M$ . What it *is* for an event to be an action (for  $b$  to be  $F$ ) is partly that reasons can be given and asked for it (*i.e.*, that  $m$  is  $M$ ).

Hence (ii) expresses a logical, conceptual or normative connection between  $M$  and  $F$ , which is of course radically different from the non-normative nomological connection (i) postulates between  $M$  and  $F$ . Davidson's mature (post-1963) view was that (ii) makes (i) impossible:

What lies behind our inability to discover [...] [causal] psychophysical laws is this. When we attribute a belief, a desire, a goal, an intention or a meaning to an agent, we necessarily operate within a system of concepts in part determined by the structure of beliefs and desires of the agent himself [posing *normative* rationality constraints on our attributions]. Short of changing the subject, we cannot escape this feature of the psychological; but this feature has no counterpart in the world of physics [in the realm of causation].  
(Davidson 1974a, 230.)

If this is correct, then Tuomela's (1998) programme cannot succeed, for the core of his argument is to *deny* that "[t]he causal powers of a mental event-token are completely determined by its material properties" (Tuomela 1998, 6), and he thinks that to do that it is sufficient to establish that "[i]f m had not been M then b would not have been [F]" (Tuomela 1998, 6). However, if von Wright, Stoutland and Davidson are correct (and I think they are), the latter statement is about conceptual (constitutive) rules and not about causation, and hence it is perfectly compatible with the former statement, which is a direct denial of mental causation.

In III.1.3 we saw that in 1976 Tuomela actually saw this even more clearly than Davidson by explicitly accepting the Logical Connection Argument and, consequently, admitting that the causal backing of intentional explanation must be due to a physical redescription of the event-tokens in question, to be subsumed under a *physical* law (Tuomela 1976, 203). I also showed how Davidson's views imply a fundamental incoherence of Elster's version of the Standard View by showing how Tuomela's clause (i) (which for the present purposes is equivalent to Elster's condition (iii)) cannot be accepted. Tuomela (1976) saw this, but, surprisingly, Tuomela (1998) returns to the Standard View. The reason is the – in my view very unfortunate – new orthodoxy of seeing the Multiple Realisability Argument as a way (or *the* way) to defend "a weak kind of emergentism or 'nonreductive materialism'" (Tuomela 1998, 4) that nonetheless is supposed to deliver strong mental causation and, therefore, the Standard View.

To see the problem of mental causation in its proper light it must be kept in mind that the contemporary problem of mental causation is a problem for materialists. The task is, as Tuomela explicitly admitted, to defend full-blown, irreducible mental causation in the world that is thoroughly physical. The philosophers I am mainly concentrating on in this Chapter (including, *e.g.*, Fodor, Kim, Papineau, Pettit and Tuomela) understand the "thoroughly physical" nature of the world as the combination

of the thesis that everything there is *supervenies*<sup>127</sup> on the physical and the principle of the causal closure of the physical domain.

The closure principle says that all physical events have sufficient physical causes. I shall not argue for this principle here, but let me point out that in the view of the mentioned philosophers the principle is (i) presupposed by our best science, (ii) accepted by all serious participants in the contemporary debate on mental causation and (iii) an essential aspect of general ontological naturalism and materialism, since its denial would entail that some physical events do not have sufficient physical causes but are magically caused by non-physical substances – and this would certainly be a form of Cartesian dualism (however, in III.2.5 we shall see that all these claims can be questioned). David Papineau (2002), for example, argues that the principle of the causal closure is a cornerstone of the contemporary world-view, since together with the causalist thesis that at least sometimes mental causes have physical effects it forms the ultimate argument for materialism. If the mental really has physical effects, then the mental must, ultimately, be also physical, since the physical domain is causally closed. If the mental is not ultimately physical, it cannot have physical effects.<sup>128</sup>

Thus, in the context of these assumptions we face the following dilemma. Either there are non-normative, causal psychophysical laws (*e.g.*, connecting reasons and actions) or there are not. If there are, then the principle of the causal closure of the physical domain seems to dictate that the mental causes are not really mental, or at least that their causal powers are reducible to physical causation, and hence the mental events are not causally efficacious qua mental. This kind of reductionism would prevent the possibility of strong mental causation, refuting thus also the Standard View. On the other hand, if there are no (causal) psychophysical laws, we are back with Davidson's anomalous monism, which "permits mental properties no causal role" (Kim 1989b, 270). As Kim sees, if the mental is causally inefficacious and we maintain a causalist view of

---

<sup>127</sup> I shall not enter into the debate on the correct technical formulation of supervenience (see, *e.g.*, Kim 1993) here. Important as that debate may be, for the present purposes a general notion, according to which *M*-properties are said to supervene on *P*-properties if it is not possible that two things should be identical in respect of their *P*-properties without also being identical in respect of their *M*-properties, is sufficient.

<sup>128</sup> In fact, Papineau (2002) thinks that we must add a third claim stating that there is no massive causal overdetermination: it is not the case that in addition to the sufficient physical cause a physical effect must have, there is always present also another sufficient, non-physical cause of the effect in question. Such massive overdetermination would certainly sound strange, and I think we can join Papineau in accepting the thesis of the non-existence of massive causal overdetermination. Kim (1993, 1998) also argues that there is no such overdetermination, but he thinks that this follows directly from the principle of causal closure – the counterfactual situations where we remove the physical cause the effect would nonetheless be there, since also the non-physical cause is assumed to be sufficient for the effect, and this would violate the principle of causal closure. See III.2.5 – where also accounts of mental causation that reject the closure principle (in particular, Dupré (1993) are examined – for further discussion.



human agency and intentional action, “anomalous monism [...] is a doctrine virtually indistinguishable from outright eliminativism” (Kim 1989b, 270), which, needless to say, would be just as devastating for the Standard View as reductivism.

Hence, the only way to rescue human agency and the possibility of intentional actions is to avoid the dilemma. As III.1.3 made clear, my view is that von Wright has got it right: we must admit that there are no causal psychophysical laws and hence the mental indeed is causally inefficacious qua mental. But this concession does not imply eliminativism regarding agency and intentional action, for we reject also the dogma of seeing agency and action as causal affairs. Kim seems to make room for this view by offering the following as the best way to coherently hold the Davidsonian position.

I suggest the following line of reconciliation: on Davidson’s account the mental can, and does, have its own ‘laws’; for example, ‘laws’ of rational decision making. The crucial point, though, is that these are *normative* rather than *predictive* [causal] laws.  
(Kim 1985, 211.)

This means that one way in which one might try to eliminate the incompatibility [between materialism and the relevance of the mental] is to interpret rationalizing [intentional] explanation as a fundamentally *noncausal* mode of understanding actions. I believe that this is an approach well worth exploring: a rationalizing explanation is to be viewed as a *normative assessment* of an action in the context of the agent’s relevant intentional states.  
(Kim 1989a, 240, Footnote 4).<sup>129</sup>

The fundamental motivation for the Multiple Realisability Argument is the view dominating contemporary philosophical community, according to which none of the three possibilities I have described above is acceptable. The claim is that we must reject reductivism, eliminativism and non-causalism. In other words, we must find a way to defend the Standard View. And we have to do all this without compromising materialism and the principle of the causal closure of the physical world. This huge – in my view, impossible – task is, the contemporary received view goes, attainable with the Multiple Realisability Argument. In what follows I argue that the argument fails. Whether we like it or not, as long as we remain materialists in this sense (as opposed to, for example, Dupré’s (1993) pluralism, which denies the principle of causal closure – see III.2.5) the choice between reductionism, eliminativism and non-causalism is all we have.

---

<sup>129</sup> In his 1998 book Kim nonetheless repeatedly confirms his commitment to the causalist programme and the importance of mental causation (although he argues that no acceptable defences of these positions exist thus far), implying that the non-causal normative view cannot in his mind be satisfactory.

According to the Multiple Realisability Argument, the dilemma of reductionism and eliminativism is based on a false dichotomy. It is not the case that either mental states are fully reducible to physical states (full-blown reductionism) or completely disconnected from them (psychophysical anomalism). What is needed is a *weak anomalism* where a physical state is still nomologically sufficient for a mental state it instantiates, but not *vice versa* (Kim 1989b, 273). In other words, the mental must be seen to supervene on the physical. This, the argument goes, can be achieved when we replace the so-called type-identity theory (e.g., Smart 1959) with token-identity theory.

The type-identity theory says that in a sense we can pick out the same natural kind by using both physicalist and mental concepts. The concepts, although having different senses, are nonetheless extensionally equivalent. Hence, according to the type-identity theory, when we have a psychophysical law connecting a mental state  $M$  with an action  $F$ , which seems to violate the causal closure of the physical domain (for  $M$  is not a physical predicate), there is a physical predicate  $P_1$  which is co-extensional with  $M$  as well as a physical redescription  $P_2$  of the action  $F$  such that  $P_1$  and  $P_2$  are connected by a physical law. Hence, according to type-identity theory  $M$ -events are identical with  $P_1$ -events and  $F$ -events are identical with  $P_2$ -events, and since  $P_i$  are proper physical events connected by a physical law, the connection  $M \rightarrow F$ , which first appeared to threaten the principle of causal closure, is perfectly compatible with the principle after all. The situation is as presented in Figure 1 below, where vertical lines represent (type) identities and horizontal arrows causal relations.

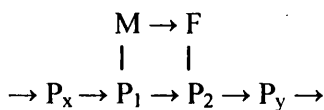


Figure 1: Type-Identity Theory

Here causation remains safely within the physical domain, and thus the principle of causal closure is not violated. But, of course, also mental causation is lost. In this picture the mental is not even epiphenomenal; it is wholly reducible into normal physical causation. Clearly this kind of reductivism cannot deliver the kind of *sui generis* strong mental causation Tuomela (1998) and other causalists are after.

This is where the Multiple Realisability Argument gets underway. As Fodor (1974, 689) points out, ontological materialism, including the principle of the causal closure of the physical domain, does not require mental predicates to be co-extensional with physical predicates. Instead of type-identity, let us try token-identity and a theory

that can be called token physicalism: “Token physicalism is simply the claim that all the events that the sciences talk about are physical events” (Fodor 1974, 689). All events are physical, but nonetheless non-physicalist concepts may reveal patterns which are real, objective features of the world but which cannot be captured if we stick to physicalist concepts (cf. Dennett 1991b). The Multiple Realisability Argument asserts that there are multiple ways in which a given mental event-type can be realised. In other words, where the type-identity theory requires that there are identity relations connecting mental types with physical types, the token-identity theory (or token physicalism) says that there are identity relations connecting only event-tokens.

To put it in Tuomela’s (1998) terminology introduced above, when we have a case in which an event-token  $m$  is an instance of a decision to perform an action of the type  $F$ , *i.e.*,  $m$  is an instance of  $M$  (and when  $m$  indeed causes an event-token  $b$  instantiating the event-type  $F$ ),  $m$  and  $b$  are identical with physical events. In other words, in addition to being instances of  $M$  and  $F$ ,  $m$  and  $b$  are also instances of some physical types  $P_l$  and  $P^*_l$ , respectively, although it is not the case that all instances of  $M$  are instances of  $P_l$  (and, similarly, all instances of  $F$  are not instances of  $P^*_l$ ), since the mental predicates carve up the world differently from physical predicates (although all instances of  $P_l$  are instances of  $M$ , and similarly for  $P^*_l$  and  $F$ ).

This way of putting the issue is important, since we should not let the language of identity relations *between* event-tokens lead us astray here: token physicalism is a form of ontological *monism*, which implies that when we focus on the cause-event  $m$ , there is only *one* concrete particular present,  $m$ , and the talk about token identity simply means that we can describe that one event-token as a mental event (as an instance of the mental type  $M$ ) or as a physical event (as an instance of some physical type  $P_l$ ). In particular, to say that in fact there are two particulars – a physical event and a mental event – would be a form of Cartesian ontological dualism.

Hence, the token-identity theory implied by the Multiple Realisability Argument holds that our use of mental types allows us to express law-like connections (such as the one between  $M$  and  $F$ ) that cannot be expressed in terms of physical types, for there are no physical types  $P_n$  and  $P^*_m$  that would be co-extensional with  $M$  and  $F$ , although all instances of  $M$  and  $F$  are instances of some physical types. Thus, Figure 1 is replaced with the following Figure 2.

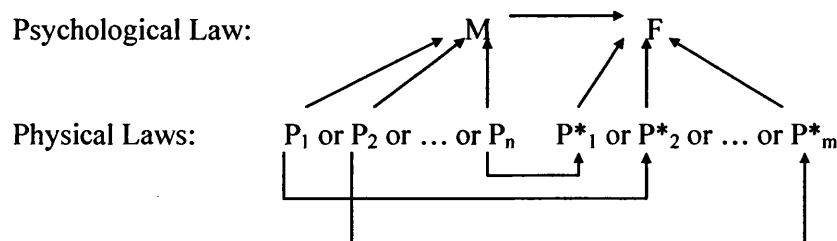


Figure 2: Token Physicalism with Multiple Physical Realisations of Mental Properties (Adapted from Fodor 1974, 695)

In Figure 2 the connections between the physical types  $P_i$  and  $P^*_j$  and the mental types  $M$  and  $F$  are not identity relations anymore, because that would, by definition, amount to the type identity theory. Rather, we must say that  $P_i$  and  $P^*_j$  provide multiple realisations of  $M$  and  $F$ , respectively. In accordance with the supervenience thesis,  $P_i$  and  $P^*_j$  determine the mental properties, but not *vice versa*. Hence we do not have the biconditionals connecting physical types with mental types required for traditional Nagel-reduction (Nagel 1961). None of the physical laws is identical with the mental law. However, the principle of the causal closure is not violated, for all *instances* of the law “if  $M$  then  $F$ ” are also instances of some physical law “if  $P_i$  then  $P^*_j$ ”, but none of the physical laws realising the psychological law is co-extensal with the psychological law.

In other words, because the mental types have multiple realisations, the mapping from physical types to mental types is many-to-one, not one-to-one as required for Nagel-reduction. However, the reason why the Multiple Realisability Argument is bound to fail to deliver the kind of strong mental causation where mental states really are causally efficacious qua mental, and which is required for the Standard View (and as we saw, for example Tuomela (1998) is very explicit about this requirement), is clearly visible here. The causalist view says that a particular is an instance of  $M$  in virtue of the causal powers it has. The principle of the causal closure of the physical world, however, requires that an instance of  $M$  has the causal powers it has *solely in virtue* of the fact that it is also an instance of whichever  $P_i$  applies to it. But this is simply another way of saying that all the causal work is done at the physical level, and hence it follows that “[t]he causal powers of a mental event-token are completely determined by its material properties” (Tuomela 1998, 6).

This implication, however, is of course precisely what the Multiple Realisability Argument was supposed to deny, for its denial is a necessary condition of strong mental causation and the Standard View. As Tuomela (1998, 6) puts it, the Standard View and

mental causation can be defended only if “[t]he causal powers of at least some mental events are not completely determined by their material properties”. My claim is, *pace* the Standard View, that also the Multiple Realisability Argument fails to deliver this, and hence I think the Standard View cannot be defended in terms of the Multiple Realisability Argument. There is no such thing as non-reductive but causal theory of mental states that stays within ontological naturalism and materialism that comprises causal closure and supervenience.

If mental predicates such as *M* and *F* are indeed irreducible to physical predicates (*i.e.*, if type-identity fails) but we nonetheless remain materialists in the present sense, the irreducibility shows at most that there is a way of *talking about* physical causation that cannot be translated directly into physics. There may be good practical reasons to keep the folk psychological language of mental predicates, but the Multiple Realisability Argument shows that we should not imagine that the language of mental causation amounts to anything more than to imprecise talk about normal physical causation. I think this is rather obvious, but since the denial, *i.e.*, the idea that the Multiple Realisability Argument allows us to combine strong mental causation with the principle of causal closure, is today so universally accepted, it is worth the extra ink to spell out this argument in detail. Before I do that, however, I want to return to Tuomela’s (1998) version of the Multiple Realisability Argument and demonstrate how it smuggles in the Logical Connection Argument. If I am right, Tuomela’s thesis implicitly presupposes the anti-causalist view the paper explicitly denies.

As we saw, the core of Tuomela’s account of mental causation is the claim that there are robust connections between mental types such as *M* and *F* that are not equivalent with (and reducible to) any physical connections. Fodor’s Multiple Realisability Argument seems to deliver that. Now Tuomela’s thesis *T\** adds to this the claim that although all the particular instances of the mental law (such as event-tokens *m* and *b*) are subsumable under some physical law (*m* and *b* are also instances of some physical types), nonetheless “[i]f *m* had not been *M* then *b* would not have been [*F*]” (Tuomela 1998, 6). In other words, Tuomela’s claim is that the irreducible connection between *M* and *F* guarantees that if *m* had not been *M*, it might have caused *b*, but not *b qua F*. I fail to see how he can hold this view if he really sticks to the non-normative conception of action.

The core of the problem is that within the present context of supervenient token physicalism it does not make any sense to imagine *m* being there at all without it being also *M*, and thus the counterfactual “if *m* had not been *M*, *b* would not have been *F*” is

rather futile.<sup>130</sup> The token identity theory says that each token of a mental type is identical with the particular, physical event-token that realises the mental property in that situation. There are not two particulars here, only one. In this instance the event of the occurrence of the mental state  $M$  of the person in question *is* the event-token  $m$ , where  $m$  (identified qua a physical state) *is* the neurophysiological state of the person or, if wide theories of content are correct, a wider state including the person and her environment.

According to the supervenience thesis, the fact that  $m$  (an instance of a  $P_i$ ) occurs is sufficient (although not necessary) for  $M$  to occur. Under the assumption of causal closure, to deny this is to subscribe either to Cartesian dualism or to Davidson's anomalous monism – both views that the Multiple Realisability Argument was designed to reject. Thus, the aspiration to hold to the causal closure and token physicalism in the sense of the mental supervening on the physical while at the same suggesting that supervenient causal relations would be independent of physical causation is deeply confused.

Let us assume that  $m$  causes  $b$ . Let us further assume that  $m$  instantiates a mental type  $M$  and  $b$  instantiates an action type  $F$ . According to token physicalism,  $m$  and  $b$  must have also physical descriptions. Let us thus assume that  $m$  instantiates also  $P_i$  and  $b$  instantiates  $P^*_i$ . Hence in this case there are only two event-tokens present,  $m$  and  $b$ , and thus although  $M$  is not identical with  $P_i$ , in this particular case *the instantiation of  $M$  simply is the instantiation of  $P_i$*  (and similarly for  $F$  and  $P^*_i$ ). Now recall Tuomela's central thesis  $T^*$ . According to it, if the connection between  $M$  and  $F$  holds, *i.e.*, if " $Q(m, M, b, F)$ " is true, there is no physical predicate  $P$  so that " $Q(m, P, b, F)$ " is also true. Tuomela even emphasises that there is "especially no such predicate in the base supporting  $M$  [...] such that [...] [ $m$ ] *qua*  $P$  caused  $b$  *qua*  $F$ " (Tuomela 1998, 12). However, it seems to me that  $P_i$  is exactly such a predicate. By assumption,  $P_i$  is in the base supporting  $M$ , and  $m$  *qua*  $P_i$  causes at least  $b$  *qua*  $P^*_i$ , and in this particular case  $b$  *qua*  $F$  is  $b$  *qua*  $P^*_i$  – unless we accept von Wright's Logical Connection Argument.

So in this situation at least the following three claims are, by assumption, true: " $C(m, b)$ ", " $Q(m, M, b, F)$ " and " $Q(m, P_i, b, P^*_i)$ ". More precisely, we have one concrete particular,  $m$ , causing another concrete particular,  $b$ , *i.e.*,  $C(m, b)$ . This is a brute fact about the world. The other claims are more complicated claims in the sense that their

---

<sup>130</sup> With a sufficiently rich theory of the semantics of conditionals and counterfactuals (*e.g.*, David Lewis' or Robert Stalnaker's theories) we can of course make sense of this counterfactual. Typically, however, such theories hold that since the antecedent of the counterfactual is nomologically impossible, the counterfactual is vacuously true regardless of the truth-value of the consequent. Hence this line of thought can hardly help to defend the causal efficacy of mental properties (cf. Sober 1999, 548).

truth depends on  $m$  causing  $b$  and  $m$  and  $b$  being event-tokens such that our predicates  $M$ ,  $F$ ,  $P_1$  and  $P^*_1$  apply to them. If intentional predicates  $M$  and  $F$  are used simply to pick out the causal powers of the particular they refer to, then there is no reason why we could not in Fodor and Tuomela's system – *in this particular case* – use, for example,  $F$  and  $P^*_1$  interchangeably.

It is important to understand this point correctly. The mental causation thesis holds that what matters here are the *description-independent* causal powers of the particulars picked out by our predicates, and in this case the physical and the mental predicate pick out the same particular. Thus, *in this case* and *for this purpose* the predicates can be used interchangeably. However, of course the predicates cannot be used interchangeably in our language games, for although the extension *in this case* is the same, the *types* as such are not identical – not in terms of extension and certainly not in terms of intension. Thus, if one follows Davidson and von Wright (and, ultimately, Wittgenstein – see Chapter III.3 onwards) to think that the rules governing the use of predicates is constitutive of the mental qua mental, the reductive thesis does not follow (because of the no echo thesis). But the Multiple Realisability Argument rejects explicitly this line of thought and defends the irreducibility of the mental in virtue of the independent causal powers of mental-tokens (qua mental).

To recapitulate, it is the very same token-event,  $b$ , that *is*  $F$  and that *is*  $P^*_1$ . Having  $P^*_1$  is sufficient (although not necessary) for having also  $F$ , but as it happens *in this case* to have  $F$  (to be a certain action) nonetheless simply is to have  $P^*_1$  (to be a combination of certain physical movements) – unless we accept the Davidsonian view of holistic constructivism. Hence, in the present context  $m$  causing  $b$  is sufficient for  $m$  causing  $b$  qua whatever predicate applies to  $b$ . And in the non-Davidsonian picture  $b$  is what it is, regardless of what predicates apply or have been explicitly applied to  $b$ 's causes. I conclude that we have a counterexample to Tuomela's thesis  $T^*$ . The situation I have described is such that " $Q(m, M, b, F)$ " and " $Q(m, P, b, F)$ ", where  $P = P_1$ , are both true.

One could perhaps try to avoid my conclusion by insisting that the nature of the causal history of the action we want to explain ( $b$  qua  $F$ ) matters to the acceptability of the explanation. After all, Davidson's climber example (III.1.3) showed that for a physical behaviour to be an action (for  $b$  to really be  $F$ ), it must have been caused by the reasons *in the right way*. However, this line of defence cannot help a causalist, since it will only get us back to the arguments of III.1.3 where I showed that a causalist cannot accommodate this line of thought. If one takes this route, she ought to admit that one of Tuomela's core claims concerning the relevance of the mental, namely that if  $m$  had not

been  $M$ , then  $b$  would not have been  $F$  (Tuomela 1998, 6) is based on a logical, conceptual or normative connection between  $M$  and  $F$ , just as von Wright explained, Tuomela (1976) acknowledged and the post-1963 Davidson was forced to admit.

In fact, the return to the Logical Connection Argument is of course the best way to make sense of Tuomela's strange-sounding suggestion that we could imagine a counterfactual situation in which  $m$  (as an instance of  $P_1$ ) is not  $M$ , even if all the other instances of  $P_1$  are also instances of  $M$ . If reasons, propositional attitudes and mental states are seen as naturalistic phenomena belonging to the causal order of the world, then the Multiple Realisability Argument and token physicalism, despite their antireductionist aspirations, make it nomologically impossible to imagine that the instances of  $P_1$  would not be also instances of  $M$ , for to be an instance of  $M$  is nothing but to be an instance of  $P_1$  or  $P_2$  or ... or  $P_n$ , as Figure 2 above shows. Therefore, statements such as that if  $m$  had not been  $M$  then  $b$  would not have been  $F$  make more sense in the anti-causalist normative framework of the Logical Connection Argument where  $b$ 's being  $F$  is not a brute fact of nature but rather  $F$  is a normative status partly constituted by norms of rationality (and similarly for  $m$ 's being  $M$ ).

In such a case it would be up to us to refuse to assign the *status* of  $F$  to  $b$  if we do not assign also the status of  $M$  to  $m$ . But then the connection between  $M$  and  $F$  is clearly not a causal but conceptual or normative relation. Indeed, von Wright's (1971) verificationist argument (III.1.2) for the acceptability of the Logical Connection Argument was precisely that whether or not a certain behaviour ( $b$  qua  $P^*_j$ ) counts as an action (whether it is  $F$ ) does not depend solely on  $b$ , but on whether or not  $m$  was  $M$ . As we saw Stoutland (1976) emphasising, to see  $b$  qua  $F$  is not to point out a naturalistic property of  $b$ , but rather to embed it, via an intentional redescription, to a normative, conceptual framework.

### III.2.3 KIM ON EXPLANATORY EXCLUSION

Thus, Tuomela's 1998 attempt to return to the Standard View via a defence of irreducible, *sui generis* mental causation in terms of the Multiple Realisability Argument fails. The causal closure of the physical world leaves no room for strong mental causation.<sup>131</sup> On the contrary, the argument actually elucidates further the derived or secondary sense

<sup>131</sup> In fact, towards the end of his article Tuomela (1998, 28-29) appears to acknowledge this. He suggests that to make room for strong mental causation we may have to reject the closure principle and straightforward mind-brain supervenience. The result would be something very close to Dupré's view (III.2.5). But then Tuomela (1998) would not give an *argument for* mental causation anymore, but an examination of what the acceptance of mental causation presupposes. Dupré is very explicit about this.



in which intentional explanations can be thought of as causal explanations. Instead of strong mental causation that Tuomela (1998), Elster (1985) and most contemporary philosophers of mind who accept the Standard View are after, perhaps we could defend a much more modest view, which Kim (1984) calls epiphenomenal supervenient causation. To spell out this line of thought, let us return to Fodor's version of the Multiple Realisability Argument.

In the Fodorian Figure 2 the law  $M \rightarrow F$  is co-extensional with the following conditional:  $(P_1 \text{ or } P_2 \text{ or } \dots \text{ or } P_n) \rightarrow (P^*_1 \text{ or } P^*_2 \text{ or } \dots \text{ or } P^*_m)$ . Why does Fodor think that the mental law is not reducible to this implication? After all, if the mental law is true, then the conditional is true, too. Fodor's (1974, 695-696) point seems to be that even if the conditional is true, from the point of view of physical science it is an accidentally true generalisation rather than a law. Neither  $(P_1 \text{ or } P_2 \text{ or } \dots \text{ or } P_n)$  nor  $(P^*_1 \text{ or } P^*_2 \text{ or } \dots \text{ or } P^*_m)$  expresses a type recognisable to physical science. Perhaps the only salient thing  $P_i$  (and similarly for  $P^*_j$ ) have in common is the fact that they are realisers of  $M$  (and  $F$ ), and hence grouping  $P_i$  (and  $P^*_j$ ) together is motivated solely from the perspective of psychological theory. Physical theory will not tell us whether or not a given physical type  $P_k$  should be included in the conditional. The only criterion is whether or not  $P_k$  realises the relevant mental property.

If this is the case, the conditional  $(P_1 \text{ or } P_2 \text{ or } \dots \text{ or } P_n) \rightarrow (P^*_1 \text{ or } P^*_2 \text{ or } \dots \text{ or } P^*_m)$  can be formed only on the basis of the mental law  $M \rightarrow F$ , especially since the number of possible realisers may well be infinite. Hence the mental law is conceptually prior to the physical implication. However, this would again imply a move towards the Logical Connection Argument: the intrinsic causal properties of  $P_k$  would not be sufficient for it to count as mental, for that *status* involves a conceptual element that depends on the rational assertability conditions of our mental concepts.

However, Fodor's arguments do seem to justify his main conclusion, namely that the Multiple Realisability Argument blocks the possibility of a Nagel-reduction of mental explanations. Type-identity is replaced with token-identity and the mental supervenes upon the physical. Now the question is whether or not this is really sufficient for mental causation? I have argued that it is not. After all, the very reason why we can hold that  $M$  is causally connected to  $F$  is precisely that each and every instance of  $F$  following  $M$  is actually an instance of some process of normal physical microcausation (Kim 1984).

But this, in turn, means, as Kim puts it, that if we say that mental causation is real, it is nonetheless real only in an *epiphenomenal* sense and not in the strong *sui*

*generis* sense Tuomela (1998) and other advocates of the Standard View require: “Mental causation does take place; it is only that it is epiphenomenal causation, that is, a causal relation that is reducible to [token by token, not in the sense of full-scale Nagel-reduction], or explainable by, the causal processes taking place at a more basic physical level” (Kim 1984, 107). According to Kim, the combination of the causal closure of the physical domain and the non-existence of massive causal overdetermination renders mental causation essentially epiphenomenal. Under these assumptions, talk of mental causation is nothing but a way to talk about normal physical causation.

The explanations in terms of mental causation cannot amount to anything more substantial, since in fact monism with causal closure commits us to the following thesis, which Kim labels the thesis of *explanatory exclusion*: “No event [token] can be given more than one *complete* and *independent* [causal] explanation” (Kim 1989a, 239). Let us look at the following two explanations, (*N*) and (*R*) (Kim 1989a, 240).

- (*N*) Whenever an organism of structure *S* is in neurophysiological state *q* it will emit movement *m*. Organism *O* of structure *S* was in neurophysiological state *q*.  
Therefore, *O* emitted *m*.
- (*R*) Whenever an organism has goal *G* and believes that behavior *B* is required to bring about *G*, *O* will emit *B*. *O* had *G* and believed *B* was required for *G*.  
Therefore, *O* emitted *B*.

The thesis of explanatory exclusion applies to (*N*) and (*R*) only if (*N*) and (*R*) are seen to share the same *explanandum*. As we saw in III.1.3, von Wright’s Logical Connection Argument and Davidson’s post-1963 view are committed to the view that they do not; (*N*) is a causal non-intentional explanation of behaviour and (*R*) a non-causal intentional explanation of action. Hence, if the non-causalists are correct, Kim’s thesis of explanatory exclusion does not apply to (*N*) and (*R*); they can both be accepted as genuine, independent explanations (cf. Stoutland 2005, 141).

The Standard View, including Fodor’s Multiple Realisability Argument, on the other hand, is committed to the view that (*N*) and (*R*) are both causal explanations, and even if their *explanandum* statements are not equivalent, token-identity theory implies that in every concrete case they nonetheless pick out or describe the same event-token (Kim 1989a, 242). Hence the exclusion thesis applies to (*N*) and (*R*) to the extent that (*R*) is seen as a causal explanation. This, of course, means that if the exclusion thesis is

correct, then (*R*) cannot be a *sui generis* causal explanation.<sup>132</sup> This, in turn, would mean that the Multiple Realisability Argument fails to avoid reductionism (even if it does avoid straightforward Nagel-reduction). In other words, if we see agency, action and intentional explanations as essentially causal issues and if the exclusion thesis holds, then agency, action and intentional explanations are reducible to physics. However, as human beings we are all committed to mental realism in the sense of the irreducibility of notions such as of agency, action and intentionality. Hence, if we are forced to accept Kim's exclusion thesis, it is a very strong argument for the non-causalist position of von Wright (1971) (or indeed for rejecting the assumption of causal closure or overdetermination – see III.2.5).

Kim points out that the exclusion thesis seems to amount to the following claim (*I*) (Kim 1989a, 243):

- (*I*) If *C* is sufficient for a later event *E*, then no event occurring at the same time as *C* and *wholly distinct* from it is necessary for *E*.

Obviously, the Standard View is committed to the falsity of (*I*). Again, we should be careful not to confuse types with tokens here. For it might seem that if *C* is a physiological type such that its instances bring about instances of a behavioural type *E*, then the supervenience thesis implies that whenever an instance of *C* occurs, also a mental event of the type *C\** occurs being thus necessary for *E*. And since we have rejected the type-identity theory, *C* and *C\** are not identical. Thus this would seem like a counterexample to (*I*). However, if one accepts the token-identity theory, it becomes clear that in all cases where a *C*-event occurs, there is no *C\**-event *wholly distinct* from the *C*-event, for in fact the *C\**-event, by definition, *is* the *C*-event. In other words, if the mental type *C\** is realised in virtue of a physical realisation base *C*, the causal powers of this instance of *C\** must be identical with the causal powers of *C* (Kim 1992, 326), for this particular instance of *C\** *is* *C*. Kim calls this the Causal Inheritance Principle, and points out that the denial of the principle “would be to accept emergent causal powers: causal powers that magically emerge at a higher level and of which there is no accounting in terms of lower-level properties and their causal powers and nomic connections”

---

<sup>132</sup> It is absolutely crucial to understand that the Standard View is committed to seeing (*N*) and (*R*) as independent of each other (reasons causing actions irreducibly qua reasons, not qua physical states). If they are not independent in this strong sense of *sui generis* mental causation, then (*N*) and (*R*) can of course coexist even if the exclusion thesis holds; (*N*) and (*R*) can coexist as causal explanations if, say, the explanatory efficacy of (*R*) derives from the explanatory efficacy of (*N*). (Kim 1989a, 241.)

(Kim 1992, 326, see also 1998, 55-56). Needless to say, this would clearly violate the principle of the causal closure of the physical domain.

Kim's (1989a, 246) conclusion is that attempts to challenge (*I*) (and hence the attempts to challenge the exclusion thesis) fail because the two explanations (*N*) and (*R*) are not independent of each other. If *C* and *C\** are nomologically equivalent to each other, what we have is a reductionist type-identity theory subject to full-blown Nagel-reduction – and hence there really is only one explanation. If we accept the Multiple Realisability Thesis and say that *C\** supervenes on *C* (a mere token-identity), explanations in terms of *C\** gain their explanatory or causal efficacy from the connection between *C* and *E*, and hence the explanations are not really distinct from one another. Avoiding the traditional type-identity theory and Nagel-reductionism is not sufficient for avoiding the kind of token-by-token reductionism that is just as disastrous for mental causation and the Standard View of seeing reasons as *sui generis* causes.

Thus, to the extent that neurophysiological and intentional explanations of *E* are both seen as causal explanations (and we are operating under the assumption of causal closure), I think Kim is right in saying that this is what we must conclude:

The two explanations differ only in the linguistic apparatus used in referring to, or picking out, the conditions and events that do the explaining; they are only descriptive variants of one another. They perhaps give causal information about *E* in different ways, each appropriate in a particular explanatory context; but they both point to one objective causal connection, and are grounded in this single causal fact.

(Kim 1989a, 248.)

This is of course nothing like Tuomela's 1998 aspiration of irreducible, *sui generis* mental causation or the Standard View of reasons causing actions qua reasons, because the role left for mental predicates is merely that of practical, instrumental aids in making predictions. Hence, for a causalist eliminativism is a real option: mental states and mental causation are not *sui generis* features of the world – whether or not we want to keep talking about them depends solely on the instrumental value of mentalist language.

Kim offers a helpful analogy. Consider a set of minerals. Here each base-property *P<sub>i</sub>* can be seen as the property of being mineral *X<sub>i</sub>*, for every mineral kind *X<sub>i</sub>*. Now we can introduce a predicate *M*, "being jade", which "can be thought of as the second-order property of being a mineral that is pale green or white in color and fit for use as gemstones or for carving. This second-order property has two known realizers, jadeite and nephrite." (Kim 1998, 20.) Let us call the first-order property of being jadeite *P<sub>1</sub>* and being nephrite *P<sub>2</sub>*. In such a situation two of the *P<sub>i</sub>*'s, namely *P<sub>1</sub>* and *P<sub>2</sub>*, satisfy the

conditions for  $M$ . Hence, all the instances of  $P_1$  (jadeite) are also instances of  $M$  (jade), but not *vice versa*, for some instances of  $M$  are instances of  $P_2$  (nephrite). The second-order property  $M$  has multiple realisations, and hence it is not Nagel-reducible to the first-order properties  $P_i$ , exactly as Figure 2 says.

It seems obvious that the introduction of the predicate  $M$  does not assign  $P_1$  and  $P_2$  any independent causal powers they would not have qua  $P_1$  and  $P_2$ . However, as was explained in Part I,  $P_1$  and  $P_2$  can well be assigned a new social status  $M$ , and the assignment of this status may give  $P_1$  and  $P_2$  a new institutional (normative) role  $P_1$  and  $P_2$  do not possess independently of the practices of maintaining that status. In the case of minerals this is indeed clear. Assume a collective accepts that henceforth instances of  $M$  (jade) will count as money. Surely pieces of  $P_1$  (jadeite) and  $P_2$  (nephrite) cannot receive new *causal* powers as a result of such a collective performative speech act, for it would indeed be magical if our speech acts had such powers. At the same time, however, it seems clear that the social and normative role of pieces of  $P_1$  and  $P_2$  in the collective in question has changed dramatically. New *normative* or *institutional* powers (or, to use Searle's (1995, 100) terminology, new *deontic* powers) have been created. I think the situation is essentially the same if instead of minerals and money we think of physical states and propositional contents. Contents assigned to physical states can enter in normative relations, as the Logical Connection Argument holds, but such an assignment does not bring about new causal powers. How exactly this might work is explained in Chapter III.3 onwards.

Obviously, this line of thought is not available for a causalist. For her,  $M$  is interesting only in virtue of its causal powers. And in the name of naturalism no instance of  $M$  can have causal powers other than the powers of its physical realisation, since ontologically speaking the instances of  $M$  are nothing but instances of  $P_1$  and  $P_2$ . In fact, this is the only way of understanding supervenience naturalistically:

Why is it that whenever  $P$  is realized in a system  $s$ , it instantiates mental property  $M$ ? The answer is that by definition, having  $M$  is having a property with causal specification  $D$ , and in systems like  $s$ ,  $P$  is the property (or one of the properties) meeting specification  $D$ . For systems like  $s$ , then, having  $M$  consists in having  $P$ . It isn't that when certain systems instantiate  $P$ , mental property  $M$  magically emerges [...]. It is rather that having  $M$  for these systems, simply is having  $P$ . We might even say, using a familiar if shopworn reductive idiom, that having  $M$ , for these systems, is "nothing over and above" having  $P$ .  
(Kim 1998, 24.)

Thus, contrary to the Standard View of thinking of mental causation as a *sui generis* form of causation, talk about mental causation is simply *functionalist* talk about normal

physical causal relations. And as argued in Part I, functions are always observer-relative normative notions, not ontologically independent facts of the world.

The mistake repeated again and again in contemporary philosophy of mind is to think that since the Multiple Realisability Argument denies the existence of one-to-one bridge laws between the physical and the mental, preventing thus the possibility of a traditional Nagel-reduction, the argument is strong enough to secure a place for *sui generis* strong mental causation. But the question of mental causation is a metaphysical problem, and hence the relevant notion of reduction must be that of a *metaphysical* token-by-token reduction (token physicalism), not *methodological* Nagel-reduction that is ultimately about intertranslatability of languages. Again, Kim has got it right: “the reducibility of a property critically depends on its *functionalizability* – whether or not it can be construed as a second-order functional property over properties in the base domain – not on the availability of bridge laws. Bridge laws are neither necessary nor sufficient for reduction.” (Kim 1998, 27.)

The failure of the Multiple Realisability Argument to defend the causalist position can be summarised with the following argument (adapted from Kim 1998, 39-46). Let us start with a truism:

- (1) Either the mental supervenes on the physical or it does not.

If the mental does not supervene on the physical, then, if the view is still ontologically monistic, the mental is either nomologically connected to the physical, in which case (Nagel-) reductionism follows, or the mental is completely anomalous, in which case it is causally impotent. In both cases mental causation is an illusion. As I explained above, this dilemma was precisely the reason why, *e.g.*, Fodor introduced his Multiple Realisability Argument, according to which the mental is neither connected to the physical via biconditionals nor is it completely unconnected to it. Instead, the mental is said to supervene on the physical. Supervenience may of course fail also if ontological dualism is accepted, but then the mental cannot have physical effects without violating the causal closure of the physical domain.

Thus:

- (2) If supervenience fails, there is no way of understanding the possibility of mental causation (under the causal closure assumption – see III.2.5).

Let us suppose that there is mental causation:

- (3) Suppose that an instance of mental property  $M$  causes an instance of another mental property  $M^*$ .

From (2) it follows directly that

- (4)  $M^*$  has a supervenience base  $P^*$ .

Let us now look at the following assertion:

- (5)  $M^*$  is instantiated on this occasion because either (a)  $M$  caused  $M^*$  to be instantiated or (b) because  $P^*$  is instantiated on this occasion.

Under the assumption (3), (a) is true. However, under the supervenience thesis (4), (b) is true. Hence it seems that the only way for (a) to indeed be true is that  $M$  caused  $M^*$  via  $P^*$ , for  $P^*$  is what instantiates  $M^*$  on this occasion. Hence:

- (6) On this occasion,  $M$  caused  $M^*$  by causing  $P^*$ .

This means that under the supervenience assumption mental causes can have mental effects only in virtue of the mental cause causing a physical effect. But from (2) it also follows that:

- (7)  $M$  has a supervenience base  $P$ .

But now it seems that both  $P$  and  $M$  are causally sufficient for  $P^*$ , which would violate the dual thesis of the causal closure of the physical domain and the non-existence of massive causal overdetermination, *i.e.*, the thesis of explanatory exclusion. To save the assumption (3), it appears that we must accept that:

- (8)  $P$  caused  $P^*$ ,  $M$  supervenes on  $P$  and  $M^*$  supervenes on  $P^*$ .

Now, as Kim points out, (8) “explains the observed regularities between  $M$ -instances and  $M^*$ -instances, and those between  $M$ -instances and  $P^*$ -instances. These regularities

are by no means accidental; in a sense they are law-based, and may even be able to support appropriate counterfactuals.” (Kim 1998, 45.)

Has Kim found a respectable way to defend mental causation and the Standard View? The answer, as Kim is well aware of, must be in the negative, because where the connections between the supervenience bases  $P$  and  $P^*$  are “genuine, productive and generative causal processes” (Kim 1998, 45), the observed regularities featuring mental predicates are mere “noncausal regularities that are observed because they are parasitic on real causal processes” (Kim 1998, 45). Thus we are forced to accept that:

- (9) The  $M$ -to- $M^*$  and  $M$ -to- $P^*$  relations are not genuine causal relations at all; they are regularities parasitic upon an authentic causal process from  $P$  to  $P^*$ .

Hence Kim’s conclusion:

- (10) “If mental-physical supervenience fails, mental causation is unintelligible; if it holds, mental causation is again unintelligible. Hence mental causation is unintelligible.” (Kim 1998, 46).

Introducing the Multiple Realisability Argument and mental-physical supervenience has not, contrary to the received wisdom in contemporary philosophy of mind and action theory, helped us to save the Standard View of seeing reasons as causes.

### III.2.4 CONCLUSION

The Multiple Realisability Argument has left us in exactly the same situation we found ourselves at the end of Chapter III.1: Intentional explanations can be causal explanations only in the parasitic sense that where they work they do so because of the genuine, physical causal processes allow also such talk to succeed. The psychological laws are not causal laws in their own right, and hence it seems that when Fodor (1974) emphasises that the Multiple Realisability Argument shows that we cannot translate our mental talk into the language of physics, this is not, *pace* Fodor, an argument for the special sciences operating with higher-order laws such as mental laws. Rather, this is exactly what eliminativists such as Quine or the Churchlands want to hear: if the language of psychology prohibits us from talking about the genuine causal processes out there in the world, we had better eliminate this misleading language and replace it with a physicalist



language suitable for capturing the genuine causal laws (those between  $P_i$  and  $P^*_j$  in the Fodorian Figure 2).

Even if we wish to resist eliminativism, causalists ought to admit that the much-celebrated Multiple Realisability Argument has not taken us much further from John Stuart Mill's view of psychology. Mill held that the laws of psychology are Empirical Laws, *i.e.*, inductive generalisations of observed connections (such as that  $F$  is constantly observed to follow  $M$ ). An Empirical Law is simply a generalisation for which we have not observed counterexamples. The real causes supporting the generalisation, on the other hand, may be very different from the observed generalisation. Thus, whether the generalisation will hold also in unobserved cases depends on the real causal laws that maintain the observed connection.

An Empirical Law [...] is an uniformity, whether of succession or of co-existence, which holds true in all instances within our limits of observation, but is not of a nature to afford any assurance that it would hold beyond those limits, either because the consequent is not really the effect of the antecedent, but forms part along with it of a chain of effects, flowing from prior causes not yet ascertained, or because there is ground to believe that the sequence (though a case of causation) is resolvable into simpler sequences, and, depending therefore on a concurrence of several natural agencies, is exposed to an unknown multitude of possibilities of counteraction. In other words, an empirical law is a generalisation, of which, not content with finding it true, we are obliged to ask why is it true? knowing that its truth is not absolute, but dependent on some more general conditions, and that it can only be relied on in so far as there is ground of assurance that those conditions are realised.

(Mill 1865, Book 6, Chapter 5, §1.)

Thus, if we do not have access to the real causes behind the Empirical Laws of psychology, we might find it convenient to use Empirical Laws for practical purposes. But we should not delude ourselves into thinking that the Empirical Laws of psychology are *sui generis* causal laws. Rather, as I have argued above, Empirical Laws and mental properties inherit whatever causal explanatory efficacy they may have from the real causal properties – and the laws covering them.

Unlike many contemporary philosophers, Mill was able to see very clearly that the Empirical Laws of psychology must not be confused with true causal laws of science.<sup>133</sup>

---

<sup>133</sup> Admittedly, for Mill the true causal laws were also psychological laws (*e.g.*, maximisation of expected pleasure and so on), and hence my appeal to Mill may be somewhat misleading. But for the contemporary naturalists I have discussed it is clear that the basic causal laws are physical laws, and thus what we ought to learn from Mill is that the (Nagel-) irreducibility of Empirical Laws is not an argument for thinking that they are *sui generis* causal laws.

The empirical law derives whatever truth it has from the causal laws of which it is a consequence. If we know those laws, we know what are the limits to the derivative law; while, if we have not yet accounted for the empirical law – if it rests only on observation – there is no safety in applying it far beyond the limits of time, place, and circumstance in which the observations were made.  
(Mill 1865, Book 6, Chapter 5, §1.)

Moreover, Mill makes it clear that empirical laws, as opposed to true causal laws of science, may well have exceptions.

This agrees very well with Fodor's picture, for also Fodor (1974, 696) considers one of the main virtues of his token physicalism to be that it explains how the laws of the special sciences can have exceptions. Fodor (1974, 696) argues that it is conceivable that there may be some properties  $P_i$  in the supervenience base of  $M$  that do not bring about an effect in the supervenience base of  $F$ , even if most realisers of  $M$  do have an effect that realises  $F$ . This is also one of the main reasons why Fodor thinks that the conditional  $(P_1 \text{ or } P_2 \text{ or } \dots \text{ or } P_n) \rightarrow (P^*_1 \text{ or } P^*_2 \text{ or } \dots \text{ or } P^*_m)$  cannot be regarded as a physical law (to which the psychological connection  $M \rightarrow F$  would be reducible): the implication does not in fact hold for all  $P_i$  and hence the conditional is not a law, for true laws are supposed to give sufficient conditions for the effect. Where Mill regards this as evidence that Empirical Laws are not proper causal laws of science (and people such as Quine or the Churchlands as a strong argument in favour of eliminativism), Fodor thinks that his argumentation portrays the laws of the special sciences as proper causal laws – and that the fact that his theory allows the laws of the special sciences to have exceptions simply makes his view more realistic.

I think that in fact this is simply more bad news for the causalist view on agency and action. If causation is what matters, then surely we ought to prefer the neurophysiological framework (the laws  $P_i \rightarrow P^*_j$ ) to psychology, since even in the cases where we observe  $M$ , psychological theory is not going to tell us whether this instance of  $M$  is going to have an instance of  $F$  as its effect or not. Neurophysiology, on the other hand, operates (Fodor assumes, cf. III.2.5) ultimately with exceptionless causal laws. A special science that is content with Mill's Empirical Laws, such as Fodor's causalist psychology, is in Fodor's own terms exactly like the "science" of Plato's prisoners in their cave (the *Republic*, Book VII), who may be able to construct a wonderfully sophisticated system of empirical generalisations concerning the sequences of shadows cast on the wall of the cave. Their generalisations may even support some counterfactuals, and the observed regularities are not random or accidental. This is why we can *use* Empiri-

cal Laws as *indicators* of causal processes in our explanations even if as such they do not capture fundamental causal processes (cf. Pettit 1996).<sup>134</sup>

Nevertheless, to the extent that we are interested in detecting genuine causal processes, studying shadows is not enough. There is no (direct) causal connection between a shadow at one moment and a shadow at one instant later, for both are effects of the objects casting the shadows. Whether one calls descriptions of shadow-sequences laws of the special sciences (Fodor), mental causation (Tuomela 1998), epiphenomenal supervenient causation (Kim), programme explanations (Jackson and Pettit) or Empirical Laws (Mill), a serious causal science should not be about them, because “[t]he really scientific truths [...] are not these empirical laws, but the causal laws which explain them” (Mill 1865, Book 6, Chapter 5, §1).

To conclude this discussion we could say that a core problem of the contemporary philosophy of mind is that it is very difficult to see how to defend an acceptable causalist view of action coherently. The philosophers discussed here want to subscribe to physicalism in the sense of accepting the principles of causal closure and token-identity. However, as agents we are as strongly committed to mental realism: mentality – intentionality, contentful mental states and meaningful actions – is a real feature of the world. According to Kim, however, the failure of the Multiple Realisability Argument, together with the fact that Davidson’s anomalous monism cannot portray reasons (so described) as causes, shows that these two fundamental principles, *i.e.*, physicalism and mental realism, are mutually incompatible. “So if I am right, the choices we face concerning the mind-body problem are rather stark: there are three – antiphysical dualism, reductionism, and eliminativism” (Kim 1989b, 267). Anti-physical dualism is not an option, but the alternatives seem to render mental realism impossible.<sup>135</sup>

However, we are left with Kim’s three intolerable options only if we accept his two fundamental premises: (i) physicalism in the sense of the combination of causal closure and token-identity and (ii) the conviction that mental realism is a causal affair. In

---

<sup>134</sup> Pettit, however, builds on his (and Frank Jackson’s) model of programme explanations (see the essays in Jackson, Pettit and Smith 2004) that seeks to secure a more fundamental role for Empirical Laws: The presence of a mental state featuring in an Empirical Law is said to *non-causally programme* it to be the case that a causal process instantiating the Empirical Law will be present. However, “non-causal programming” is either an utterly unnatural and incomprehensible notion or a purely instrumentalist or pragmatic notion in Mill’s sense (for also Pettit accepts the causal closure principle, cf. III.2.5).

<sup>135</sup> Hence I think the persistent claims, according to which – despite their explicit denials – thinkers such as Dennett (*e.g.*, 1991b) who accept the purely instrumental reading of mental causation and intentional explanations are really eliminativists, are by no means unfounded. Dennett’s case is particularly interesting, for in what follows I argue that Dennett’s (2003) recent views might be interpreted such that in his view intentional explanations are purely instrumental only qua causal explanations (opening the door for reductionism/eliminativism), but we can also treat them as non-causal rationalising explanations, in which case mental states would matter qua contentful states – and mental realism would be rescued. This would be a view not very different from the one defended in this dissertation.

In III.3 onwards I wish to argue that the causalist reading of mental realism is unmotivated: we can and should accept the anti-causalist, normative interpretation of mental realism of G. H. von Wright's (and, perhaps, post-1963 Davidson's) Logical Connection Argument. However, before I seek to undermine the motivation for causalism, I want to say something about the other possibility of resisting Kim's conclusion, that of relaxing one's notion of physicalism.

### III.2.5 POSTSCRIPT: THE CAUSALIST'S LAST HOPE

Contrary to the received wisdom in contemporary philosophy of mind, the Multiple Realisability Argument fails to make causal mental realism compatible with physicalism that accepts the principles of the causal closure of the physical domain and token-identity instead of type-identity. However, as such this does not imply the failure of mental causation. Rather, the argumentation shows that one cannot coherently accept the following five theses (adapted from Crane 1995, 229) that the advocates of the Standard View typically want to accept:

- (A) Causes have their effects in virtue of (some of) their features or properties.
- (B) There is *sui generis* mental causation.
- (C) The physical domain is causally closed.
- (D) There is no massive causal overdetermination.
- (E) Mental and physical causation are of the same kind.

I have argued that the Multiple Realisability Argument cannot deliver (B) if it sticks to the other theses. Tuomela (1998), for example, opens his paper by claiming to defend (B) while keeping the other theses, but towards the end of his paper he suggests that perhaps keeping (B) requires us to abandon (C). This, however, is not a *defence of mental causation* anymore, but an explication of what mental causation requires.

Davidson's defence of (B), discussed in III.1, is essentially based on a nominalist rejection of (A). In Davidson's view it does not make sense to say that causes have their effects in virtue of their physical or mental features. Rather, we must say that a cause is a concrete particular that causes its effects qua the particular it is, regardless of how we happen to pick it out. I argued that this does not suffice for the Standard View, for in Davidson's view the mental *exists* only relative to the essentially normative and non-causal mentalist language game that *constitutes* the mental qua mental, and thus

psychology cannot be a causal science (Davidson 1974a). In other words, in the context of the Standard View one *cannot* have (*B*) without (*A*): As Elster (1985) shows, for the Standard View the mental must be causally efficacious in virtue of one of its features, *i.e.*, its *content*, which in the context of Davidson's holism exists only relative to the non-causal language game. Thus Davidson cannot deliver (*B*) in the sense required by the Standard View.

Further, in III.2.4 I briefly mentioned (especially Footnote 134) how Jackson and Pettit's (see Jackson, Pettit & Smith 2004) theory of programme explanations rejects (*B*), but when this is seen to threaten mental realism they revert to saying that mental states can nonetheless non-causally "programme" things to happen. Without an explanation of what this programming is, this view is but an unintelligible attempt to rescue (*B*) by rejecting (*E*) without violating (*C*).

Hence, we must reject either (*B*), (*C*) or (*D*). Notwithstanding some exceptions (*e.g.*, von Wright, Stoutland and the present dissertation), the consensus appears to be that (*B*) "is the last assumption we should reject" (Crane 1995, 230). However, no-one seems to be willing to reject (*D*) either. Kim (1989b, 281), for instance, calls the rejection of (*D*) absurd. To hold that in cases of intentional agency there always *happens* to be two mutually independent sufficient causes presupposes such a gigantic coincidence that we simply cannot consider rejecting (*D*) (recall Footnote 128).

Thus, rejecting (*C*), the principle of the causal closure of the physical domain, is the causalist's last hope. This, however, is something most philosophers do not even consider. In their view contemporary science simply compels us to keep (*C*) (*e.g.*, Papineau 2002). But then we cannot have (*B*), *sui generis* mental causation, and, therefore, the Standard View.

However, some philosophers (*e.g.*, Cartwright 1999, 32-33 and Dupré 1993, 184; 2001, Chapter 7) think that the acceptance of (*C*) is simply an uncritical dogma. John Dupré defends this view explicitly in the context of the problem of mental causation, and thus in what follows I concentrate on his version of the idea.

Interestingly, the spirit of Dupré's discussion is very similar to the animating themes of this study. For example, my criticism of mainstream rational choice approach in Part II (which I see as closely connected to my views on intentional explanation and mental causation – see III.5.1 onwards) is very close to, and indeed partly motivated by, Dupré's (2001) arguments. Similarly, I argue (III.5.3) that the non-causal view of agency and action I defend allows me to reinforce Dupré's (2001) criticism of evolutionary psychology. Further, a major motivation for Dupré's rejection of (*C*) seems to

be (e.g., 1993, 90 ff.) his willingness to resist naïve fetishisations of fundamental physics, *i.e.*, the view that one day physics could replace the human sciences. Exactly the same motivation underlies my rejection of causal theories of action and agency.

Moreover, when it comes to explanations of actions, Dupré (1993, 152) underlines that more often than not we are interested in *what* someone is doing rather than *how*, exactly, the action takes place. Explanations in terms of fundamental physics cannot answer such what-questions, essentially because actions “take place in social contexts that have much to do with determining what kinds of actions they are” (Dupré 1993, 153). The (largely sociological) question of *what one does* cannot be answered in terms of a physical account of *how*, exactly, the physical movement that instantiates the action happens. This could serve as a summary of my non-causal view too. Similarly, Dupré’s denial of (mind-brain) supervenience is based on seeing the content of mental states as determined largely by “contextual social factors that go beyond the internal structure of the individual” (Dupré 1993, 157). Dupré even hints (1993, 156; 2001, 35) that the reason for this is Wittgenstein’s rule-following argument. Once again, this is *exactly* the line I take in the following Chapters. However, where I think this view goes hand in hand with my non-causalist, normative view of action and agency, Dupré is a firm causalist.

Crucially, however, Dupré agrees with Crane, Kim and the present study that supervenience as such cannot suffice for the Standard View (Dupré 1993, 96-97 & 101). But where I reject (B) and the Standard View, Dupré rejects (C) (concerning which I remain neutral). Dupré’s (1993) main arguments against (C) are based on his observation that – *pace*, e.g., Papineau (2002) – far from supporting the principle, our best science in fact speaks against it. In particular, determinism, which is the natural fellow traveller of the principle, “seems almost entirely, or perhaps entirely, devoid of empirical support” and, accordingly, “our most successful scientific theories describe a probabilistic rather than a deterministic world” (Dupré 1993, 184). The idea is that the exclusion thesis (III.2.3; recall especially how the exclusionist principle (I) and the step from (7) to (8) built explicitly on the physical causes being *sufficient* for their effects) and the argument from overdetermination (III.2.2.), which lie at the core of my Kim-inspired criticism of (B), are based on a deterministic reading of the principle of causal closure. If indeterminism is true, and antecedents in the laws of fundamental physics are insufficient for their consequences, the *sui generis* mental causation that rejects (C) does not challenge (D).

Although Dupré's view stands or falls with his rejection of (C), assessing the (un)acceptability of determinism and causal closure is beyond the scope of this dissertation. Hence, I adopt a different line of argumentation. First, I wish to point out that Dupré agrees, contrary to the prevailing view, that causal closure renders (B) and the Standard View impossible, regardless of the Multiple Realisability Argument. However, just as most advocates of the supervenience orthodoxy, also Dupré finds the rejection of (B) intolerable. Thus, he wants to turn "the reductionist *modus ponens* (causal completeness [closure] requires reductionism) into [...] [an] antireductionist *modus tollens* (the failure of reductionism implies the falsity of causal completeness)" (Dupré 1993, 102). As most contemporary philosophers, also Dupré is not so much interested in *defending* mental causation as such; rather, his goal is to demonstrate how one can consistently *accept* mental causation.

What I am getting at is that the reason why Dupré and others insist on keeping (B) is, I think, their commitment on mental realism. Crucially, however, my non-causalism is not meant to challenge mental realism, but the further assumption that mental realism is a causal notion. My point is that although Dupré may well be correct in his criticism of determinism and, thereby, causal closure, in any case his causalist mental realism is committed to much stronger presuppositions than my non-causalism. In particular, Dupré's defence of mental realism *presupposes* indeterminism and the failure of (C).<sup>136</sup> At least the latter presupposition remains unacceptable to the majority of contemporary philosophers of science and mind. My non-causal mental realism, in contrast, is compatible with both views: The normative view is a compatibilist notion of agency that is not tied to the fate of indeterminism and the principle of causal closure.

All this is, of course, still merely a negative argument in favour of non-causalism: I have argued that the causalist view has implications many will find unattractive. However, perhaps there are causalist theories that avoid these implications, although I doubt it (recall also Footnote 124). Be that as it may, instead of criticising more causalist theories I think it is time to finally offer a positive argument in favour of non-causalism.

The non-causal, normative view of mental realism holds that the constitutive principles of the mental realm are essentially different from causal relations regardless of our specific views on causation (as long as we take causal relation to be essentially non-normative – see III.1 and Appendix). Dupré, on the other hand, thinks that mental

---

<sup>136</sup> This, of course, does not worry Dupré, since in his view we have further, independent arguments for these presuppositions (see Part III of Dupré 1993).

states are essentially causal, as his *modus tollens* testifies. Thus, although I have not refuted Dupré's view, if I can show that mental states exist only relative to normative (constitutive) rules essentially different from causal relations, the motivation to accept causalism even in Dupré's sense is seriously compromised. The talk of the mental would belong to the logical space of normative reasons, not to the space of causation (the rejection of (B)). This is the task undertaken in the rest of this dissertation.



## CHAPTER III.3:

WHAT IS INTENTIONALITY ANYWAY? THE PROBLEM OF RULE-FOLLOWING<sup>137</sup>

## III.3.1 INTRODUCTION

The challenge undertaken in this Part of my dissertation is to show that the intentionalist framework adopted in Parts I and II is justified. Intentionality and intentional action must form *sui generis*, irreducible features of the world. After Chapters III.1 and III.2 above, the jury is still out. III.2 showed that the celebrated attempts to defend the framework of intentional agency on the basis of the Multiple Realisability Argument fail in their own terms. The failure stems from the steadfast commitment to the principle of the causal closure of the physical world (e.g., Papineau 2002) and to deterministic causation. Rejecting these widely-held commitments appears to offer the most promising defence of the causalist interpretation of the intentionalist framework (III.2.5).

Another way of defending the intentionalist framework was set up in III.1, where I showed that contrary to the view dominating contemporary action theory (e.g., Mele 2003, 82), Davidson's alleged refutation of the Logical Connection Argument fails to refute the non-causalist view of von Wright and others. If von Wright is right, reasons and actions belong to what Wilfrid Sellars (1963) called the logical space of normative reasons, not to the logical space of non-normative causal processes. This move rescues the intentionalist programme, since it is only contentful states and intentional actions (in virtue of their propositional contents) that can stand in normative, conceptual relations. And they do it precisely qua contentful, qua intentional. Thus, non-causalism is another viable option for a defence of the intentionalist framework. Which should we choose?

Thus far the argumentation of this Part has been almost exclusively negative in nature. I have shown that the non-causalist view survives Davidson's alleged refutation. Similarly, I have argued that Dupré's metaphysical pluralism is the only hope a causalist has left. But I have not explained how, exactly, either programme works. I have not argued *for* causalism or non-causalism. This is what I want to do next. More precisely, I want to examine the nature of intentional phenomena (such as meaningful actions, sentences, thoughts and languages) qua intentional and ask what it is in these phenomena

---

<sup>137</sup> This and the following Chapter build largely on Saaristo (2004b).

that makes them intentional. My goal is to show that explicating the nature of intentional phenomena tips the scales in favour of normative non-causalism – and, ultimately, yields a novel theory of collective intentionality in the sense of Parts I and II.<sup>138</sup>

As part of my defence of the normative view I need also to explain how, exactly, the *sui generis* normative space of reasons is possible in the world that ultimately consists of non-normative, blind physical processes governed by causal laws. Again, it seems that I face a dilemma. Either the required normativity is based on independently-existing Platonist norms, in which case the ontological naturalism I have been advocating is compromised, or it is a socially constructed social fact, in which case – given the intentionalist analysis of social facts given in Part I – the attempt to appeal to normativity as the saviour of intentionality seems blatantly circular.

Thus, I need to give an account of normativity that avoids the dilemma. I argue that our most fundamental social institutions – which, echoing Wittgenstein, I call social bedrock practices – constitute simultaneously both normativity and intentionality, avoiding thus the looming circularity. These are social *bedrock* practices, for they do not presuppose intentionality, but all intentional notions presuppose them (including the practices Part I talks about). I argue that the social practice view allows us to see normativity as an *external* and *objective* feature of the world without assuming anti-naturalist Platonism (recall the intersubjective but ontologically naturalist definitions of externality and objectivity constructed in Part I). In the terminology of Kusch (2006), the normativity I talk about is *intersubjective* normativity, not metaphysical or semantic.

This combination of naturalism and objective normativity, in a sense, presupposes that I perform the ultimate philosophical magic trick and derive “ought” from “is”. Again, I argue that the theory of social practices and the intersubjective but non-independent notion of objectivity defended in Part I comes to the rescue. The normativity I defend will be relative to the social bedrock practices and, hence, *objective for and external to* individuals participating in the practices (but dependent of the practices) and invisible for others. This kind of *bootstrapped* normativity is the strongest kind naturalism can allow (cf. Appendix), and I hope to show that it suffices for the intentionalist programme. In the course of my argumentation I hope to establish the importance of normativity for actions and agency – and by doing so I wish to undermine the motivation behind causalist approaches to intentional agency.

---

<sup>138</sup> However, the non-causalism I defend is not a rejection of causalism *tout court*, for instead of claiming that causal factors have nothing whatsoever to do with intentionality and agency, the goal is rather to specify the precise scope of causalism.

To do all this requires me to say a lot about the ontological nature of normativity, meaning, mental contents and, most importantly, intentionality. The task, enormous as it may be, is nonetheless worth the effort, for its successful completion (i) delivers a naturalistic defence of the intentionalist programme so badly needed, (ii) allows us to sharpen our understanding of social practices and social institutions, (iii) reveals the true nature of human action and agency and, finally, (iv) ties together the main themes of all the three Parts of this dissertation. So, let me roll up my sleeves and get to work. The job is colossal, but still – or precisely for that reason – it is best to start from the foundations. So what is intentionality? How is it possible?

Intentionality in the general sense – as *aboutness*; certain things such as contentful thoughts, meaningful sentences or linguistic signs etc. being *directed at, about* or *of* something else, *i.e.*, having *contents* – is a deeply astonishing phenomenon. However, for something to be intentional it is neither sufficient nor necessary for that something to be causally connected to its object. After all, we can have meaningful thoughts about unicorns as well as abstract objects such as numbers. Hence, intentionality cannot amount to a mere straightforward causal relation between stimulus and reaction (the fundamental flaws of this popular view are explained in detail in III.3.2). Intentionality is directedness that can take different forms. An intentional state can amount to *taking* something as true (beliefs), hoping that something *would be* true (desires), aiming to *make* something true (intentions) and so on. We are capable of entertaining ideas without being committed to their truth (without believing them). Whatever theory of intentionality one is going to advocate, the theory must be able to accommodate these different aspects of intentionality and intentional states.

I may as well lay my cards on the table. I think that in order to take into account the different features of intentionality we must concede that for the theory of intentionality the primary question is not about the nature of the *relation* between thoughts and the world, but about the nature of *content*, *i.e.*, how there can be contentful thoughts in the first place.<sup>139</sup> Intentionality is not made possible by certain kind of connection be-

---

<sup>139</sup> Brandom (especially 2002, 21 ff.) thinks that this kind of emphasis marks a departure from Cartesian philosophy that treats intentionality and content (and representation) as unproblematic and consequently concentrates on *epistemology*, *i.e.*, on the success of our representations. The animating thought of my approach is then rather a *semantic*, Kantian question: I wonder how we can represent anything in the first place, rightly or falsely? In Appendix I follow Brandom and, *e.g.*, Putnam (recall his *semantic* argument against the essentially Cartesian Brains in a Vat scenario in Chapter 1 of Putnam 1981) in suggesting that the semantic problem is not only conceptually prior to Descartes' epistemological problem, but also that a successful solution to the semantic problem allows us to avoid Cartesian problems in epistemology. Solving the semantic problem solves, in a sense, epistemological scepticism for free. This approach is essentially anti-representational; *pace* the Cartesians, representation is not treated as a primitive semantic notion but a highly problematic one (note that my anti-representationalism is nonetheless of Brandom's kind

tween, say, a human brain and physical objects, but in virtue of certain states or things being contentful. It is the intentional, meaningful content that makes it possible for something to be directed at something else. Moreover, the core of the theory of meaning or intentional content is already given in Part I of this study, where it was argued that the facts that certain pieces of paper are (count as) money or that certain marks on a paper are (count as) meaningful sentences are to be analysed in terms of *normative statuses* collectively assigned to the pieces of paper and the marks on the paper.

Part I argued that ontologically speaking such normative statuses boil down to collectively accepted and required patterns of behaviour. Similarly, this Part argues that meaningful (intentional) states such as beliefs and desires are normative statuses that get their content from the patterns of behaviour attached to them in virtue of collective acceptance (*i.e.*, in virtue of a normative connection interdefining and (partly) constituting beliefs, desires, actions etc. – recall III.1). Forming a belief amounts to undertaking a *commitment* to something being the case, and such commitments bring with them certain collectively accepted *rights* and collectively accepted *duties*. Similarly, desires and intentions bring with them different commitments, as do thoughts about unicorns and mathematical objects. It is, for example, collectively required that, *ceteris paribus*, in the presence of cows I undertake the commitment to use the sentence “there are cows present” (from the belief) as a premise in my practical reasoning (leading to actions) and theoretical reasoning (leading to other commitments, such as the commitment that there are mammals present).

This view, which in what follows will be argued for in detail, implies that my account of intentionality subscribes to the Wittgensteinian idea of seeing *meaning as use* in the sense that meaning, content and intentionality are constituted by collectively sustained use of assertions (including appropriate performances of actions). Meanings reside in normative, social practices. However, I cannot simply apply the theory of Part I to intentional states such as beliefs, desires and intentions, since that theory was formulated *within* the framework of intentional agency. In Part I the analysis of social facts in terms of normative statuses built on full-blown intentional attitudes that ground the statuses. The intentionality of marks on a paper etc. was analysed as *derived* from the more fundamental intentionality of human agents, which was simply taken as given. Indeed, (collective) intentionality was identified as one of the fundamental building blocks of social reality.

---

in the sense that I, like Brandom but unlike, *e.g.*, Rorty (1979), think that the notion of representation is extremely important, although not primitive (cf. Appendix).

The present Part, in contrast, focuses on exactly the kind of *original* intentionality that the intentionalist framework of Part I presupposes, and hence to avoid vicious circularity I will need to build on pre-intentional (or sub-personal) building blocks. Thus it should be kept in mind that although in what follows I for sake of simplicity talk mainly about meaning and content in general – or rather carelessly about sentences, thoughts, concepts and linguistic signs – the target of my argumentation is the original intentionality that makes the intentional contents of mental states possible.

Crucially, it is clear that for anything to be directed at something else, it cannot be directed at everything or at randomly chosen targets. This is very clear in the case of linguistic terms. They apply to certain things precisely to the extent they do not apply to everything or randomly chosen things. In this sense, concepts are essentially *rules* stating how terms may be applied: for a linguistic entity or a mental state to have conceptual content – to be intentional – its use must be rule-governed.<sup>140</sup> Basically, our thoughts have content when they are normatively connected with specific states of affairs and propositions. If this is not the case, then we cannot talk about a contentful (and thereby intentional) thought or sentence. The meaning of a sentence or thought is constituted by the norms governing the appropriate applications of the sentence or thought in question, *i.e.*, its assertability conditions. This kind of *inferential* or *normative role semantics* is, I think, the core of Wittgenstein's insight of seeing meaning as use.<sup>141</sup> Consequently, a behaviour is an action precisely to the extent that it has meaningful content, *i.e.*, to the extent that we can ask for reasons (and not only causes) for the behaviour.

Thus, a thought or sentence has meaning to the extent that it is embedded in a *normative system* (this, of course, is the thesis of *meaning holism* we saw to be an essential part of the Logical Connection Argument); when it can serve as a premise or

---

<sup>140</sup> I am of course alluding to the well-known problem of rule-following here. The problem is usually located in the so-called later philosophy of Wittgenstein (1953 in particular), but more often than not the discussion concentrates on Kripke's (1982) reaction to Wittgenstein's arguments. This essay is not the place for any kind of general review of the rule-following literature. Rather, I stick to my own task of constructing a positive argument *for* the non-causal view of intentional action and agency and draw from the rule-following discussion only when it is directly relevant for this quest. The key reactions to Kripke's arguments are collected into Miller & Wright (2002). Kusch (2006) analyses and answers virtually all the criticisms directed to Kripke's book. Another good critical discussion of the suggested solutions to the rule-following problem is Haukioja (2000).

<sup>141</sup> Thus I also think that the fundamental semantic unit is a move in a language game (Wittgenstein) or a judgement one can be responsible for (Kant), *i.e.*, a sentence or a proposition, and the sub-sentential entities such as singular terms should be explained in terms of the basic units (*e.g.*, rules governing the use of terms – concepts – determine to *what* one's judgement or sentence commits one); see Brandom (1994, 2000 & 2002 and Pagin 2002 for criticism). Since this Chapter is principally interested in the normativity common to all semantic notions, especially actions, and not in orthodox philosophy of language, I will talk rather carelessly about terms, sentences, concepts etc. without explicating these differences (*cf.*, however, A.1.2).

conclusion in inferences – including practical inferences featuring actions (as conclusions) and observations (as premises). Indeed, the inferential role *is* the content. I think that by now my sudden interest in semantics should be clear: the conditions for a sentence or thought to be intentional and to have meaningful content are exactly the same as the conditions for a behaviour to be an intentional action.

Actions are, by definition, meaningful behaviours, intentional states are meaningful states, words are meaningful utterances and so forth. To be intentional *is* to have meaningful content. To have meaningful content *is* to have a normative role or status in an inferentialist framework. Hence to show how meanings are possible in our ontologically naturalistic world *is* to show how there can be intentional actions in the non-causalist normative sense of the Logical Connection Argument in our ontologically naturalistic world. This, I argue, requires *social* holism.

### III.3.2 MEANING AND RULES

My first approximation for capturing the nature of meaning shall be the claim that any theory of meaning must be able to accommodate the following three conditions (adapted from Williams 1999, 159 and Esfeld 2001, 73-74 and explained and analysed in detail below):

- (1) Meaning is something that we understand immediately and completely.
- (2) Meaning determines the future applications.
- (3) Meaning sets the normative standards of correctness for the future applications.

These conditions are based on observations made by Wittgenstein. In his view, Condition (1) simply states an obvious fact about our use of meaningful language: “we *understand* the meaning of a word when we hear or say it; we grasp it in a flash” (Wittgenstein 1953, §138). Condition (2), in turn, requires that meaning is extended in time and space. The meaning of the term “cow” is not exhausted by a single application; it determines also which other objects are cows and which are not, and whether or not the term applies in situations one faces in the future. Thus, to call a dog a cow is a *mistake* (Condition (3)). In this sense meanings (concepts) are essentially *rules*, and meaningful, intentional behaviour, *i.e.*, action, is essentially *rule-governed behaviour*.<sup>142</sup>

---

<sup>142</sup> Hence Kant’s paradoxical-sounding statement that for one to be a truly free intentional agent capable of intentional actions one’s behaviour must be bound by rules is literally correct.

Take an example where someone learns the meaning of a term by being shown a sequence of examples of correct applications of the term. The motivation behind Conditions (1)-(3) is to account for problems created by the obvious point that *on some interpretation* any way to continue the finite sequence of examples can be seen to be compatible with the sequence (Kripke 1982, see also Chapter III, “The New Riddle of Induction”, in Goodman 1973).<sup>143</sup> Thus, if I have learned a concept on the basis of a sequence of examples, how can I be *justified* in thinking that I know how to go on to apply the concept? How am I to know that I have picked the *correct* rule out of infinitely many possibilities on the basis of the finite sequence of examples? What if the rule capturing the meaning of, say, the concept *cow* is that the term “cow” is to be applied to cows until 2007, and to horses thereafter? Or perhaps the point of the sequence was not that the term “cow” should be applied to cows, but to any four-legged, horned ruminants. “Whatever I do is, on some interpretation, in accord with the rule [*i.e.*, the concept governing the use of the term]” (Wittgenstein 1953, §198).

Condition (2), of course, is designed to rule out these kinds of problems. As Wittgenstein expresses the idea behind Condition (2), “[t]he rule, once stamped with a particular meaning, traces the lines along which it is to be followed through the whole of space” (Wittgenstein 1953, §219).<sup>144</sup>

In addition to this, a peculiar feature of meaning is that the use of concepts is *rule-governed*; it must be possible to apply a concept *incorrectly*. In other words, the theory of meaning must allow for cases when I am applying the concept *cow* incorrectly (say, to horses on a dark night). The rule cannot be identical with actual applications, for we must be able to say that in calling the horses cows I was nonetheless applying the concept *cow*, albeit mistakenly. This *problem of error* is captured by Condition (3). “Following a rule is analogous to obeying an order” (Wittgenstein 1951, §206). I must be able to think that I am following a rule correctly while actually failing to do so.

<sup>143</sup> Note, however, that Kripke’s and Goodman’s problems are not identical: Goodman is challenging inductive reasoning, whereas the target of Kripke’s argument is still a deeper issue, the ontology of meaning.

<sup>144</sup> In what follows this thesis is interpreted such that the extension of the rule (future applications) is determined via a social process. The *meaning finitism* of the so-called Edinburgh School sociologists of science emphasises this feature in a very provocative way: “the established meaning of a word does not determine its future applications” (Bloor 1983, 25) and hence “[t]he future applications of terms are open-ended” (Barnes, Bloor & Hendry 1996, 55). Condition (2) appears to deny this kind of meaning finitism, and even my final interpretation of (2) holds on to the idea that meanings *are* – at least in most cases, cf. Footnote 146 below – determined for each individual. However, the difference between my view and that of orthodox meaning finitism may be one of mere emphasis rather than a substantial disagreement, for also in my view the future applications are determined by a social practice (which indeed *is* open-ended), not by the term itself, individual interpretation or meaning as an abstract object. More on these issues in III.3.3 and Appendix.

Wittgenstein's goal is to show that understanding a meaning cannot be grounded on a mental act of consciously *interpreting* a sign. His argument is meant to refute both the Platonist and psychological theories of understanding meanings. In short, Platonism treats meanings as external objects that agents understand by connecting to them. The connection may be either purely causal or based on a non-causal interpretation. Psychologism, in turn, holds that meaning is based on a conscious interpretation of an external meaning, be that a Platonist entity, social use, or what have you. The core of Wittgenstein's view of meaning as use is that Platonism, psychologism and indeed all theories that dissociate the practice of applying a rule from the rule itself are bound to fail. In Wittgenstein's view meanings reside in our social practices, and to understand a meaning is to be able to *participate* appropriately in the social practices (the language game) that constitute the meaning, not to contemplate them from outside. Meaning is identical with (social) use.

The core of Wittgenstein's argument is that Platonism and psychologism cannot satisfy Conditions (1)-(3) simultaneously. Wittgenstein's general hostility towards philosophical theories suggests that conceivably in his view no explicit theory can do so (cf. McDowell's *quietism* in III.3.3), for perhaps (1)-(3), albeit unavoidable, simply are inconsistent. My goal, however, is to argue that (1)-(3) can be made consistent without challenging Wittgenstein's lasting intuitions concerning the essence of meaning. Nonetheless, also my view renders Platonist and psychological theories of meaning implausible and subscribes to a social practice view of meaning.

The starting point of Wittgenstein's argument is his observation that Conditions (1) and (2) are in fact incompatible: "we understand the meaning of a word when we hear or say it; we grasp it in a flash, and what we grasp in this way is surely something different from the 'use' which is extended in time!" (Wittgenstein 1953, §138). Thus, Wittgenstein formulates what has become known as the *Infinity Problem* of rule-following. The very point of learning rules (concepts) is that the same rule (concept) is thought to apply to infinitely many new cases. The question is, then, that when a person encounters a new situation to which she has never applied the rule, how is she to know how the rule applies to this new case? The obvious answer that the rule should be applied *in the same way* as in the earlier cases is not satisfactory, because the problem is precisely what counts as the same way (Kripke 1982, 8).

Thus, the Infinity Problem raises what Meredith Williams (1999, 159) calls the (Infinite) *Regress Argument* against rule-following and meaning. Condition (1) requires that meanings must be grasped in a flash. But surely nothing that can be grasped (in the



sense of a conscious interpretation) in a flash is capable of determining all the possible applications of what is grasped, for they are always open for new interpretations.

This, however, is exactly what Condition (2) requires of any acceptable theory of meaning. As Williams says, the Regress Argument “shows that in the end meaning must be something other than an act of interpretation or that which requires interpretation” (Williams 1999, 159). Thus the Regress Argument challenges psychological theories of meaning and understanding (including those versions of Platonism that build on interpretation). The idea is that if understanding a meaning amounts to an act of interpretation (and requires that the meaning is grasped in a flash), then meaning cannot extend beyond the very act of interpretation, or Condition (1) remains unsatisfied. But this means that Condition (2) remains unsatisfied. In other words, there seems to be a critical trade-off between Conditions (1) and (2). If understanding a meaning is based on a psychological act of interpretation, then this interpretation determines further application only on a new interpretation, and so on *ad infinitum*.

This is how Wittgenstein himself puts the argument: “Suppose, however, that not merely the picture of the cube [satisfying Condition (1)], but also the method of projection [satisfying Condition (2)] comes before our mind? [...] But does this really get me any further? Can’t I now imagine different applications of this schema too?”, Wittgenstein (1953, §141) asks. The answer he gives is that interpretations cannot indeed solve the problem, since “any interpretation still hangs in the air along with what it interprets, and cannot give it any support. Interpretations by themselves do not determine meaning.” (Wittgenstein 1953, §198.) If understanding a concept is based on conscious interpretation, it seems that every concrete application of the concept requires one to re-interpret also the method of projection – and whichever way one does it, one’s solution will be compatible with earlier applications on some interpretation, and Condition (2) remains unsatisfied.

However, it seems to me that the Infinity Problem is a problem precisely to the extent it gives rise to the Regress Argument. Consequently, if we can avoid the Regress Argument, then the Infinity Problem is not necessarily a problem for us. Thus it is crucial to notice that an essential component of the Regress Argument is the explicit commitment to psychologism, *i.e.*, the idea that to understand a meaning is to give a conscious interpretation or to grasp the meaning consciously. It is *psychologism* that leads to the infinite regress of interpretations.

In other words, the Regress Argument is simply another way of expressing the circularity problem I have formulated already several times above. What we are after is

an account of *original* intentionality *constitutive* of meaningful thoughts and human psychology, and hence to explain meaning in terms of conscious psychological interpretations would simply beg the question – we would need to ask the very same questions concerning the intentionality of the language of thought. The problem is to have a theory of meaning and intentionality, and thus to build on intentional and meaningful notions simply begs the question.

This is why the Regress Argument is a strong argument against all psychological (and interpretatively Platonist) accounts of meaning (*i.e.*, accounts building on intrinsically meaningful notions). We want to know how *original* intentionality is possible – an account of *derived* intentionality is already given in Part I of this study. The Regress Argument shows that essentially psychological theories are bound to simply beg the question when it comes to the explanation of original intentionality.

However, I do not think the Infinity Problem will remain with us no matter what. We can avoid the problem by rejecting the commitment to treat understandings of meaning as acts of conscious interpretation. This would evade the Regress Argument. I think this is also the route taken by Wittgenstein:

It can be seen that there is a misunderstanding here [*i.e.*, in thinking that to understand a meaning is to give an interpretation] from the mere fact that in the course of our argument we give one interpretation after another; as if each one contended us at least for a moment, until we thought of yet another standing behind it. What this shows is that there is a way of grasping a rule which is *not* an *interpretation*, but which is exhibited in what we call “obeying the rule” and “going against it” in actual cases.  
(Wittgenstein 1953, §201.)

So what is needed is an account of understanding as something that is not an instance of interpretation but shows in actual applications of concepts and in other meaningful events. But what could this kind of understanding be?

Wittgenstein offers useful guidance: “Try not to think of understanding as a ‘mental process’ at all. – For *that* is the expression which confuses you.” (Wittgenstein 1953, §154.) To understand a meaning is to obey a rule, and “‘obeying a rule’ is a practice” (Wittgenstein 1953, §202) and hence to understand a meaning is not a mental act but an ability to behave appropriately: “To understand a language means to be master of a technique” (Wittgenstein 1953, §199). The Regress Argument, and hence the Infinity

Problem, can be avoided only if we are absolutely clear that rule-following (understanding) is not a psychological concept at all.<sup>145</sup>

This requirement is the rationale behind Wittgenstein's famous dictum, according to which "[w]hen I obey a rule, I do not choose. I obey the rule *blindly*." (Wittgenstein 1953, §219.) G. P. Baker and P. M. S. Hacker sum up this position nicely: "the point of the argument was [...] to show that rule-following, and hence a language, is a kind of customary behaviour, a form of *action*, not of thought" (Baker & Hacker 1984, 21 – although here "action" cannot be read in the customary sense of *intentional* action, for that would again beg the question).

The task is, then, to give an account of rule-following such that rules are followed *blindly* or, in other words, so that understanding amounts to a *practical technique* requiring no contentful deliberation. One popular and prominent trend in the rule-following literature (*e.g.*, Blackburn 1984, Boghossian 1989, Forbes 1983-4 and Millikan 1990) is to hold that to avoid the Regress Argument we must say that rule-following, and thus understanding, is ultimately based on a biological *disposition* characteristic of the human species to continue sequences (to follow a rule) in a certain way.

Recall that the core of the rule-following problem is to explain how we are to pick out one rule out of infinitely many rules instantiated by a finite sequence of examples. The dispositional solution holds that it is a mistake to concentrate on the *logical* point that whatever one does is logically compatible with the sequence of examples. Rather, say the dispositionalists, we should concentrate on the fact that as biological beings it is often the case that one of the infinitely many logically possible ways to continue a sequence simply strikes us as the natural way to continue the sequence.

Perhaps evolution has equipped us with homogeneous dispositions to continue sequences of examples uniformly. The suggestion is that our shared biological nature makes us to share the (logically unjustifiable) sense of what counts as obviously the same. We know, for example, that animals are able to classify objects they face into those who belong to the same species with them and to those who do not on the basis of a very small sequence of examples. In Philip Pettit's words, "although any finite set of examples *instantiates* an indefinite number of rules, for a particular agent [or for members of particular species] the set may *exemplify* just one rule" (Pettit 1990, 36; my italics).

---

<sup>145</sup> Needless to say, most cases of everyday understanding *are* "mental processes" or even acts of interpretation. Recall, however, that what we are after here is understanding and intentionality at the most fundamental level of original intentionality that makes the psychological everyday concept of understanding possible in the first place.

In this manner, the dispositional suggestion goes, Kripke is right to hold that there is no logically compelling solution to the problem of why to continue a sequence in one way rather than another, but often enough there nonetheless is a specific way in which we as members of the species *Homo Sapiens* are *disposed* to continue it (cf. Esfeld 2001, 81). Assuming the existence of such a disposition (or a system of such dispositions) is, I think, the sensible way to understand the talk about an intrinsic, biological *language instinct* all humans are supposed to share. To put this suggestion in the terminology introduced above, the dispositional solution says that “my understanding of a concept *at* a time is a dispositional state” (Forbes 1983-4, 21), and hence we ought to replace Condition (1) with the following Condition (1\*):

(1\*) Meaning amounts to (or is constituted by) a disposition to react in a certain way (to continue a sequence of examples in a certain way).

This, I think, delivers exactly what we required. (1\*) captures the motivation behind (1), for certainly a directly activating disposition counts as grasping something in a flash. Hence we have a solution to the core of the rule-following problem: there are infinitely many ways to continue a sequence, but in virtue of our shared biological make-up we are disposed to pick exactly one of them. Moreover, since we are talking about a directly activating biological (causal) disposition, when we continue a sequence we do not consciously *choose* (interpret) but continue the sequence *blindly*, exactly as Wittgenstein (1953, §219) required. The act of picking out one of the infinitely many possibilities is not based on logical considerations or interpretations but on a biological disposition. Surely this counts as blind obedience exactly in the sense required to avoid the Regress Argument.

Finally, (1\*), unlike (1), *is* compatible with (2), since to have a causal disposition to continue a sequence in a certain way just *means* that the future applications are in most cases<sup>146</sup> determined for the person who has the disposition, which is precisely what Condition (2) requires (Blackburn 1984). Thus, replacing (1) with (1\*) explicates, I think, the motivation behind Wittgenstein’s way of avoiding the Regress Argument by appealing to blind obedience: “How can he know how he is to continue a pattern by himself – whatever instruction you give him? – Well, how do I know? – If that means

---

<sup>146</sup> But not in all cases. The view of meaning I am building here is naturalistic in the sense that there is no *a priori* guarantee that our concepts will apply smoothly to all possible situations. Perhaps in some outlandish situations our dispositions simply do not work and we cannot find anything to say.

‘Have I reasons?’ the answer is: my reasons will soon give out. And then I shall act, without reasons.” (Wittgenstein 1953, §211.)

The dispositional solution seems undoubtedly to be heading in the right direction. The question is, then, whether or not it *suffices* for meaning and intentionality? Note that if the dispositional solution indeed is sufficient, then causalist theories of meaning, and thereby of action and agency, unquestionably have the upper hand. In III.3.1 I argued that the rule-following problem captures the ultimate nature of meaningful action and agency. Thus, if the solution to the problem is essentially causalist, as biological dispositionalism is, surely causalism about intentional action and its explanation must take priority over non-causalism.

However, the dispositional solution to the problem of rule-following is inescapably insufficient. The reason is that the Infinity Problem, and the corresponding Regress Argument, is only the *first* part of the problem of rule-following. The second, more fundamental problem of rule-following is the *Normativity Problem* (cf. Esfeld 2001, 73).

The Infinity Problem was raised by the mutual incompatibility of Conditions (1) and (2). The incompatibility was resolved by replacing (1) with (1\*). The Normativity Problem, on the other hand, is created by the mutual incompatibility of (1) and (3), and replacing (1) with (1\*) serves only to make the problem worse. Recall that Condition (3) was assumed to capture the possibility of *failing* to follow a rule while trying to do so, *i.e.*, the possibility of making mistakes. As Kripke puts this, “[t]he relation of meaning and intention to future action is [also] *normative* [Condition (3)], *not* [only] *descriptive* [Condition (2)]” (Kripke 1982, 37).

The obvious problem is that if we solve the rule-following problem by appealing to a blind, causal disposition to simply continue a sequence in one way without reasons, then our solution to pick that particular way to continue the sequence cannot be *rationaly justified*. Reasons cannot be given for that solution; it is simply what we do (Wittgenstein 1953, §217). This is indeed what the Infinity Problem and the Regress Argument appear to require us to hold. But then the view falls prey to Kripke’s question: “Is not the dispositional view simply an equation of performance and correctness?” (Kripke 1982, 24). The essential normativity of meanings, captured by Condition (3), seems to have been lost.

The causal “blind” solution succeeds to make (1) and (2) mutually compatible, but renders the question of the normative aspect of rule-following completely unanswerable. I think Williams has got this right. “Though causal factors are relevant to our

understanding of rule-following, a fully causal account cannot make space for the basic normative distinction, that between correct and incorrect actions. [...] There are constraints on behavior that [...] are not purely causal.” (Williams 1999, 168.) I think this is also Wittgenstein’s view. He very clearly says that the dispositional solution is the *first* and *necessary* part of the solution to the problem of rule-following, but that as such it cannot be *sufficient* (Wittgenstein also emphasises – correctly, see III.4.2 – that the dispositions must be socially mediated):

“But how can a rule shew me what I have to do at *this* point? Whatever I do is, on some interpretation, in accord with the rule.” [...] “Then, can whatever I do be brought into accord with the rule?” – Let me ask this: what has the expression of a rule – say a sign-post – got to do with my actions? What sort of connexion is there here? – Well, perhaps this one: I have been trained to react to this sign in a particular way, and now I do so react to it [the dispositional solution]. But that is only to give a causal connexion; to tell how it has come about that we now go by the sign-post; not what this going-by-the-sign really consists in. On the contrary, I have further indicated that a person goes by a sign-post only in so far as there exists a regular use of sign-posts, a custom. (Wittgenstein 1953, §198.)

As Sellars (1963, 327) puts this, a mere “pattern governed” behaviour (the dispositional solution) cannot suffice for linguistic practices, for they are essentially rule-governed and, in Sellars’ terminology, require “rule obeying” behaviour that goes further than mere pattern governed behaviour. Sellars shares my reading of Wittgenstein in thinking that the normative rule obeying behaviour *involves* the dispositional, pattern governed behaviour, but the two cannot be identified with each other. Biological dispositions producing pattern governed behaviour are causal mechanisms to react to certain kinds of input with certain kinds of output. As Kripke sees, there is no room for a distinction between performance and correctness here, and hence they cannot suffice for the construction of linguistic practices.

Some dispositionalists see this problem. Ruth Garrett Millikan (1990), for example, builds on the distinction between *selection* and *selection for* (e.g., Sober 1993, 83). To use Fodor’s (1990, 71 ff.) example, evolution has *selected* frogs with a mechanism that causes them to snap their tongue at any little black thing flying around them, but evolution has nonetheless *selected for* a mechanism which snaps at flies. Millikan’s claim is that since it is snapping at flies that is causally responsible for the evolution of the mechanism, we can say that snapping at flies is the *proper function* of the mechanism and that when the mechanism snaps at non-flies, the mechanism is *malfunctioning*,

and hence the solution building on evolved, biological dispositions does leave room for normativity.

I think the crucial ingredient here is that *we* can redescribe biological dispositions as normative functions. As such, biological dispositions are simply blind, non-normative causal mechanisms. In Part I it was argued that Searle is correct to hold that biological functions are not intrinsically normative; their normativity is always assigned to them in virtue of being embedded into a system of values constructed and maintained by intentional agents. Similarly, earlier in this Part I have explained how von Wright and Davidson thought that normative notions follow a unique logic and hence “have no echo” in the realm of blind, causal mechanisms. The anti-naturalist claim that the biological world is intrinsically normative is, paradoxically, the (in my mind unacceptable) core of the allegedly naturalistic programmes of, *e.g.*, Millikan and Dretske (1988). Their idea is not far from, say, objective idealism where the natural world is assumed to be largely conceptual and rational so that we, as rational beings, can have epistemic access to it (see Appendix).<sup>147</sup>

In short, intentional phenomena cannot be identified with causal factors (including causalist Platonism), for that would fail to resolve the Normativity Problem. “No one ever acts incorrectly in the sense of violating his or her own dispositions. Indeed, to talk of ‘violating’ dispositions is illicitly to import normative vocabulary into a purely descriptive context.” (Brandom 1994, 29.) The dispositional solution commits a fundamental philosophical category mistake, because it is not sufficiently sophisticated to distinguish between Conditions (2) and (3) or, in other words, between *causal* determination and *normative* determination. As Robert Brandom (1994) likes to express this, meaning cannot be understood in terms of mere causal *properties*; what matters for meaning are the *proprieties* of use. Or as Peter Winch explains, “the notion of following a rule is logically inseparable from the notion of *making a mistake*. If it is possible to say of someone that he is following a rule that means that one can ask whether he is doing what he does correctly or not.” (Winch 1958, 32.)

---

<sup>147</sup> My objection to Millikan proceeds, in a sense, at the meta-level. For a decisive criticism of Millikan in her own terms, see Haukioja (2000, 35-38).

### III.3.3 KRIPKE'S SCEPTICISM, THE NAÏVE COMMUNITARIAN VIEW AND MCDOWELL'S QUIETISM

The situation I have arrived at is the core of Kripke's (1982) *meaning scepticism*. Kripke thinks that to have a positive theory of meaning, Conditions (1)-(3) must be satisfied, but the Infinity Problem and the Normativity Problem together appear to show that the conditions cannot be met:

The sceptical argument, then, remains unanswered. There can be no such thing as meaning anything by any word. Each new application we make is a leap in the dark; any present intention could be interpreted so as to accord with anything we may choose to do. So there can be neither accord, nor conflict.  
(Kripke 1982, 55.)

However, it seems to me that the precise nature of Kripke's scepticism is often misunderstood. He does not think, contrary to what the quotation above seems to imply, that there are no meanings – “the sceptical conclusion is insane and intolerable” (Kripke 1982, 60).

Kripke is not a sceptic about meaning, but about a certain approach to the theory of meaning. In Kripke's view the rule-following considerations show that any *fact-based* theory cannot capture the nature of meaning. Kripke's point is that if we see meanings as independently existing facts in need of interpretation, we run into the Infinity Problem (the incompatibility of (1) and (2)). If we see meanings as based on naturalistic facts about us, such as dispositions, we run into the Normativity Problem (the incompatibility of (1\*) and (3)). Kripke thinks that the sceptical challenge, *i.e.*, the mutual incompatibility of Conditions (1), (2) and (3), cannot be answered insofar as we think that our account of meaning must be *fact-based*: “Now if we suppose that facts, or truth conditions, are of the essence of meaningful assertion, it will follow from the sceptical conclusion that assertions that anyone ever means anything are meaningless” (Kripke 1982, 77).

However, in Kripke's view this is not surprising, for he thinks that the central message of *Philosophical Investigations* is to give “a picture of language based, not on *truth conditions*, but on *assertability conditions* or *justification conditions*: under what circumstances are we allowed to make a given assertion?” (Kripke 1982, 74). Thus, in Kripke's mind an appropriate answer to this question must not be stated in terms of truth-conditions, for the question does not amount to a naturalistic search for a brute fact. Rather, the question is inherently *normative*: when am I *entitled* to make a given



assertion, or when am I *required* to make it? The question is about a normative status, rights and duties, not about a brute fact of the matter. Meanings are norm-based, not fact-based. When we want to find out what is the meaning of a given judgement (or a word or concept), we are not searching for an independent fact that establishes the meaning. Rather, we have to find what is the appropriate way of using the judgement (word, concept) in our language game. We are searching for a role or status, not a fact.

When we acknowledge this, no sceptical conclusion follows (Kripke 1982, 77):

If Wittgenstein is right, we cannot begin to solve it [the sceptical problem] if we remain in the grip of the natural presupposition that meaningful declarative sentences must purport to correspond to facts; if this is our framework, we can only conclude that sentences attributing meaning and intention are themselves meaningless. [...] The picture of correspondence-to-facts must be cleared away before we can begin with the sceptical problem.

(Kripke 1982, 78-79.)

It is very difficult to see what, exactly, Kripke argues for here. It seems to me that his line of thought must be the following. Take the following judgement in Finnish: “Kissa on matolla”. How could someone who does not know Finnish know that she has understood the meaning of this judgement? She may be given a sequence of examples of when this sentence is correctly applicable. And although there are infinitely many ways of coherently (under some interpretation) continuing the sequence, she may find it natural to concentrate on the fact that in all those cases there was a cat on a mat in front of the person uttering the sentence. Thus, she might become disposed to apply the sentence to cases where a cat is on a mat.

Kripke’s point seems to be that there cannot be an independent brute fact of the matter as to whether the learner is now in possession of the meaning of the sentence. In particular, the disposition she has acquired cannot constitute such a fact, for this view would run into the Normativity Problem. Similarly, her knowing the meaning cannot be based on the learner consciously grasping *the idea* (a Platonist meaning) behind the sentence, for this would fall prey to the Regress Argument. It seems to me that Kripke’s emphasis on (later) Wittgenstein’s rejection of truth-conditional semantics suggests that in his view we cannot describe *independently existing* (in terms of Part I) conditions that would unambiguously dictate whether the learner has grasped the meaning.

The best we can do, Kripke thinks, is to see whether we find her use of the sentence acceptable. If we do, we should be willing to treat her as someone who understands the sentence, *i.e.*, assign her a certain status and allow her to participate in the practice of using the sentence. Crucially – and this is the radical core of Kripke’s argu-

ment – if we recognise her usage as acceptable, this is not evidence for any practice-independent *fact* that she indeed understands the meaning. Rather, this kind of social recognition of her competence to participate in language games is what understanding a meaning consists in.

Thus, Kripke writes, “Wittgenstein’s sceptical solution concedes to the sceptic that no ‘truth conditions’ or ‘corresponding facts’ in the world exist that make a statement like ‘Jones, like many of us, means addition by ‘+’ true. Rather we should look at how such assertions are *used*.” (Kripke 1982, 86). The meaning of the assertion is its normatively defined place in our language games. Its meaning is given by the conditions of its appropriate assertability, and the rule-following considerations guide us to see that the answer to this question cannot state a brute fact. The answer – and therefore the theory of meaning avoiding the sceptical conclusion – must be an essentially *normative* judgement.

According to Kripke (1982, 74 ff.), this changes the way in which we should approach intentional phenomena. To know the meaning of a sentence is to be able to use it appropriately. This is of course very close to the view of von Wright and Davidson (III.1) who thought that the meaning of a judgement is constituted by the norms governing the appropriate *use* of the judgement, *i.e.*, normatively defined role in a language game that specifies when one is allowed to form the judgement and what follows from it. The main difference between Kripke and von Wright/Davidson is that, contrary to what the rule-following considerations might seem to imply, von Wright and Davidson insist that even a rule-governed use of a single judgement (not to mention a single word) cannot be enough for the judgement (or word) to have conceptual content. Content requires that not only is the use of the judgement rule-governed, but also the rational relations between several judgements must be rule-governed. This is the aspect of the *Quinean meaning holism* in von Wright and Davidson’s view. I think the point is easy to understand by looking at an example by Brandom (who also accepts the view):

You do not convey to me the content of the concept *gleeb* by supplying me with an infallible gleebsness tester that lights up when and only when exposed to gleeb things. I would in that case know what things were gleeb, without knowing what I was saying about them when I called them that, what I found out about them or committed myself to.  
(Brandom 1994, 122.)

Mastery of assertability conditions does not suffice for conceptual content. Rather, we understand the meaning of a judgement when we recognise (in practice) in which situations we are entitled to endorse it *and* what other judgements (including

meaningful actions) we become entitled and committed to if we endorse it, and what other judgements are excluded by endorsing the judgement we are contemplating. Thus, it is a mistake to see the Wittgensteinian slogan that meaning is use as a commitment to behaviourism (cf. Williams 1999, 242, Stoutland 1988). As Brandom explains, the Wittgensteinian normative practice view implies a crucial difference between “a spectrophotometer [...] hooked up to a tape recorder in such a way that it produces a noise of the acoustic type ‘That is red’ when and only when it is irradiated with light of the proper frequency” and “a fanatical human red-reporter nearby [...] [with] just the same responsive dispositions to produce those noises” (Brandom 1994, 88).

The machine simply reacts to the input in accordance with its dispositions. For the human, on the other hand, the dispositional story is not the whole story. To the extent that she is capable of contentful thoughts and language, her emitting the noise counts as a *commitment* that (normatively) brings with it other commitments and excludes still some others – and it is precisely such normative (logical) connections that make the noise a *contentful judgement* and not a mere series of sounds. Part of the meaning of the judgement “That is red” is that by endorsing the judgement one is committed to holding that the object in question is colourful, that it is not green *etc.* Similarly, endorsing the judgement may commit one to certain actions (if, for example, one has previously become committed to red-related desires). These normative relations give the judgement its conceptual content (Brandom 2000, 48).

The *Quinean meaning holism* at the centre of the view of Brandom, von Wright and Davidson already presupposes that we are able to follow rules, *i.e.*, that there is a solution to Kripke’s problem. Thus the rule-following considerations add to the meaning holism (i) a further argument as to why purely causal (dispositional) solution cannot suffice for meaning and (ii) a requirement of also *social holism*. Or so I argue.

However, the topic of this study is not the Quinean holism (although I return to it in Appendix). Rather, for the present purposes it suffices to see that the fundamental Kripkean practice theory of meaning that I defend does not take *representation* or correspondence to facts to be a primitive semantic notion. That role is reserved for a position in a normative linguistic practice. This is often referred to as Wittgenstein’s replacement of his earlier representational and descriptive picture theory of language with his later view of seeing language as normative practice.<sup>148</sup> This view is both controversial and

---

<sup>148</sup> As I argue below (especially in Appendix), this – *pace, e.g.,* Rorty (1979) – should not be seen to imply that representation and truth conditions are not important. They are. The claim is rather that they are not primitive: representation and truth conditions are to be explained in terms of a normative, linguistic practice and not *vice versa*.

complex: I hope the rest of my dissertation offers an explanation of what this view really amounts to in the case of meaningful actions (including not only speech acts but all intentional actions).

First of all, it is important to notice that at precisely this point Kripke begins to build a social element into his theory of meaning. Although we need some kind of dispositional solution that provides the *blindness* required for the Infinity Problem, the practice within which an assertion can be meaningful must be a *social* practice, for only the social element can give a normative standard of correctness to the assertions of individuals (Kripke 1982, 88). This, Kripke suggests, is what also Wittgenstein had in mind in *Philosophical Investigations*:

[T]here is a way of grasping a rule which is *not* an *interpretation*, but which is exhibited in what we call “obeying the rule” and “going against it” in actual cases.

(Wittgenstein 1953, §201.)

And hence also ‘obeying a rule’ is a practice. And to *think* one is obeying a rule is not to obey a rule. Hence it is not possible to obey a rule ‘privately’: otherwise thinking one was obeying a rule would be the same thing as obeying it.

(Wittgenstein 1953, §202.)

In my view Kripke’s social theory of normative semantics is basically correct. Unfortunately, it is not at all clear what, exactly, the theory is. In the next Chapter and Appendix I give my own argument as to what the best way of explicating the view is. However, to motivate that argument we need to see why a popular reading, which I call the *naïve communitarian view*, cannot be an acceptable reading of Kripke.

In short, according to the naïve communitarian view, Kripke holds that individuals just do what comes naturally to them (and thus bypass the Infinity Problem), and then *compare* their individual applications of a rule to the applications of the members of their linguistic community. The normative element is brought into the picture by this comparison. An application (a judgement) is correct if it agrees with the applications of the members of the community, and it is incorrect if it differs from the communal use.

This, however, is *not* what Kripke has in mind. We saw before that Kripke (1982, 111) very explicitly rejects all theories (social or otherwise) that try to describe truth-conditions for attributions of meaning. Wittgenstein’s message is that we must look for practical assertability conditions, not factual truth conditions.

The naïve communitarian view was, it seems to me, held in the early 1980s by, for example, Crispin Wright (1981) and Christopher Peacocke (1981).<sup>149</sup> Here is how Wright formulates the view:

None of us unilaterally can make sense of the idea of correct employment of language save by reference to the authority of securable communal assent on the matter; and for the community itself there is no authority, so no standard to meet. [...] we shall reject the idea that [...] a community goes right or wrong in accepting a particular verdict on a decidable question; rather, it just goes.  
(Wright 1981, 105-106.)

In this picture the *communal* practice is neither constrained by any facts about meaning nor is it subject to any kind of rational criticism, because it itself is the ultimate constituent of meaning and rationality.<sup>150</sup>

My view is that this approach is on the right track. Whatever solution to the problem of rule-following is to be accepted, the solution must be able to accommodate the dispositional solution to the Infinity Problem. However, also the Normativity Problem must be taken into account, and I think Kripke and others are right to think that this requires the theory to be essentially social.<sup>151</sup> The 1981 naïve communitarianism of Peacocke and Wright is nonetheless unacceptable for at least two reasons.

Not surprisingly, the first reason why the naïve communitarian view fails is that it subscribes to the crude *summative* view of collective acceptance and (linguistic) groups that was criticised and rejected in Part I of this study. As was argued in detail

<sup>149</sup> Some formulations of the meaning finitism of the Edinburgh School, e.g., Bloor (1983) and Barnes, Bloor & Hendry (1996) appear to be committed to the naïve communitarian view too. Below, however, I argue that at least Barnes (2000), Kusch (1999) and, perhaps, Bloor (1997) can be read as moving towards a more acceptable position.

<sup>150</sup> Recall that the reason why Collin and Niiniluoto (I.3) wanted to reject the kind of theory of social facts I defend was precisely that they thought that when applied to language it implies the naïve communitarian view. However, in this Chapter I reject the naïve communitarian view, in the next I show that the communitarian view that follows from my theory is very different from the naïve version and finally Appendix argues that my view of social practices avoids the problems identified by Collin and Niiniluoto.

<sup>151</sup> Baker and Hacker (1984), as well as, e.g., Haukioja (2000), Pettit (1993) and Tuomela (2002) think that the crucial point is that meanings are based on practices, and it is at most a contingent fact that the linguistic practices of social creatures such as humans happen to be *social* practices. I, however, think that, e.g., Barnes (2000), Bloor (1997), Brandom (1994), Esfeld (2001), Haugeland (1990), Kripke (1982) and Kusch (1999) are right in thinking that the Normativity Problem shows that the practices must be social. We need an external standard of correctness, and only a social aspect can deliver it in a way that does not run into either the Infinity or the Normativity Problem all over again. It seems to me that the main motivation for the non-social view is Blackburn's (1984) claim that whatever a social dimension can deliver, the same can be achieved by letting the different time-slices of one individual to play the role the different individuals play in the social picture. I have already argued against this kind of view in Part II, and I think that in the present context it is even clearer that this cannot work. There is no interaction between the time-slices and hence the correct and the factual are always equated – in which case normativity is in fact lost (see Esfeld 2001, 89-91 for a similar argument and Kusch 2002, 181 ff. for a detailed classification of the degrees of communitarianism and a criticism of the intuitive arguments for the non-social view). Admittedly Blackburn's argument may apply to the naïve communitarian view which, as I explain below, is essentially individualistic and thus privileges an individual's perspective in the same way as Blackburn's time-slice view privileges the present time-slice.

there (I.1.5 and, in particular, I.2.3), summative accounts cannot capture the foundations of human sociality. In the present context, the summative identification of correctness with a generalisation about what the others typically do confuses performance with correctness just as much as the individualist dispositional view does (Kripke 1982, 111). This is how Paul A. Boghossian voices this objection:

The community [...] will be disposed to call both horses and deceptively horsey looking cows on dark nights ‘horse’. [...] The communitarian, however, cannot call them [*i.e.*, applications of ‘horse’ to sufficiently horsey looking cows] mistakes, for they are the community’s dispositions. He must insist, then, firm conviction to the contrary notwithstanding, that ‘horse’ means not *horse* but, rather, *horse or cow*.  
(Boghossian 1989, 173.)

The naïve communitarianism would imply a view according to which the community is infallible, and as Boghossian insists, this is hardly plausible (cf. I.3).

Moreover, the naïve communitarian view is not Wittgenstein’s view: “From its *seeming* to me – or to everyone – to be so, it doesn’t follow that it *is* so” (Wittgenstein 1969, §2; see also 1953, §241). Or as Williams puts this, “[a]n empirical generalization about what most people do is not the same as a norm standing for what people ought to do” (Williams 1999, 165; see also Baker & Hacker 1984, 71). To put this in Kripke’s (1982) terminology, the infallibility of the community, which amounts to an intolerable form of cultural relativism (cf. Appendix), is due to the naïve communitarian view’s commitment to a truth-conditional or fact-based (the fact being the generalisation) view of meaning attributions. An acceptable version of communitarian semantics must *not* imply “that the answer everyone gives to an addition problem [Kripke’s main example of rule-following] is, by definition, the correct one” (Kripke 1982, 112).

Since in what follows I defend an essentially Kripkean, social theory of semantics, I must be able to demonstrate that Kripke is right in thinking that the rejection of fact-based accounts of rule-following allows us to resist the strong relativism of the naïve communitarian view. This is done in Appendix. As can be anticipated, the remedy I favour is the one used in Part I to develop an account of social groups and social action that overcomes the problems of the summative view. Normativity, I argue, cannot be assimilated with a statistical notion. Normativity comes from collective acceptance, which, as I argued in Part I, is not a summative (statistical) notion at all.

The second reason why the naïve communitarian approach fails is perhaps conceptually more interesting. Recall that the core of the Regress Argument is that to understand a meaning cannot be based on an interpretation of a rule external to the prac-

tice of following the rule. In the naïve communitarian picture, however, we have the individual application on the one hand and the norm as a statistical generalisation on the other. As we saw Kripke emphasising, this is a programme doomed to failure. Meanings are based on normative practices, not on non-normative representations and facts. Put differently, the naïve communitarian view accounts for applications of a rule in terms of the *rule* “an application is correct if it agrees with the communal use, and it is incorrect if it disagrees with the communal use”. To account for rule-following in terms of rules is, of course, blatantly circular.

Thus, as Baker and Hacker (1984, 96) emphasise, to avoid the Regress Argument we must concede once and for all that the practice of applying a rule and the rule itself cannot be separated from one another. In their terminology, the connection between them is *internal* and *grammatical*. If, as I have argued, both the Platonist proposal of appealing to an external, universal idea that can be *interpreted* to determine meaning and the direct psychologism that builds on conscious interpretations of signs lead to an infinite regress of interpretations, surely the same regress follows regardless of whether the object of interpretation is a statistical generalisation, a sign or a Platonist idea. A generalisation can be a norm only if it is *interpreted* as one, and hence we are back where we started.<sup>152</sup>

To conclude this Chapter, let me explicate the position we have arrived at. First, we saw that to account for intentionality (including intentional action), we need to show how meanings, or contentful thoughts and mental states are possible in our naturalistic world. This requires us to formulate a theory of meaning (or content) that succeeds in accommodating Conditions (1), (2) and (3). It seems clear that to avoid the Infinity Problem we need to replace (1) with (1\*), *i.e.*, to accept the animating thought of the dispositional solution. But this very move seems to make Condition (3), the essential normativity of meaning, inaccessible. John McDowell captures this dilemma well:

Wittgenstein’s problem is to steer a course between a Scylla and a Charybdis. Scylla is the idea that understanding is always interpretation. This idea is disastrous because embracing it confronts us with the dilemma of §4 above [*i.e.*, what I have called the Infinity Problem, the Regress Argument or simply the mutual incompatibility of (1) and (2)] [...]. We can avoid Scylla by stressing that, say, calling something ‘green’ can be like crying ‘Help!’ when one is drowning – simply how one has learned to react to this situation [*i.e.*, the dispositional solu-

---

<sup>152</sup> It is of no help to say that the Regress Argument was raised by the Infinity Problem, and the naïve communitarian view appeals to dispositions in that context, and to interpretations only in the context of the Normativity Problem. Surely any application agrees with the communal practice on some interpretation, and hence the naïve communitarian view has only managed to raise the Regress Argument *also* in the context of the Normativity Problem.

tion of replacing (1) with (1\*). But then we risk steering on to Charybdis – the picture of a basic level at which there are no norms; if we embrace that, [...] then we cannot prevent meaning from coming to seem an illusion. (McDowell 1998a, 242.)

Moreover, we have seen that the attempt to avoid the dilemma in terms of the naïve communitarian solution serves only to make things worse. It leads to unacceptable relativism and, further, re-invokes the Regress Argument in another context. However, just as we need to keep the core of the dispositional solution on board, I think we need to remain faithful also to the key insight of the communitarian view, namely that the solution must be essentially social, since only interaction between an individual and her community can give rise to the required normativity in a way that is able to defeat the two problems of rule-following and deliver objective meanings (III.4).<sup>153</sup>

Again, McDowell is on the right track.

The [...] key to finding the indispensable middle course is the idea of a custom or practice. How can a performance both be nothing but a ‘blind’ reaction to a situation, not an attempt to act on an interpretation (avoiding Scylla); and be a case of going by a rule (avoiding Charybdis)? The answer is: by belonging to a custom [...], practice [...], or institution [...]. (McDowell 1998a, 242.)

What I have claimed might be put like this: Wittgenstein’s point is that we have to situate our conception of meaning and understanding within a framework of communal practices. (McDowell 1998a, 243.)

As Williams (1999, 168-169) explains, what is needed is a *social* solution where the relation between an individual application and the communal practice is *not* based on interpretation. Rather, we need a view that allows a *practical* agreement of the community to be displayed *in* the behaviour of individuals (Williams 1999, 175-177). As Wittgenstein puts it, “[t]o use a word without justification [*i.e.*, blindly] does not mean to use it without right [*i.e.*, non-normatively]” (Wittgenstein 1953, §289).

In other words, I think Kripke, McDowell and Williams are absolutely correct in thinking that intentionality and meaning must, ultimately, be based on our *bedrock* rule-following practices that are essentially both *social* (making room for normativity) and *blind* (avoiding the Regress Argument). The problem is that Williams has nothing precise to say concerning what such bedrock practices could be like. She can explain what

<sup>153</sup> Note that this conviction reflects straightforwardly the argumentation in Part I, where I argued that social institutions, which are ultimately *norms of appropriate behaviour*, must be assumed to be *collectively* upheld in order to secure the *epistemic objectivity* of the norms in question.



the bedrock practices must deliver, but she cannot say how they do it. In particular, she does not explain *why* the emphasis on sociality solves the Normativity Problem.

McDowell has similar problems. Also he knows what is required, but he does not know how to answer to this requirement. When explaining the nature of the bedrock practices he, like Williams, repeatedly falls back on vague illustrations and metaphors, such as his statement that “a linguistic community is conceived as bound together [...] by a capacity for a meeting of minds” (McDowell 1998a, 253). McDowell is well aware that the naïve communitarian solution is unacceptable *and* that he does not really offer an alternative way of thinking about social bedrock practices. However, McDowell thinks that this is inevitable and, to support this, he asks us to recall what Wittgenstein says about the bedrock rule-following practices.

“How am I able to obey a rule?” – If this is not a question about causes, then it is about the justification for my following the rule in the way I do.  
If I have exhausted the justifications I have reached bedrock, and my spade is turned. Then I am inclined to say: “This is simply what I do.”  
(Wittgenstein 1953, §217.)

McDowell thinks that Wittgenstein is telling us that we should not even try to dig below the bedrock. Accordingly, McDowell concludes that we have to treat social bedrock practices as primitive constituents of the human form of life that cannot be analysed further, since all rational discussion already presupposes them. Thus McDowell’s position is that of *quietism*. As agents we cannot deny the reality of meaning and intentionality (since to hold *any* view, including the view that there are no meanings, *presupposes* meanings), and the rule-following arguments show that they presuppose social bedrock practices that are both normative and blind. Hence we must assume the existence of such practices even if we cannot say anything more about them.

However, I fail to see how this position is any more acceptable than the naïve communitarian view. After all, McDowell simply holds that an acceptable solution presupposes something about which we can say nothing whatsoever. I think Haukioja’s criticism of McDowell captures this accurately:

The important question that remains is: *how* does a practice give rise to meaning? [...] we need to know *how* shared membership in a community equips us to make our minds available to one another, *how* shared command of a language equips us to know one another’s meaning without interpretation. To say that it *just does* so equip us is really only to say that we *just do not know* how it does so.  
(Haukioja 2000, 52.)

Unfortunately, McDowell gives no answers to Haukioja's important questions. Thus it seems to me that "social bedrock practices", as McDowell and Williams use the expression, is at risk of turning into the kind of empty phrase which in fact explains nothing whatsoever and which Wittgenstein called a philosophical superlative. "You have no model of this superlative fact, but you are seduced into using a super-expression. (It might be called a philosophical superlative.)" (Wittgenstein 1953, §198.)<sup>154</sup> Hence, if the choice between the quietism of McDowell and Williams, the naïve communitarian view and the sceptical conclusion really is the last word on intentionality and meaning, the chances of the intentionalist programme certainly do not look good. In fact the essence of social science, the study of *meaningful* behaviour, is at risk.

The rule-following considerations show that the causalist approach, which accounts for contentful mental states in terms of their causal role, cannot be sufficient (because of the Normativity Problem). In III.2 I argued that – Davidson's criticisms notwithstanding – non-causal, normative view remains a viable option to causalism. Now, however, it seems that the non-causalist is running into other problems. In particular, in line with McDowell's quietism the non-causalist appears to be compelled to assume that the required normativity is created by social bedrock practices, even if we cannot tell how. Perhaps we must simply accept as given that human action, thought and all other meaningful activities just include both a non-normative blind element and a normative element. Maybe this kind of unanalysable dualism of the normative on the one hand and the non-normative on the other is not as unacceptable as full-blown Cartesian mind-body dualism, but still it would heavily undermine the general naturalism that I have advocated throughout my dissertation.

Fortunately, however, the theories developed in Parts I and II of this study show the way towards an account required for an acceptable solution to the problem of rule-following. The solution is not to discard completely the dualism of the normative and the non-normative, for in order to save intentional agency an irreducible space of reasons (the Kantian Kingdom of Ends<sup>155</sup>) is required. Thus our goal must be to understand

---

<sup>154</sup> To appeal to Wittgenstein when criticising McDowell's quietism is of course risky in the sense that McDowell (especially 1994) thinks that proper understanding of Wittgenstein amounts to understanding that questions such as Haukioja's miss the point fundamentally and thus do not deserve any other answer than "a shrug of the shoulders" (McDowell 1994, 178). Although McDowell's view may be justified as an explication of how Wittgenstein in fact thought, my aim in what follows is to show that a view that both acknowledges McDowell's achievements *and* answers to Haukioja can be constructed.

<sup>155</sup> Thus I subscribe to the Kantian idea that human agency can be acknowledged only if we admit that qua agents we are not inhabitants of the empirical world (Sellars' logical space of causation and the natural sciences) but the Kingdom of Ends (Sellars' logical space of normative reasons). My view is nonetheless essentially anti-Kantian in the sense that in this Chapter I have argued that the rule-following considerations demonstrate that the Kingdom of Ends *cannot* be understood in terms of universal, Platonist

how such a space is constructed and maintained within social practices without reducing the space into non-normative notions. An acceptable theory of meaning must be essentially normative. In other words, we need as much dualism as we can have within general naturalism, and this requires a social theory of rule-following with an intersubjective notion of normativity. We have to accept naturalised methodological holism.

---

principles that we reach via reason and interpretation (because of the Regress Argument). Such a view renders agency just as impossible as dispositional causalism does. In the following Chapters I seek to replace the Kantian noumenal principles with blindly (implicitly) normative social practices. One could say that the Kantian aspect of my argumentation corresponds to what Rorty has identified as the need for analytic philosophy to move from its Cartesian phase to a Kantian phase and the anti-Kantian aspect to Rorty's insistence that analytic philosophy ought not to halt at the Kantian phase but to move on to Hegel and Wittgenstein's philosophy in the sense of preferring historical practices to the timeless principles of Kant and Plato (see Brandom 2002 for a similar line of thought). In fact, I think the so-called continental tradition in contemporary philosophy has been much more able to appreciate the socio-historical nature of the principles of rationality than the Anglo-American analytic tradition, which has all too often seen the naturalistic anti-Platonism as a form of objectionable relativism (see Appendix).

## CHAPTER III.4: DIGGING BELOW THE BEDROCK

### III.4.1 INTRODUCTION

David Bloor (1997, 64-65) argues that Kripke's conclusion that all fact-based theories of meaning are bound to fail is seriously unfounded. The failure of theories based on psychological, Platonist or dispositionalist (causal) facts leaves room for a fact-based theory that builds on an essentially *social* fact, when social facts are understood as something different from mere aggregations of facts about individuals (Part I).

I think this disagreement between Bloor and Kripke is largely based on a misunderstanding. In particular, what seems to be at stake is confusion concerning what, exactly, the *sceptical conclusion* is. Bloor appears to read it as the claim that in Kripke's view objective meanings are impossible, full stop. If this is Kripke's sceptical conclusion, then Bloor's plea for social facts seems justified. Part I argued that there is a class of social facts that are essentially different from individual facts and their aggregations (thus, this is not dispositionalism or psychologism) such that these facts are, despite their ontological dependence on individuals taken collectively (thus, this is not Platonism), epistemically fully objective and external to all individuals taken singly. Therefore, it seems, *pace* Kripke's insistence on the impossibility of fact-based accounts of meaning, that a possibility of a theory building on social facts remains, for Kripke does not appear to acknowledge this class of facts.

However, this is not what Kripke means by a sceptical conclusion (III.3.3). His sceptical conclusion holds that a fact-based account in the sense of an account building on non-normative features cannot succeed. However, he thinks that objective meanings are real. This implies in Kripke's view that an acceptable theory of meaning must build on collective practices governing appropriate behaviour. However, as we saw in Part I, *this is exactly what social facts ultimately amount to*. Thus, I conclude, to give a fact-based account of meaning in Bloor's sense, *i.e.*, an account building on truly social facts, *is* to explicate precisely the kind of social, normative practice theory of meaning that Kripke argues to be our only hope for a theory of meaning that avoids the sceptical conclusion. The aim of this Chapter is to provide such a theory.

The animating thought behind the theory I present is that that a satisfactory solution of the rule-following problem must be essentially social. The social aspect is required for addressing successfully the Normativity Problem: a rule-follower needs an

external, objective standard to differentiate between performance and correctness. But the standard can be neither a fact (be it social or Platonist) in need of interpretation nor a mere cause affecting the follower's behaviour with no room for normativity. The standard must be such that it assesses an individual's solution as correct or incorrect without appealing to interpretation. The argument I shall give is that the only way of fully appreciating this is to see rule-following essentially as a social practice.

Before I proceed to my argumentation I need to bring in a caveat. I of course have not analysed each and every particular attempt to resolve the rule-following problem in a non-communitarian way, for this dissertation is not the right place for that (cf. Kusch 2006 that does a pretty good job in this respect). Instead, I have tried to show at a more general level why *any* view that subscribes to psychologism, Platonism, naïve communitarianism or purely causal dispositionalism is bound to fail to account for either the Infinity Problem or the Normativity Problem of rule-following. Moreover, in this Chapter I show that a more sophisticated form of communitarianism can triumph over such problems. However, it is conceivable that one day a profoundly new *kind* of (i.e., such that we cannot at the moment envisage it at all) non-social solution to the rule-following problem may nonetheless be developed. Thus, although I have argued that the essential features of the rule-following problem I have analysed make non-social theories unsuitable for solving the problem, perhaps there are possible ways of resolving the problem that are different from mine, although I cannot imagine what they would be like.<sup>156</sup>

However, this small, unavoidable disclaimer should not be seen to undermine the importance of this Chapter too much. Recall that the main aim of this and the preceding Chapter is to provide a positive argument for the normative view of agency and mental realism after the negative arguments of III.1 and III.2, and not to do general philosophy of language. In particular, my aim is to use the social solution to the rule-

---

<sup>156</sup> Cf. Kusch (2006, 182), who argues that all arguments in favour of the social solution must be open in this way and, thus, since the equation of performance with correctness closes the door for non-social but non-fact-based theories of rule-following, “[t]he strongest conclusion to be drawn is that given currently available versions of meaning determinism [fact-based theories in my terminology, be they private as in dispositionalism or social as in the naïve communitarian view], private rule-following is impossible” and the only option we have left is the view defended in this Chapter, i.e., the combination of communitarianism and the rejection of fact-based views. However, although my conclusion must remain open in this sense, I should add that my argument is nonetheless stronger than what Kusch (2006, Ch. 6) calls “the official road to intersubjectivity” that builds on analyses of factually existing rule-following practices and thus has a problem in making the inference from contingent facts of what *particular* rule-following practices *are* like to modal claims of what *all* rule-following practices *must be* like. By analysing the essential features of the problem (rather than particular practices and examples) I have taken Kusch’s “improved road to intersubjectivity” and hence possible views that challenge my conclusion cannot simply offer a new analysis of contingent practices or disconnected examples; rather, they must answer my philosophical arguments (cf. Kusch 2006, 183).

following problem – or rather, my reconstruction of the strongest possible version of it (III.4.2 & III.4.3) – to provide a constructivist account of the framework of intentional agency and to show how this account not only favours the normative view (III.5.1) but also provides a novel way to understand the foundations of the theory of collective intentionality (III.5.2) and, finally, how the social solution ties together the different themes of this dissertation (III.5.3).

Those who think that my argument is not able to establish the correctness of the social solution beyond all reasonable doubt are welcome to think of this Chapter as providing (i) a novel, positive angle to the debate between causal naturalism and interpretive understanding that was found to be so badly needed by presenting (ii) a strongest possible version of the social solution to the problem of rule-following and (iii) an examination of what implications that solution, if correct, has for the philosophical foundations of the human sciences in general and the theory of collective intentionality in particular. As said, however, I think that by constructing a view that combines the dispositional solution to the Infinity Problem with a social solution to the Normativity Problem the argument provided is as strong a defence of the social solution to the problem of rule-following as one can hope for.

#### III.4.2 THE INDISPENSABILITY OF COLLECTIVE AGENCY

My claim is that the problem of rule-following cannot be solved unless we adopt the anti-individualistic view defended in Parts I and II of this study. Recall that I argued that we must accept we-mode behaviour as a primitive form of behaviour and, moreover, that certain biological (causal) dispositions that are vital for social action are essentially *group-level behaviours* (and group-level adaptations) and thus, in a manner of speaking, dispositions of a group rather than of individuals. In this Chapter I argue that these notions allow us to defend an account of rule-following practices that portrays the practices primarily as social practices such that, first, the practices are *blind* (the Infinity Problem) and, second, the individual applications the practices consists in are *derived* from the group-level framing of the situation (the Normativity Problem).

I mentioned that this kind of view is suggested by Bloor. However, my argumentation moves immediately beyond Bloor's framework, since Bloor (1997, 17) appears to think that the required social practices are based on monitoring, controlling and sanctioning of others. Part II argued against attempts to account for truly social action and practices in terms of sanctions, at least insofar as sanctioning is understood as indi-

vidual-mode action. The relevant practices must be practices of a group in a stronger sense than the sum of individual practices. In short, the practices cannot be based on individual-mode action. A plural subject consisting of agents behaving in the collective we-mode is required.

However, since we are talking about the bedrock practices that make rule-following, and thus meaning and intentionality, possible, the notion of the we-mode applicable here cannot be that of full-blown collective *intentionality*, for that would again beg the question. This is the main reason why I cannot be satisfied with Bloor's account:<sup>157</sup> As Kusch (2004) points out, Bloor (1997) uses intentional language in his characterisations of the constituents of rule-following practices, and thus Bloor is guilty of this kind of circularity.<sup>158</sup> We must be absolutely clear that individuals capable of rule-following (and thus, ultimately, intentionality, meaning and action) possess pre-intentional (or sub-personal) social dispositions to constitute social practices already at the non-intentional (causal) level. Such individuals must have social dispositions to cooperate and harmonise their behaviours with other individuals. And all this must be done *blindly*, *i.e.*, in virtue of pre-intentional dispositions. However, the Normativity Problem prevents us from *identifying* rules with such dispositions. They are necessary but insufficient.

These theoretical considerations are compatible with our empirical understanding of the innate mechanisms that play a crucial role in human language acquisition.<sup>159</sup> Human babies share many inborn dispositions with the great apes such as chimpanzees. For example, both are capable of adopting new patterns of behaviour on the basis of examples. However, empirical studies have established that pre-linguistic infants, unlike chimpanzees, are disposed to react appropriately to purely *co-operative*, communicative gestures, such as simply showing something or pointing something out. Chimpanzees, in contrast, recognize agentive gestures only in a *competitive* setting. Humans, unlike Chimpanzees, appear to be intrinsically disposed to co-operatively *harmonise* their behaviour with others as opposed to merely interacting with others. Accordingly, as long as the interaction involves co-operation and collaboration, human infants, unlike more

---

<sup>157</sup> Indeed, in my view Bloor's theory *cannot* be fully satisfactory, for it is closely connected to Bloor's treatment of intentional explanations as *causal* explanations and Bloor's *reductivist* view of meaning. As I have explained, I cannot accept these claims; normativity must remain in a central role.

<sup>158</sup> The circularity is especially clear in Bloor (2001). Although Kusch (1999 & 2004) is very clear about the importance of avoiding such circularity, it seems that some of his earlier works (especially 1997) may be guilty of it too.

<sup>159</sup> Discussions with Marja-Liisa Kakkuri-Knuuttila, who has applied the social theory of rule-following as defended in Saaristo (2004b) to empirical studies of language acquisition (in her presentation at the annual conference of the Philosophical Society of Finland, January 2006), have been very helpful to me regarding these issues.

economically inclined chimpanzees, are fascinated with games and other interactive situations that are pointless from the individually instrumentalist point of view of achieving a further end. *Offering* things to others for no further purpose is very characteristic of pre-linguistic infants learning to participate in social practices. This is something chimpanzees never do. (Hare & Tomasello 2004, Tomasello & Camaioni 1997, Tomasello 2004, Rakoczy & Tomasello 2006.) The characteristically human, intrinsic inclination towards co-ordination and social harmonisation of behavioural dispositions is, it will turn out, a crucial step from Sellars' pattern governed behaviour to real rule-obeying action.

This sits very well with our understanding of the biological basis of meaningful language. Dunbar (2002), for example, concludes that the exchange of information could not have been the main factor in the evolution of language, since large scale exchange of information is possible only when we already have a well-developed language. Evolution, however, is a gradual process, and benefits to be gained exclusively from full-blown language cannot explain the evolution of less developed forms of language. Hence, argues Dunbar, we have reason to believe that language evolved to construct and maintain social bonds between individuals and to promote group cohesion. If this is the case, no wonder meaningful language is based on social, co-operative dispositions.

Similarly, Snowdon (2002, 209) emphasises that the nature and evolution of language cannot be understood if we do not focus on the social function of language in creating and constructing social bonds between individuals.<sup>160</sup> Further, Sterelny (2003) argues that co-operation (and indeed group selection – recall Part II) has been a crucial factor in the evolution of human cognition. Although I argue (III.5.3) that the tendency of the evolutionists to interpret their studies as confirming the view of the human brain as intrinsically meaningful and capable of representations is heavily confused, I think that as such their results lend strong support to my theoretical arguments.

To introduce this perspective to the present philosophical discussion I follow Barry Barnes' (2000) terminology and say that individuals capable for social rule-following must exhibit *collective agency* (cf. II.2). Barnes acknowledges the Wittgensteinian point that under some interpretation any way of continuing a sequence of examples can be made consistent with the sequence. There is nothing in the sequence itself – or indeed in any explicit formulation of a rule – that would determine the correct

---

<sup>160</sup> Correspondingly, sociologists studying dialogues have argued forcefully that in verbal communication the conveyance of propositional contents is only a small part of the ongoing activity: what is really at stake is the construction of a social bond that promotes co-operation (e.g., Scheff 1990).



extension. However, if one simply does what strikes one as the natural way to proceed, the crucial distinction between correctness and actual performance is lost (III.3).

Thus, Barnes (2000, 54) argues, the extension of a sequence can be said to be correct *only* if the relevant community *agrees tacitly in practice* on which way to continue the sequence counts as correct. The agreement must be collectively sustained such that the members of the collective seek actively to harmonise their respective individual dispositions to extend the sequence. Such a process creates, Barnes thinks, a *shared* sense of what is the right way by ordering the “[d]iverse individual inclinations in the [...] application of rules [...] into a coherent collective practice” (Barnes 2000, 54).

Moreover, Barnes is very explicit on the requirement that the agreement in question, and the process of seeking to harmonise the individual dispositions, cannot be understood as deliberative processes that already presuppose intentional states and actions, for that would beg the question. Thus, says Barnes, “[s]ocial agents are necessary here, agents with a prior non-rational [pre-intentional] inclination toward agreement and coordination, agents who by virtue of this inclination possess collective agency” (Barnes 2000, 56). This leads to Barnes’ notion of the fundamentally *collective agency* (and, of course, also the core of Barnes’ argument against seeing rational choice theory as capable of explaining fundamental sociality – recall I.1.5): Intentional individual agency *presupposes* that agents are highly social in the sense of non-consciously seeking to harmonise their individual dispositions at the pre-intentional level. In this sense the social is prior to the psychological. This, I think, gives exactly what we need without building circularly on intentional notions. The task, then, is to make the notion of collective agency more precise.

An interesting attempt for such an explication is given by Esfeld (1999, 2001), Haugeland (1990), Haukioja (2000), Kusch (1999) and Pettit (1993, 2002). Their central idea is that to explain rule-following we must assume that individuals are equipped with two kinds of dispositions. First, they have the familiar *first-order* disposition to continue a sequence in a certain way. This is argued to solve the Infinity Problem. Second, individuals are assumed to have *second-order* dispositions to monitor the first-order dispositions and to make them match the first-order dispositions of others. These second-order dispositions are understood as exactly the kind of dispositions towards coordination and synchronisation that Barnes talks about.

Kusch, to use his argumentation as an example, thinks that this line of thought can be articulated by stating that the problem of rule-following shows that individuals must be assumed to be equipped with an *imitation device*, a *sanctioning device* and an

*adjustment device* (Kusch 1999, 267-268, see also Esfeld 2001, 81 ff.). The imitation device is a disposition to seek to continue a sequence of examples (to classify things) in the same way as most others do. The sanctioning device is a disposition to (a) sanction those individuals who classify things differently from the collective practice and (b) sanction those individuals who fail to do (a). Finally, the adjustment device is a disposition to adjust the functioning of all three devices in accordance with the received sanctioning from others. Note that these are all blind dispositions; *interpretation* does not enter the picture.

Kusch thinks that individuals who possess both the first-order disposition to continue a sequence blindly and the three devices required for collective synchronisation form a “proto-normative system”:

In such a system, the collective consensus and the process of sanctioning constantly shape and determine one another. The consensus sets the standard for the sanctioning, and the sanctioning protects and recreates the consensus. Neither phenomena can be reduced to the other without distorting the overall process. (Kusch 1999, 270.)

Similarly to Sellars, Kusch concludes that although an isolated individual can engage in activities that may look like concept application (by simply continuing a sequence of examples blindly on the basis of her internal first-order disposition), such a crude dispositional solution can never deliver meaningful classifications (concept applications), since, as we have seen (III.3), an isolated individual has no access to the normative standards of correctness required for meaningful actions. Hence “[c]oncepts are ‘possessed’ primarily by normative systems of individuals, and they are possessed by individuals only in so far as they are parts of such systems” (Kusch 1999, 270).<sup>161</sup> Or, as Winch puts this, “all meaningful behaviour must be social, since it can be meaningful only if governed by rules, and rules presuppose a social setting” (Winch 1958, 116; see also Williams 1999, 147 & 168 and Kripke 1982).

I think Kusch is essentially correct. His proto-normative system is a system that goes on blindly (solving the Infinity Problem), but since it is primarily the *collective* that possesses the rule, the collective sets (proto-)normative constraints on individual

---

<sup>161</sup> In this matter Esfeld (1999, 2001) and Haugeland (1990) agree with Kusch. Haukioja (2000) and Pettit (1993, 2002), on the other hand, defend a version of the second-order dispositional view according to which the second-order disposition does not have to be social – in their view normativity emerges from the tension between the first-order disposition determining the applications and the second-order disposition monitoring the consistency of the applications (cf. Coates 1997), be the second-order disposition social or not. I think this cannot work. Without a standard exterior to the individual we cannot reach a solution that acknowledges the normative nature of meaning and intentionality: without a social aspect the Normativity Problem can never be solved, for in the non-social picture performance and correctness (of the second-order disposition) remain equated.

members of the collective (solving the Normativity Problem). However, the norm consists in the totality or blind functioning of the system. Rules, as Brandom's (1994) slogan has it, are *implicit in social practices*. We must not postulate an *explicit* rule (either a Platonist timeless truth or the generalisation of the naïve communitarian view) that would be in need of interpretation (and would lead to the Regress Argument). Rather, the system *embodies* the norm (Kusch 1999, 268; Haugeland 1990, 405).

Hence to follow a rule is to *participate* in a social *practice*, not to learn to *state* explicitly what exactly the rule says (Williams 1999, 205). What matters is a rule-instantiating practice, not an explicit rule. Of course, when we *have* meanings and language, we can seek to *make* explicit the rules that *are* ultimately implicit in our practices (and move from implicit proto-normativity to explicit normativity). This, however, is a task for linguists, logicians and other theorists; it is not relevant for laypeople or children learning their first language, for they are learning to participate in social practices, not a set of explicit rules.

However, it seems to me that Kusch's (and Haugeland's) way of describing the social aspect of rule-following may be somewhat misleading in the sense that they may appear as sharing Bloor's implicit individualism. As Williams explains, "community agreement is constitutive of practices, and that agreement must be displayed in action" (Williams 1999, 176), and thus "[w]hat Wittgenstein is really emphasising is not even defeasibility so much as our *agreement* as human beings. The very emphasis that commentators have placed on corrective behavior is out of place." (Williams 1999, 175). The problem is that the emphasis on *corrections* (sanctioning) gives still too individualistic (recall II.3.1) a picture of rule-following.<sup>162</sup> We must find a way of understanding social practices so that the practice is indeed primarily social and not a result of individual-mode corrective actions. "The point of learning bedrock practices is to come to *share the same sense of the obvious*" (Williams 1999, 180; my italics), not to learn to avoid sanctions.

Sharing the same sense of the obvious is meant to capture the *blindness* of bedrock practices: bedrock practices can ground rule-following precisely because the social element of sharing the same sense of the obvious that establishes normativity is *direct* and not based on more fundamental individual-mode considerations. To produce normativity, the social must somehow be prior to the individual applications. Although the

---

<sup>162</sup> Moreover, Glüer and Wikforss (2006, 26), for example, argue that the emphasis on sanctions as corrective *actions* – something that can be done correctly or incorrectly – leads to a regress that renders, e.g., Brandom's account of rule-following unacceptable. It is important to notice that the present view does not imply such a regress: for Brandom (1994, 44) rule-following is "norms all the way down", whereas I build explicitly on non-normative dispositions; on behaviour, not on actions.

talk of sanctions seems fit to capture proto-normativity, it is too individualistic for an account of collective agency in Barnes' sense.

### III.4.3 SOCIAL PRACTICES AND IMPLICIT NORMATIVITY

In short, bedrock rule-following practices must correspond to the kind of social behaviour I have analysed in detail in Part II in the context of the group-selection theory and the social identity approach in social psychology (and which was explicitly denied to be *based* on sanctioning but rather to *provide* sanctioning). Thus, although I think Kusch intends his three devices to capture this kind of strongly social behaviour, in my mind the best way to capture the requirements of bedrock rule-following practices is the theory that was used to capture the core elements of strongly social behaviour in the earlier parts of this study, *i.e.*, the theory of collective intentionality.

However, since we are here analysing (preconditions of) original intentionality, the account must be spelled out in strictly non-intentional terms. In the earlier parts, however, I have operated with Searle's (1995) notion where the *content* of an individual application as part of a collective task is deliberately derived from considerations at the collective level, or with Tuomela's (2002, 26) notion of a *reason-based we-intention* where *X* has a reason-based we-intention to participate in a social practice *P* (where *P* is a social action type such as following a rule) iff

- (i) *X* intends *P*
- (ii) *X* believes that everybody in the relevant collection of agents intends *P*
- (iii) *X* believes that there is a mutual belief in the collection that everybody intends *P*
- (iv) (i) at least partly because of (ii) and (iii).

Searle's and Tuomela's accounts work well when we are analysing social institutions in a setting that already presupposes, as it were, the framework of intentional agency. To apply Searle's account for the present purposes, however, we must reject the notion of *consciously deriving* the *meaningful contents* of individual applications from collective-level considerations and replace this with *blind* framing of the situation. Similarly, to apply Tuomela's account the *intentions* and *beliefs* mentioned in the *analyses* must be interpreted in the non-propositional sense in which we attribute intentions and beliefs to non-linguistic animals and machines. Hence the account cannot be reason-based but purely causal, and consequently the connection expressed by (iv) must be seen as a causal connection stating how the second-order social dispositions (ii) and (iii)

monitoring the behaviour of others regulate the first-order disposition (i) to continue a sequence of examples.

This reading of Tuomela gives us the following, essentially social (normativity) and blind (infinity) account of rule-following:

For any individual  $X$  in a community  $C$ ,  $X$ 's performance  $p$  (an event-token) in a situation  $s$  (a unique situation) is rule-governed (by the rule "do  $P$  in  $S$ ") in the proto-normative sense iff

- (i)  $X$  is disposed to treat  $s$  as an instance of a situation-type  $S$ .
- (ii)  $X$  is disposed to perform  $P$  (an event-type) in  $S$ .
- (iii)  $X$  performs  $p$  as an instance of  $P$ .
- (iv) The members of  $C$  share the dispositions (i) and (ii) and go along with (iii) (the members treat – tacit agreement in practice –  $p$  as a token of  $P$ ).
- (v) (i) – (iii) hold at least partly because of (iv) (other causes include  $X$ 's individual, biological first-order dispositions).

As said, the causal mechanism in (v) is to be understood in terms of Barnes' (2000, 56) notion of "non-rational inclination toward agreement and co-ordination" which collective agency consists in and which creates "agreement in practice" (Barnes 2000, 54) and which Kusch seeks to explicate with his three devices. Moreover, the types  $S$  and  $P$  are constituted by the dispositions and thus do not commit the analysis to the existence of Platonist universals or anything similar. With these clarifications, the result is essentially a pre-intentional version of my final explication of Tuomela's notion of *we-intentionality* in 1.2.4.

Further, the dispositions are essentially social in the strong sense that although they are features of individuals, their evolution requires group-level selection and, correspondingly, the dispositions can be seen as instrumentally rational only when rationality is understood as a collective-level concept. These notions were analysed in detail in the context of the evolution of social behaviours and collective intentionality in Part II.

As I explained above, this kind of normativity and rule-following is normativity *in practice*. Rules and norms have their home in what is *done* rather than in what is *said*. Brandom (1994, 22, 100-101 & 206) and Winch (1958, 55-57) illustrate this by appealing to Lewis Carroll's classic article "What the Tortoise Said to Achilles" (Carroll 1895) in which Carroll famously argues that there is no non-circular way to justify the fundamental principles of rationality (the case at hand is the conditional in logic) if we

insist on appealing to explicit rules. This is because, ultimately, the rules are implicit in our practices of performing and accepting inferences. An appeal to an explicit rule – to ground the practice on a fact – would lead to the “insane and intolerable” sceptical conclusion.

Thus, explicit rules depend on tacit norms that are implicit in our practices. This applies to all rule-governed behaviour, be it applying a concept or drawing logical inferences. As Carroll elegantly shows, we can formulate the logical rule of the conditional *because* we already recognise certain inferences as valid in practice and not *vice versa*. In other words, the conditional qua an explicit rule is an explication of a conceptually prior implicit practice of treating certain inferences as acceptable. To think otherwise is to commit the interpretationist mistake and, consequently, fall prey to the Regress Argument. Moreover, the practice must be social, or it fails to resolve the Normativity Problem.

Thus, according to this view the normativity is created by identifying the rule with the practice of the whole collective (a proto-normative system) within which individual applications are assessed. A peculiar implication of this view is that a proto-normative system is such that the distinction between *correct* and *incorrect* applications is available only *for those participating in the practice*, for to participate is to *subject one's applications to the assessment of others*. Looking from within the practice, there is normative rule-following instead of mere regularities of behaviour. Since intentionality, meaning and content presuppose rule-following, individuals can perform actions and have beliefs and intentions precisely to the extent that they participate in social, bedrock rule-following practices embodying implicit proprieties (Williams 1999, 242; Brandom 1994, 159). Or to put this in more fashionable terms, “[a]ll this entails that contents of thought are socially constructed” (Tuomela 2002, 74).<sup>163</sup> This, of course, is methodological holism *par excellence*.

The Wittgensteinian account of normative rule-following holds that the normative distinction between correct and incorrect application exists only relative to social practices and is therefore visible and significant only for those who participate in the bedrock practices. From the point of view of an external, detached observer,<sup>164</sup> our most

---

<sup>163</sup> I am not sure if Tuomela would approve of the way in which I see him agreeing with, among others, Brandom, Kusch, Williams and Winch. The theoretical background in the relevant Chapter 3 in Tuomela (2002) is, however, Sellars (1963) which, together with Wittgenstein (1953), is also the main source of inspiration of, e.g., Brandom (1994).

<sup>164</sup> Of course, in order to form contentful judgements, such an external observer would need to participate in some other social practices. According to the present view, an autonomous, non-social agent is a conceptual impossibility.

fundamental social bedrock practices consist of nothing but workings of interconnected and non-normative causal *dispositions* of non-independent individuals, evolved via group selection and generating regularities of behaviour – but no rules, for rules require participation in the bedrock practice (cf. Esfeld 2001, 88-89).

Needless to say, to aim for anything stronger would require us to derive “ought” from “is” in the strong metaphysical sense, which is not a plausible goal. The normativity in question must be intersubjective, tacit agreement in practice, not an unnaturalistic metaphysical *sui generis* quality. We must be content with a view that explains why certain things appear as normative for those participating in the rule-following practices constituting the human form of life. This, I think, is the idea behind Wittgenstein famous declaration that “[i]f God had looked into our minds he would not have been able to see there whom we were speaking of” (Wittgenstein 1953, p. 217).

In particular, physical brain cells *per se* cannot be about anything nor have propositional *contents* (the rejection of the wonder tissue theory). As such they are simply physical objects governed by non-normative causal laws. They may contain meanings only if such a normative, meaningful *status* is assigned to them in social practices embodying norms. Within social practices they can *count as* meaningful. And, as the Logical Connection Argument holds, the meanings and contents of mental states, actions *etc.* are made possible by normative *proprieties* relative to social bedrock practices. To attribute an intentional state or action to someone is to attribute her a normative status (Brandom 1994, 16-17). Such status receives its content by being rationally (normatively) connected to other statuses that ultimately, as we saw in Part I, consist of collectively accepted and required patterns of behaviour (cf. also Appendix).

However, before I can conclude this section and move on to the actual topic of this Part, intentional action and its explanation, I must address one more influential line of thought. In III.3 I argued that a straightforward, causal dispositionalism cannot solve the rule-following problem, because it runs into the Normativity Problem. However, III.3 simply followed the custom and accepted that Kripke is right in thinking that the Normativity Problem is both a real problem and indeed the hard part of the rule-following problem, and one that straightforward dispositionalism cannot resolve. However, although most writers acknowledge the importance of the Normativity Problem, there are philosophers who think that a dispositionalist needs not resolve the problem, for it is but a pseudo-problem that does not need an answer. They argue that meaning is

not normative in the sense of requiring a solution to the Normativity Problem. Rather than resolving the Normativity Problem, they aim to dismiss it altogether.<sup>165</sup>

Åsa Maria Wikforss, for example, points out that Kripke does not really give arguments as to why meaning must be normative. Rather, Kripke simply assumes this to be case and uses the Normativity Problem “as a pre-theoretical litmus test for other theories: Any theory which fails to allow for the required normativity can be rejected out of hand” (Wikforss 2001, 203). Admittedly III.3 does something similar. Most, if not all, of the philosophers I discuss in this dissertation *accept* that meaning indeed is normative in the sense of the Normativity Problem. Davidson and von Wright, for example, think (III.1) that if the system of intentional attitudes (including actions) one has do not for the most part comply with norms of rationality they simply are not intentional attitudes, for this kind normativity is *constitutive* of them (so described). In the end, in my view Wikforss and others fail to challenge the kind of normativity I have advocated, but explicating their objections – and the answers implied by what I have said above – is nonetheless worth the effort since it helps us to understand what, exactly, the position defended in this dissertation is and how, surprisingly, I am actually in rather large agreement with Wikforss.

The first thing to notice is that the normativity Wikforss and others are mainly targeting is of the kind Wikforss (2001, 203) calls “ought-implying” normativity: it prescribes what one ought to do in a given situation (Papineau (1999) is similarly very clear on this). However, as I have explained, this is not the kind of normativity that is relevant here. The normativity behind the Normativity Argument is that of *constitutive* rules, not *regulative* (although it has to be said that many formulations of Kripke (1982), as Wikforss demonstrates, give the impression of being instances of regulative normativity). As I have repeatedly explained, the rules that are relevant here are comparable to, say, rules of chess: they do not tell what one *ought to do* in certain situations within the game; rather, they constitute the very framework of rules that *is* the game (cf. I.2.1). Only when we have the system of constitutive rules constituting the game (chess, language game), can we seek to formulate regulative rules of the form “if one wishes to do *X* in this game, one ought to do *Y*”.

Part III has emphasised that the Logical Connection Argument talks about constitutive connections (recall that III.1.1 explicitly rejected the regulative reading of the

---

<sup>165</sup> The dismissal of the Normativity Problem is advocated, for example, by Bilgrami (1992), Coates (1986, 1997), Horwich (1995), Glüer & Pagin (1998), Glüer & Wikforss (2006), Pagin (2002), Papineau (1999) and Wikforss (2001). Particularly influential for this line of thought is Bilgrami (1992, 83 ff. in particular), but in what follows I concentrate mainly on Wikforss (2001), for Wikforss both develops further the other criticisms and brings them to bear explicitly on the issues I have discussed in this Part.



normativity in question) and that Kripke's arguments are ontological, not epistemic. Hence, I think Wikforss (2001, 205) is right in arguing that "ought-implying" regulative rules about how to use terms *cannot* show the essential normativity of meaning, for they already presuppose meaning. To use this line of thought to show that meanings are *not* normative is, however, equally futile, for it simply presupposes that meaning is not constitutively normative. Of course Wikforss and others are right in claiming that moral (or even epistemic) norms concerning how one ought to apply meaningful words already presuppose meaning and thus cannot be constitutive for meaning.

However, Wikforss seems to acknowledge this when she writes that "if *meaning* is to be normative, the normativity in question must be semantic in kind, not merely epistemic or moral" (2001, 205). She even conceptualises (Wikforss 2001, 215) this explicitly in terms of Davidsonian constitutive (as opposed to regulative) norms and claims that her arguments challenge the normativity of meaning also in this sense. However, as I show below, her arguments, while fatal for some positions close to the present one, do not challenge the present view. Indeed, in some crucial aspects I fully agree with her. Let me first, however, clear away a couple of objections that obviously have no bearing on the present view.

First, Wikforss says correctly that meanings cannot be normative in the sense that expressing a false judgement would by definition be always a semantic (and not only epistemic) error. Her argument is that the normative view of meaning cannot handle this, if normativity – correctness and incorrectness – is really seen as constitutive of meaning. Notwithstanding the seeming plausibility of Wikforss' argument, I find this surprising, since the social normative view was introduced explicitly to make room for the case where an individual makes an *epistemic* error by calling, *e.g.*, horses cows, but *not semantic*, since she was nonetheless applying the concept cow, albeit mistakenly. As we saw, equating actual use with meaning was a major problem both for the naïve communitarian view and the non-normative, crude dispositionalism, but not for the present view. Again, Wikforss' argument applies to theories that see explicit, regulative rules as constitutive of meaning, not to the present social practice view that treats participation in social practices as constitutive of meaning. Moreover, it seems that Wikforss may well be right in saying that her considerations show that "the normativity of meaning" cannot derive "from the connection between meaning and truth alone" (Wikforss 2001, 207, cf. Horwich 1995). This, however, is not a problem for the present

Kripkean view, which, as we saw, explicitly rejects truth-conditional semantics and builds on assertability conditions.<sup>166</sup>

Second, perhaps my view is not that different from a charitable reading of Wikforss' anti-normativism after all. She argues that "the problem of error arises for theories [...] that construe the relation between meaning and use in such a way that any difference in use implies a difference in meaning." (Wikforss 2001, 208). I agree with this. We need a distinction between actual use and the meaning, *i.e.*, the correct use. And Wikforss is correct in holding that once we have solved the problem of error in this sense, *i.e.*, bridged the gap between true meaning and actual use so that we avoid the problem of error, there is no need for further, metaphysical normativity in the theory of meaning. I have argued that the way to do this is to introduce a social element that grounds this (normative) distinction between actual use and true meaning into the theory. However, I have built the social element out of non-normative dispositions instantiating collective agency. They bring in the normativity required for solving the Normativity Problem but, crucially, do not require further normativity. The view I defend, *contra* some formulations by Kripke, is a form of second-order social dispositionalism. Like Wikforss, I have explicitly refused to see anything more normative in meaning. Indeed, I have argued that to do so would be to turn to unacceptable metaphysics. We need a theory of social practices that builds purely on naturalistic dispositions but that nonetheless has room for the normative aspects (external and objective from the point of view of those participating in the practice) of linguistic practices: In the technical parlance of Part I, I have no need for *independent* normativity.<sup>167</sup>

Thus, if the core of non-normative dispositionalism is, as Wikforss (2001, 209) says, that meanings are fully determined by actual use, and actual use can be accounted for in terms of causal dispositions, I have no quarrel with this, for it is true also of my view.<sup>168</sup> The only difference is that Wikforss (and, *e.g.*, Coates 1997) does not require the dispositions in question to have complex social and hierarchical structure, whereas,

<sup>166</sup> Wikforss (2001, 207) seems to counter this move by saying that one can of course accept pragmatist semantics, where normativity indeed is constitutive of meaning, but then the normativity thesis stands or falls with pragmatist semantics. In a sense I am happy with this, because I do accept the pragmatist view in this sense. Indeed, Wikforss concludes that "[w]hat we need to do, of course, is distinguish between what a word is true of and what it is not true of, but this is just the old problem of accounting for reference and has nothing to do with norms" (Wikforss 2001, 207). This however, is just a statement and remains unjustified in Wikforss' papers. Indeed, Appendix argues that the normative nature of meaning is precisely what is required to make room for an acceptable and naturalistic (rejection of both wonder tissue and objective idealism) solution to the "old problem".

<sup>167</sup> Kusch (2006, *e.g.*, 66) makes a similar point by explaining that the Kripkean view replaces "semantic normativity with intersubjective normativity".

<sup>168</sup> Indeed, many anti-normativists criticise the normativity argument primarily as part of a view that builds normativity – and semantics – on the notion of agents *intending* to use expressions in certain ways. Such psychologism is something I have explicitly rejected as question begging.

as I have shown, for me the social aspect and the structure of first and second-order dispositions do a considerable amount of philosophical work in contexts that do not fall into the scope of Wikforss' (2001) interests. Moreover, also Wikforss (2001, Footnote 29, p. 222 & 212) accepts explicitly that to counter further problems regarding linguistic practices (such as those discussed in Appendix) we may have to go for a social theory of meaning, but that this does not challenge forms of "non-normative pure use theories" that build on social dispositions.<sup>169</sup> Again I agree, for my theory is a socialised version of dispositionalism: as I have all the time said, in the name of ontological naturalism I am not prepared to accept anything but non-normative, physical, causal features to the fundamental furniture of the world. In particular, I do not want to assume any practice-independent norms or contents (cf. Appendix).

Further, when Wikforss (2001, 214 ff.) argues that the Davidsonian conviction, according to which the rules of rationality are constitutive of the meaningful realm, cannot ground the idea that meanings are normative, it is very important to notice that Wikforss' arguments are explicitly such that they do not challenge my argumentation in III.1: Wikforss is not attacking the animating idea behind the Davidsonian anomalism of the mental (and thus the Logical Connection Argument). Rather, her goal is to show that "[i]t is quite possible to stick to the Davidsonian view of intentionality and yet deny that meaning is normative" (Wikforss 2001, 215). Wikforss emphasises that the Davidsonian view is that unless one uses terms rationally one's uses are actually meaningless – not that if one means something with a term, then one ought to use the terms such and such a way (Wikforss 2001, 215). As said, I agree with this wholeheartedly. If the relevant rules were regulative in nature, they would already presuppose meanings.

Thus, Davidsonian normativity boils down to the constitutive nature of rationality constraints (and not to the regulative ought-normativity). In Wikforss' view this shows that the Davidsonian view does not imply that meaning includes a normative element that cannot be captured by "pure use", *i.e.*, dispositional, accounts, because Wikforss' concern is explicitly "exclusively with normativity in the sense that implicates an 'ought', a prescription" (Wikforss 2001, 203), *i.e.*, regulative normativity, which she also explicitly opposes with constitutive normativity. Indeed, Wikforss (2001, 218) endorses enthusiastically the idea that the link between use and meaning is constitutive, which, of course, is a view I have as earnestly defended in this dissertation.

---

<sup>169</sup> Here Wikforss parts company with Bilgrami (1992), who resists social theories of meaning categorically. But the sociality Bilgrami has in mind is primarily of Burge's type (see, *e.g.*, Burge 1979). However, discussion on Burge's social theory of language goes beyond the scope of this dissertation, and thus I will not address Bilgrami's anti-socialism here.

Hence, here we have the feature that allows the view defended in this dissertation to prevail over Wikforss' challenge and, in a sense, it to largely agree with Wikforss. After all, like Wikforss, I want to emphasise very strongly that in my view all the building blocks of rule-following practices must be non-normative causal dispositions; there is no room for normativity that is not constructible by using dispositional elements, *i.e.*, by a "pure use theory" in Wikforss' terminology that she contrasts with normative views. To think otherwise would be circular. Indeed, in this respect my view was explicitly argued to differ crucially from, *e.g.*, McDowell's quietism – which Wikforss (2001, 214) sees as a paradigmatic example of the kind of normative theory she resists – where an unnatural normative element remains unaccounted for (cf. also my further criticisms of McDowell in Appendix that take up precisely this point). My second-order social dispositionalism is meant to develop further naturalistic pure use theories so that they can account for the apparent normativity of linguistic practices and resolve the Normativity Problem. This is my anti-causal humanism *within* causal naturalism.

Thus, to conclude this Chapter we can say that the meaning of an action amounts to the role the intentional description of the action plays in our social practices of giving and asking for reasons. Actions are what we do, what we are *responsible* for. Causation matters in human activities for sure, but to capture a behaviour qua meaningful – qua action – we must not concentrate on the causal history of the behaviour but on whether our practices *authorise* that kind of behaviour in the particular situation and what further commitments the performance brings with it. Thus, we finally have an answer to the question of what makes behaviour action. An action is a behaviour that is assigned a certain (normative) status in our social practices.

With this answer we are able to return to the problem concerning the nature of intentional explanations of actions.

## CHAPTER III.5:

## EXPLANATION OF ACTION AND COLLECTIVE INTENTIONALITY

## III.5.1 THE TRUE FORM OF INTENTIONAL EXPLANATIONS

Behaviour is an action to the extent that reasons can be given and asked for it. This requires that the behaviour is redescribed in a way which embeds the behaviour (via the redescription) into the normative web of practical reasoning consisting of collectively acknowledged entitlements, requirements and commitments, just as von Wright (1971) says. However, as such von Wright's Logical Connection Argument remains silent about *why* meaningful actions must be seen as essentially normative in a sense that purely causal theories cannot capture (even the no echo argument of III.1.3 simply argues that *if* – or rather: *since* – actions are essentially normative, a causal theory cannot be satisfactory), *where* the normative web comes from and indeed *how* it is possible to *follow* normative requirements (rules). Answers to these questions are what I have attempted to provide with my social account of rule-following.

In particular, I have argued that the norms governing practical reasoning cannot be explicit rules in Plato's heaven, in the language of thought or what have you. In contrast, I have defended the view that they are implicit in our social practices. Thus, when we look at the standard form of intentional explanation,

1. *X* desires *D*

2. *X* believes that *A* is the best means to attain *D* under the circumstances

(*L*) If any agent, *X*, desires *D*, and believes that doing *A* is the best means to attain *D*  
under the circumstances, then *X* does *A*

*Ergo*, *X* does *A*,

we should realise that the explanation has this form because this is how rational agents are collectively required to arrange their desires, beliefs and actions in order to count as rational agents that are capable of desires, beliefs and actions. In the Wittgensteinian picture, (*L*) is an attempt to make explicit a norm that resides implicitly in our practices of giving and asking for reasons.

Similarly, the rule-following considerations show that the form of the argument in general and (*L*) in particular cannot be based on a timeless Platonist idea or Kantian universal principle in need of interpretation. But this means, *pace* Davidson, that the

belief-desire model of intentional action is not universally necessary for rational, intentional action. Rather, it is *commitments*, which in our practices are connected to other commitments, rights and duties, that are essential for intentional action, and these notions correspond naturally to *beliefs* and *intentions* – in my mind there is not always need for a *desire* or *pro-attitude*<sup>170</sup> (recall my analyses of collective intentionality and we-mode action in Part I that operated solely with beliefs and intentions and, crucially, not with desires).

As was explained, in the Wittgensteinian social practice view the form or explicit model of an acceptable inference (such as the belief-desire model) cannot be conceptually prior to accepting certain inferences as valid in practice. The rule-following considerations show that we cannot justify our rule-governed practices by appealing to explicit rules. Rather, formulations of explicit rules or models of practical inference are attempts to make the essential components of an implicit, blind practice explicit. In short, the social practice view I defend treats as an acceptable form of practical inference any form that is (tacitly) accepted as valid in social practices.<sup>171</sup>

Crucially, there are familiar cases of practical reasoning that do not operate in terms of the Davidsonian belief-desire pair. Brandom (2000, 87) uses the following example:

- (a) It is raining.
- (b) I shall open my umbrella.

According to the orthodox Humean, or mainstream Davidsonian, tradition in action theory we ought to insist that such an inference is crucially incomplete: we must assume that at least the following premises are implicitly present:

- (c) I desire to stay dry.
- (d) I believe that it is raining.
- (e) I believe that if it is raining I must open my umbrella in order to stay dry.

However, when we understand that the form of intentional explanation is determined by our own (collective) practice of treating certain forms as acceptable in prac-

---

<sup>170</sup> A closely related reading of intentional explanations is defended by Brandom (in particular 1994, 245 ff. & 2000, Chapter 2). I would also like to thank Brandom for discussing my criticism of Davidson with me at his research seminar “Concepts and Contents” at the University of Tampere (Finland) in May 2004.

<sup>171</sup> Again, this is not as relativistic a claim as it may sound. Collective acceptance is always open to criticism and corrections – see Appendix.

tice, it becomes clear that the standard insistence of adding the extra premises (c)-(e) is justified only if it *is* our practice to require them to be added. But, typically, we do not in fact (indeed in *practice*) require such additions. This, I think, is also in Anscombe's (1959) view the core of *practical* reasoning. In our practices the fact that it is raining is a perfectly good reason for opening one's umbrella. Only the most dedicated Humean/Davidsonian action theorists among us will not accept "it is raining" as full answer to the question concerning my reasons for opening my umbrella when I leave the philosophy department and face the streets of London.

A related, important point is also that if we think, as the Standard View holds, that the intentional explanation picks out a natural chain of causes and effects, then we face the problem of an indispensable *ceteris paribus* clause: besides adding (c)-(e), we must require that I do not have any desire overriding (c), that I know how to do (b) and so on. It seems that no matter how many such premises we add to the inference, a clever enough philosopher could come up with a counterexample. In other words, (b) must include an indispensable *ceteris paribus* clause.

The indispensability of the *ceteris paribus* clause is a serious problem for the Standard View of seeing the inference as a non-normative, causal process. This indispensability is, of course, simply another way of pointing out the futility of Elster's project of trying to find necessary and sufficient *causal* conditions for a behaviour to count as an action (III.1.3) and, thus, of expressing Davidson's (1973a & 1974a) post-1963 acknowledgment that the premises of a practical inference an intentional explanation of action consists in cannot give "sufficient conditions of intentional (free) action" (Davidson 1980, xvii). It seems that we can complete the inference only by *deciding* not to take into account any further complicating possibilities. The *ceteris paribus* clause is indispensable precisely because there *is no other fact of the matter* as to when further qualifications are no longer needed than the social practice of (tacitly) *accepting* certain reasons as sufficient. This, of course, is perfectly compatible with the normative, collective acceptance view of practical reasoning I defend.

This point takes us deep into the philosophy of action. We have already seen that in the case of the theoretical reasoning of Carroll's (1895) Tortoise and Achilles, to state the conditional is "to make explicit [...] what before was implicit in our practice of distinguishing some inferences as good" (Brandom 2000, 81), not to justify the practice. Thus, as Carroll saw, we cannot use the explicit rule of the conditional to justify our inferential practices, since the implicitly normative inferential practices are *prior* to the

explicit norm. The explicit rule does not function as a premise in our fundamental inferential practices which, after all, are *blind*.

Similarly, expressions of desires should not be always required to feature in practical inferences, since, just like the expressions of the conditional in the context of theoretical reasoning, such expressions function ultimately not “as a *premise*, but as making explicit the *inferential* commitment that permits the transition” (Brandom 2000, 89) from, for example, “it is raining” to “I shall open my umbrella”. In Brandom’s terms, “normative vocabulary (including expressions of preference) makes explicit the endorsement [...] of *material* proprieties of *practical* reasoning” (Brandom 2000, 89; boldface omitted). As the Logical Connection Argument states, my desire to stay dry (qua a propositional attitude) partly consists in and receives its conceptual content from my commitment to the inferential rule (or practice) taking me from “it is raining” to “I shall open my umbrella”.

Thus, the Humean belief-desire model of practical reasoning and intentional action that builds on instrumental rationality is parasitic upon the more fundamental Kantian model in which to act intentionally is to act in accordance with a norm. However, whereas for Kant such norms were explicit rules, I have argued that the norms must fundamentally be implicit in our practices of giving and asking for reasons. Indeed, the Humean model, where it works, derives its plausibility from its ability to make explicit an implicit practice. This, of course, is very different from the actual self-understanding of the Humeans, for they think that a desire as an entity logically distinct from beliefs and actions is required to causally initiate the action.

My reading of the Logical Connection Argument on the one hand and the Wittgensteinian social practice view on the other rejects this view as a twofold mistake. First, desires, beliefs and actions qua contentful attitudes are essentially tied to (and constituted by) a normative web of commitments, entitlements, acknowledgements etc., and hence it makes no sense to talk about desires as something wholly distinct from beliefs, intentions and actions. Second, intentional action and intentional explanation are not primarily causal issues, and hence there is no need for a desire to set the action in motion in a causal sense. An intentional explanation explicates why a certain action was *appropriate* in light of the commitments of the agent by subsuming the action (via a rationalising redescription) into a normative web. The point is not to show what triggered certain movements.

This view is not that far from, for example, Dennett’s views on intentional action and its explanation. After all, in his view intentionality ought to be understood in



terms of mutual acceptance of the intentional stance and the corresponding mutual ascriptions of intentionality. Dennett (2003, 251) even expresses this explicitly by arguing that the notions of intentionality, agency, rationality, free action and the like are established in the Sellarsian practice of giving and asking for reasons. This idea of seeing intentionality and the framework of agency as something that is *bootstrapped* into existence within social practices (Dennett 2003, 259 ff.) is obviously very similar to the picture I have been painting in this Part of my study. However, as for example Searle is very keen to point out, the Dennettian ascriptions of intentionality and indeed the adoption of the intentional stance appear already to presuppose the existence of more fundamental original intentionality, for *attributions* of intentionality surely require an intentional agent to perform the attributions.

I think that in a sense this criticism is correct. Dennett's account is not fully satisfactory in this respect. However, the alternative favoured by Searle, namely that certain brain states just are intrinsically intentional (ridiculed as the wonder tissue theory by Dennett), does not do any better when faced with the problem of rule-following. Actually, in this duel of two contemporary philosophical titans I think it is Dennett who is closer to an acceptable solution. It is very important to see that the Searlean accusation of circularity has a hold only if the ascriptions of intentionality that play such a crucial role in Dennett's theory are understood as *explicit* assignments (ascriptions made explicitly by individual agents). In contrast, if attributions of intentionality are made, as I have argued above, in terms of proprieties *implicit* in blind social practices, the circularity accusation fails (cf. Brandom 1994, 147).

Indeed, the social practice view seems to be the view taken by Dennett (2003) when he suggests that the original intentionality does not belong to any agent but to the social game of giving and asking for reasons which is, ultimately, constructed and maintained by the evolved tendencies (causal dispositions) of *Homo Sapiens*. This is the view also the present study subscribes to. Moreover, I think that Dennett's (2003) way of locating intentional free agency within the *normative* game of giving and asking for reasons, as opposed to the *causal* order of things, is undeniably the most promising way of defending free agency even if causal determinism is true (III.2, Saaristo 2004a).

However, more often than not Dennett seems to think that we are free to adopt either the intentional or the physical (causalist) stance, since in his view the distinction is not grounded in the objective features of the world; all that matters is the instrumental value of the chosen stance in predicting behaviour. I, on the other hand, think that as agents we are tied to the intentional perspective. I am also convinced that the theory of

intentionality I defend can capture intentionality as an *external* and *objective* (in the technical sense of Part I, *i.e.*, as nonetheless *dependent* on social practices) feature of the world and not as a mere instrumentally useful tool.<sup>172</sup> Contrary to some formulations by Dennett, I think that ascriptions of intentionality are appropriate according to their accuracy in explicating objective proprieties, not only according to their predictive utility (cf. Brandom 1994, 56-57).

### III.5.2 THE STATUS OF COLLECTIVE INTENTIONALITY

Let me next move on to discuss the case of full-blown collective intentionality. The crucial point is that the form of intentional explanation (such as the belief-desire model) is not forced upon us by some causal laws of nature or eternal ideas in Plato's heaven. The standard form of intentional explanation in terms of (1), (2) and the conclusion has the form of an acceptable practical inference because it instantiates (*L*), which in turn is an explication of a norm or propriety implicit in our practices.

Similarly, the inference from "it is raining" to "I shall open my umbrella" is acceptable because it sits well with our practices, which implicitly include a norm that could be made explicit as "if it is raining, one is rationally entitled (or even required) to open one's umbrella". Thus, no further justifications are required. In contrast, if I give the statement "it is raining" as the answer when I am asked why I do *not* open my umbrella, my answer typically is not accepted. I am not rationally entitled to such an inference. To get the others to acknowledge my entitlement to the practical inference from "it is raining" to "I shall not open my umbrella", I can express, to use Brandom's (2000, 87) example, my Gene Kelly desire to sing and dance in the rain, which *explicates* my *practical commitment* to the inference from "it is raining" to "I shall not open my umbrella".

In sum, according to the present social practice view, whether or not a practical inference is complete depends neither on any independent fact about the correct Platonist form of practical inferences nor a causal process, but simply on whether we in practice require the inference to be completed or not. Bearing this in mind, let us next consider a situation where an agent faces a situation she recognises to be a social dilemma situation where mutual co-operation is the collectively optimal (best for *us*) thing to do. Note that this is an issue I have discussed at length in Parts I and II. I argued there that

---

<sup>172</sup> This is one reason why I argued (III.4.1) that the Kripkean social practice view I defend does not contradict Bloor's (1997) claim that a *social* solution to the problem of rule-following can be fact-based in the sense of offering an *objective* ground for meaning and intentionality.

we should expect humans to have the tendency to adopt the we-mode, *i.e.*, to frame such situations as collective tasks where the individual action-intention or role is *derived* from the collective-level optimality considerations.

Thus, in the present context the question of the status of collective intentionality and collective we-mode action can be formulated as the question of what we ought to say concerning the following practical inference:

- (a) In this (social dilemma) situation co-operation is collectively optimal.
- (b) I shall co-operate.

In Part II, I argued that this inference captures the intentional action of we-mode agents capable for collective intentionality. The standard individualistic view, on the other hand, insists that the inference must be completed by adding the following premises:

- (c) I desire to realise collectively optimal outcomes.
- (d) I believe that in this situation I must co-operate in order to realise the collectively optimal outcome.

However, the mainstream obsession to treat the inference from (a) to (b) as incomplete is nothing but the tired *a priori* insistence on the priority of individual-mode notions and individual-mode psychology (II.3). In Part II it was argued that we have strong evolutionary reasons to assume that the anti-individualistic picture captured by the inference from (a) to (b) captures a *sui generis* form of reasoning which does not require individualisation in terms of (c) and (e).<sup>173</sup> In this Part we have seen that indeed there is nothing in the world that would force us to treat the inference from (a) to (b) as crucially incomplete. It is up to us to establish it as an acceptable form of intentional explanation in the game of giving and asking for reasons by *treating* it as such in practice. For example, when asked why I co-operated in a social dilemma situation I may simply answer that it is the right thing to do, explicating thus the norm that resides implicitly in our practices.<sup>174</sup> Thus, although co-operation is irrational from the point of

---

<sup>173</sup> I wanted Part II to be neutral concerning the explanation versus understanding debate and the nature of original intentionality, and thus it was left open whether the evolutionary process in question amounts to biological or cultural evolution. On the basis of this Part, however, it is clear that the process must be largely that of cultural evolution, for intentionality is an irreducibly social phenomenon.

<sup>174</sup> One of the most stable empirical findings of social identity theorists (*e.g.*, Kerr & Park 2001, 118) is indeed that people conceptualise the tension between individual-mode rationality (leading to defection) and collective considerations (requiring co-operation) in the obviously normative terms of good and bad.

view of individualistic benefit maximisation, co-operation is rational in the primary, Kantian sense of corresponding with the fundamental norms (implicit in our practices) that constitute intentionality and rationality.

Moreover, the evolutionary arguments of Part II demonstrate that communities that treat the inference as an elementary form of practical reasoning are likely to have flourished, for the practice of accepting this form of inference is a social trait that has the capacity to beat the Davidsonian strategy of insisting on the primacy of individual-mode notions in multi-level selection processes. Similarly, Brian Skyrms (1996) argues forcefully that we should expect most societies to have evolved to include an implicit norm or practice that can be made explicit precisely in terms of accepting the inference from (a) to (b).

We have also seen that an individualist cannot appeal to the idea that individual-mode intentionality would be more naturally paired with our pre-intentional, purely causal (biological) dispositions. First of all, intentional notions have no echo in the realm of causation, and hence the relevance of this line of thought is highly questionable. Nonetheless, the individualist could insist that even if there are no systematic connections between the normative (the intentional) and non-normative, the normative must eventually be constructed by the workings of our biological (blind) dispositions. After all, we have seen that normativity resides implicitly in our blind practices. However, it should by now be also clear that this line of thought speaks strongly *against* individualism and the priority of individual-mode notions.

In Part II we saw that the pre-theoretical understanding of evolution as nothing but a competition between individuals is not warranted in light of proper understanding of evolutionary dynamics. Group-level selection processes that promote co-operation between individuals are most likely largely responsible for the evolution of many of our social traits. And we saw that the products of group selection, when described in intentional terms, are highly compatible with and supportive of the theory of collective intentionality. Moreover, in this Part we have seen that the biological dispositions that make the framework of intentional agency possible must be essentially social dispositions structurally similar to the picture of intentionality implied by the theory of collective intentionality.

Thus it can be concluded that collective we-mode intentionality is real to the extent that our social practices make it so. This result may sound somewhat disappointing, since in the other parts of this study I have argued that we-mode collective intentionality is an objective, *sui generis* phenomenon, and now I am admitting that in fact it exists

only relative to our practices which bootstrap it into existence. However, it should be clear that such dissatisfaction is unwarranted.

The fact that collective intentionality exists only relative to our social practices does not make it any less an objective phenomenon than the fact that the pieces of paper in my pocket are money (Part I). Similarly, both facts – that the pieces of paper in my pocket are money and that collective intentionality is a *sui generis* form of psychology – are in an important sense *self-referential* (cf. Barnes 1983, Kusch 1999, Searle 1995, Tuomela 1995, 2002): Both facts are constituted by our (tacit) collective acceptance to treat them as true in practice, and in both cases this practice boils down to practical acceptance of certain inferences and patterns of behaviour. To treat either of these facts as a truly practice-independent fact would be metaphysically highly questionable.

Moreover, collective intentionality is based on our implicitly normative practices such that collective intentionality is a *sui generis* form of intentionality *both* in the sense that it is not based on a more fundamental individual intentionality *and* in the sense that it resides, as does all intentionality, in the Sellarsian logical space of normative reasons and not in the logical space of non-normative causation and the natural sciences. Hence, as Davidson (1974a, 230) puts it, short of changing the subject it cannot be reduced to anything non-intentional. All intentionality is practice-dependent in this sense. Apart from accepting Cartesian dualism (or indeed full-blown objective idealism with objective spirits or collective consciousnesses), we cannot require intentionality – collective or individual – to have a more independent status than the one assigned to it by the social practice theory.

### III.5.3 TYING UP LOOSE ENDS

Finally, the position developed and defended in this Part allows me to sharpen some arguments that were left somewhat open in Parts I and II. The first point I wish to make is a clarification of the distinction I drew between a subjective and objective characterisation of collective intentionality when discussing Tuomela's (2000) definition of a *reason-based* notion of collective intentionality in I.2.4.

Recall that Tuomela wrote his definition in the *objective* form, *i.e.*, as saying that the *fact* that also others have the relevant we-intention (and the *fact* that there is a mutual belief about this in the collective) is the reason why *X* too adopts the we-intention in question. In my own formulations, however, I preferred the *subjective* version of Tuomela's definitions. The subjective version states that the reason why *X* adopts the

we-intention is that she *believes* (at least tacitly) that the facts in question obtain. The reason for preferring the subjective version was that I was after the perspective of the agent *X* and an intentional explanation of her adoption of the we-mode.<sup>175</sup> Now when we have explicated the ontological nature of intentionality, it is easy to see that Tuomela's original objective formulation is in a sense ontologically more fundamental.

The objective formulation corresponds roughly to the inference from (a) to (b) above, which was seen to capture the essence of collective intentionality. After all, I did not require (a) to be turned in to the subjective form of "I believe that (a)". Hence, to the extent that co-operation is normatively required in social dilemma situations in the relevant community, *i.e.*, insofar as co-operation in social dilemma situations is a normative (perhaps implicit) practice of the community, the objective definition of *reason*-based collective intentionality is literally correct. The fact that the situation at hand is a social dilemma situation is a good (collectively accepted) *reason* for co-operation, for "this is a social dilemma situation" is a good answer (an acceptable rationalisation) to the question concerning one's reason for co-operative action. Co-operation is what is collectively required of agents in social dilemma situations in most everyday social practices. The *objective* definition of reason-based collective intentionality explicates collective *action with a reason* in social dilemma situations.

However, also the subjective formulation captures an important aspect here. Just as (L) explicates the implicitly normative practices that tie beliefs and desires to actions in the individualist case, expressing a belief that the setting at hand is a social dilemma situation (and the belief that the others treat the situation as a collective task) amounts to *undertaking* a commitment with the kind of propositional content that, given the implicitly normative practices of the collective in question, commits the undertaker also to co-operation. The *subjective* definition of reason-based collective intentionality explicates collective *action for a reason* in social dilemma situations.

In other words, the objective practice is what *justifies* the inference formulated in subjective terms. The *implicitly* normative practice *explicated* by the objective definition is what *entitles* the agent to her subjective inference. Moreover, this interplay between the subjective and the objective perspective is important not only for the proper understanding of intentional action, but also for seeing the nature of the overall picture that emerges from my social practice view (Appendix).

---

<sup>175</sup> I mentioned in Part I that Tuomela (2000, 52) acknowledges this point, but simply says that the "subjective channelling" is assumed to be built into the reason-relation. I insisted on keeping the distinction explicit, but I was not able to justify my insistence satisfactorily. Now, at last, we have enough understanding of the metaphysics of intentionality to fully appreciate the importance of the explicit distinction.

The second theme that was not brought into its proper conclusion when the issue was on the table for the first time is the connection between my theory of collective intentionality and Hacking's (e.g., 1999) notion of the *looping effect* of human kinds. I explained (II.3) how Hacking thinks that the categories of the human sciences are *interactive*, since "people [...] can become aware of how they are classified and modify their behavior accordingly" (Hacking 1999, 32).

Hacking's motivation comes from the Anscombe-Davidson view of intentional action, where the core idea is Anscombe's (1959) insight that actions are essentially actions "under a description". In Anscombe's view, for one's behaviour to be an intentional action there must be a description such that one intended to act under that description (III.1.2). Thus, actions are largely conceptual, for conceptual descriptions are (partly) constitutive of actions qua actions. Hacking's (1995a, 235) core idea is that from Anscombe's insight it follows directly that one can intentionally perform only the kinds of actions one can contentfully describe<sup>176</sup> and, consequently, "[w]hen new descriptions become available [...], then there are new things to choose to do" (Hacking 1995a, 236).

This is an important point, but in order to get to the looping effect Hacking moves quickly even further. It is not only that conceptual constructions are required for intentional actions. Rather, conceptual constructions are constitutive also of the very possibility of being a person in the sense of an agent, and thus conceptual developments can also bring in new ways to be an agent. Hacking's (1995a) interesting case study is the multiple personality disorder (cf. also Rovane 1998, 169 ff.). Hacking argues that multiple personality is undeniably both a real psychological condition – some people really do have multiple personalities – and (at least partly) brought about by conceptual developments in psychological discourse: In the 1970s multiple personality was "a mere curiosity" and one "could list every multiple personality recorded in the history of Western medicine" (Hacking 1995a, 8), but after multiple personality had become an official diagnosis of the American Psychiatric Association in 1980, cases of multiple personality disorder were found in their thousands (Hacking 1995a, 8). However, according to Hacking, it is not the case that before the year 1980 there were a lot of undiagnosed multiple personalities around. Rather, suddenly in the North America of the 1980s there was a huge outbreak of multiple personality disorder.

Hacking explains that the construction of a contentful, relatively well-defined concept of multiple personality disorder has "provided a new way to be an unhappy per-

---

<sup>176</sup> All we can *do* is to apply concepts, as Brandom's slogan puts this.

son [...] multiple personality [...] has become, to use one popular phrasing, a culturally sanctioned way of expressing distress” (Hacking 1995a, 236). Roughly put, the construction of the theoretical concept of multiple personality disorder has made it possible that there can really be several persons in one body. The construction of the concept has brought a new form of agency into being. And, crucially, the new way is not less real than other ways of being an agent, since this is what human agency consists in. *All* forms of agency are based on collectively constructed and maintained ways of being an agent.<sup>177</sup>

On the basis of this Part of my study we can see that this is a profound point indeed. Our *essence* as intentional agents is based on collectively maintained *proprieties* governing appropriate inferences (including practical inferences concluding in actions). These proprieties are implicit in our practices and they constitute the normative framework within which we can meaningfully see individuals as agents and persons, for to be an intentional agent is to be able to perform actions, *i.e.*, behaviours that one is responsible for (and which reasons can be given and asked for). The construction of the theory of multiple personality disorder has provided new acceptable ways of rationalising one’s behaviour, *i.e.*, for constituting the behaviour as an action and the actor as an agent.

This reasoning suggests that there is no *metaphysical self* in the sense of a Cartesian ego behind the social practices constituting the human form of life. Rather, persons are constituted within such practices. Practices are *prior to* and *constitutive of* agents and personhood. Pre-social (or asocial) individuals are not agents unaffected by social practices; such individuals are not agents at all but only physical objects governed by causal laws. To be an agent (or a person) requires that one is situated within the logical space of normative reasons governed by normative rules of rationality, not by causal laws of nature. Personhood is a social status, not a causal property (Saaristo 2004a).

To be an agent is to participate in social bedrock practices that constitute (the forms of) agency. Thus, when Hacking (1986) says that in our practices we “make up people”, this should really be taken literally to the extent that by “people” we mean people qua intentional agents and persons and not qua physical objects. “My claim is that we ‘make up people’ in a stronger sense than we ‘make up’ the world” (Hacking 1984, 40). What Hacking means, I think, is that we can causally modify the physical world that exists independently of our practices, but in our practices we non-causally

---

<sup>177</sup> Besides Anscombe and Davidson, in the background of Hacking’s arguments lurks the towering figure of Michel Foucault and his arguments on how new ways of being a person can be constructed in (scientific) discourse.



constitute the very possibility and form of personhood – people qua agents do not exist independently of our practices, and thus there is no original form of personhood (Cartesian ego) independent of our practices.<sup>178</sup> I read Hacking's claim that we make up people as closely related to the central thesis of Kusch (1999), namely that (folk) psychology is a social institution.

Where Hacking concentrates on the construction of *contents* of actions (new kinds of action) and on the construction of types of *agents* (e.g., agents with multiple personalities – perhaps also collectives as agents), I have focused on the *mode* in which actions are performed, regardless of the content and performer of the action. I have shown that just as the possible contents and indeed types of agents are relative to our practices, also the mode of our actions is dependent on what forms of practical inferences we treat as acceptable in our practices. In this sense it is up to us to *make* the theory of collective intentionality true or false.

Consequently, what kind of inferential proprieties we in fact reproduce in our practices has significant effects on the way the society works. Although intentional explanations are not causal explanations, our practices are such that individuals are guided to modify their behavioural dispositions to reflect normative proprieties (cf. Brandom 1994, 260 – recall also Kusch's three devices and my analysis of collective agency). This is precisely the factor that allows us to use largely prescriptive, intentional explanations also as descriptive generalisations in the sense of Millian empirical laws (III.2.4).

Thus, if our theoretical naivety leads us to disregard collective notions and insist exclusively on individual-mode accounts of agency and rationality, this is not a case of mere misrepresentation of actual practices. Rather, such one-sidedness may lead us to modify our basic normative attitudes, *i.e.*, what we expect and require of rational agents, what kind of behaviour is accepted as rational action, which kinds of rationalisations are allowed and so on. In the worst case such individualism may contaminate even our basic practices, and since such practices constitute forms of intentionality, our theoretical immaturity may turn into a self-fulfilling prophecy. Attributions of intentionality do not only *describe* but also *prescribe*.

This is also the fundamental reason why the empirical approaches reviewed in II.3 failed to unambiguously settle the debate between methodological individualism (exemplified there by the discovered preference hypothesis in experimental economics)

---

<sup>178</sup> Hacking (see especially 1995b, 362 & 364 ff.) actually resists this kind of reading – not because he thinks it is wrong, but because he does not want to discuss what he calls “deep” issues such as constructivism or the distinction between hermeneutic understanding and causal explanation.

and holism in the sense of the theory of collective intentionality (represented by the social identity approach in social psychology). Recall that we were forced to admit that in a sense individualism and holism simply construct different conceptual frameworks in terms of which actions can be conceptualised, and empirical evidence did not unambiguously determine which way is correct and which is not. Now we can see clearly why this indeed must be the case: Descriptions are (partly) constitutive of actions.

As the Logical Connection Argument and Davidson's philosophy of mind explain, we turn a bodily behaviour into an action by embedding it into a conceptual framework via an intentional redescription. Thus, there are no description-independent actions that could set the truth conditions for our conceptualisations of actions. In other words, there is no independent fact of the matter as to what is the true form of intentional actions. This is the by now familiar replacement of truth-conditional semantics with inferential role semantics based on socially established assertability conditions when it comes to assignments of intentionality. By accepting certain types of explanations as valid we constitute the social fact (practice) that indeed makes them valid (cf. C. Taylor 1985). This is the ultimate core of Hacking's theses of the *looping effect* of social kinds and the view that *we make up people*: social scientific theories are partly constitutive of their own objects, including the nature of action and agency.

The third theme left open in the previous Parts is the relationship between the view I defend and the fashionable research programme of evolutionary psychology. Part II was largely dedicated to evolutionary issues, and I have appealed to evolutionary science also in this Part. Nonetheless, I mentioned in Part II that I see my programme as highly compatible with John Dupré's (2001) vigorous criticism of evolutionary psychology. However, I had not yet accumulated sufficient conceptual tools for making the compatibility explicit. Now, on the other hand, the connection practically suggests itself.

Dupré draws our attention to the fact that a founding premise of evolutionary psychology is that it assumes the brain to "somehow contain symbolic representations" (Dupré 2001, 35).<sup>179</sup> Dupré believes that Wittgenstein's (1953) rule-following considerations provide an impeccable argument against any scientific research programme that builds on such a premise. Meanings reside in social practices, not in the brain. Original intentionality is to be found implicitly in social bedrock practices, not as instantiated in wonder tissue. Psychology operating with mental contents is a normative social science,

---

<sup>179</sup> In addition to evolutionary psychology, a good part of contemporary cognitive science also accepts this premise. Hence, my criticism of evolutionary psychology below applies, *mutatis mutandis*, also to much of mainstream cognitive science.

not a biological natural (*i.e.*, causal) science. Nonetheless, Dupré admits that “Wittgenstein’s arguments are [...] notoriously difficult and controversial, and it would be beyond the scope of the present work [*i.e.*, Dupré 2001] to examine them in any detail” (Dupré 2001, 35-36). The present Part of my study, in contrast, has sought to provide precisely such a detailed examination *vis-à-vis* Dupré’s thesis.

The rule-following considerations indeed imply that the traditional humanist framework, labelled the Standard Social Science Model, or the SSSM for short, by the evolutionary psychologists John Tooby and Leda Cosmides (1992, 23), emerges as essentially untouched from the attempts of the evolutionary psychologists to replace it with what Tooby and Cosmides call the Integral Causal Model, or the ICM.<sup>180</sup> Tooby and Cosmides (1992, 26) characterise the SSSM as locating *meanings* to the public sphere of culture such that individuals are seen to be capable for meaningful action only to the extent they participate in cultural practices. As I have argued in length above, the position of the SSSM is – *pace* Tooby and Cosmides – literally correct.<sup>181</sup>

It seems to me that the main reason why Tooby and Cosmides oppose the SSSM’s view of meaning and content is that in their view the SSSM is essentially committed to treating the human mind as a *tabula rasa* or, at most, as some kind of general-purpose information processing machine (Tooby & Cosmides 1992, 28-29). Tooby and Cosmides (1992, 24) think that the empirical neurosciences as well as evolutionary studies have established beyond all doubt that the human brain includes a number of different *context-specific* mechanisms (dispositions). In other words, the brain is neither a blank slate waiting to be written on nor a general-purpose processor. Instead, the brain instantiates a system of innate dispositions with very specific scopes.

I, of course, applaud this line of thought. A crucial step in my argument against the impossibility of human intentionality and meaning in general (the sceptical conclusion) was to assume that the brain is *not* merely a passive receiver of inputs or a general-purpose computer capable only of purely logical operations. Indeed, to solve the Infinity Problem we had to assume that the human brain contains context-specific, in-

---

<sup>180</sup> In what follows I defend only what I see as the philosophical core of the humanist SSSM, *i.e.*, the view implied by my own argumentation in this study. In particular, I am not committed to defending the specific views of any of the social theorists Tooby and Cosmides identify as paradigmatic representatives of the SSSM.

<sup>181</sup> Further, I have claimed my view to be largely compatible with that of Daniel Dennett (especially 2003). I think most theorists would agree with Don Ross (2002, 143) who argues that “Dennett’s work provides crucial philosophical background to the attack by evolutionary psychologists on what Tooby and Cosmides (1992) call the Standard Social Science Model”. I fear Dennett himself might be happier with Ross’ interpretation. However, the nature of my study is systematic rather than exegetical and, whether he likes it or not (and contrary to Ross’ claim), the lasting core of Dennett’s argumentation presents him as a defender of the SSSM camp (for a related reading of Dennett, see Kusch 1999, 340).

trinsic (and evolved) dispositions to, for example, classify objects in a certain ultimately unjustifiable way and to harmonise the first-order dispositions to match the behaviour of others.

Thus, the humanist SSSM in my sense is not only compatible with but actually presupposes the view that the human mind (qua biological) consists largely of functionally specialised mechanisms (dispositions) to produce behaviours that solve particular adaptive problems, such as preconditions for language, co-operation and so on. Moreover, these dispositions must be assumed to be sufficiently richly structured for constituting a complex architecture of first and second-order dispositions. In short, I have shown that far from rejecting the picture of the biological mind as a collection of functionally specialised causal dispositions, philosophically mature versions of the SSSM actually presuppose the view Tooby and Cosmides say is implied by the brain sciences.

Thus, Tooby and Cosmides may be right in pointing out that the traditional empiricists' picture of the biological mind as either a blank slate or a general-purpose computer is based on very primitive understanding of the natural sciences in general and evolutionary biology and neuroscience in particular. However, where empiricist social scientists may have somewhat unsophisticated understanding of the natural sciences, Tooby and Cosmides themselves are astonishingly naïve when it comes to philosophical semantics and the metaphysics of intentionality. They think that the rejection of the *tabula rasa* and the computer pictures of the biological mind implies a view where the context-specific, innate causal dispositions are intrinsically meaningful, *i.e.*, the ICM. This, of course, is nothing but the familiar hubris of imagining that solving the Infinity Problem suffices to solve the problem of rule-following (III.3.2).

The mechanisms (dispositions) Tooby and Cosmides talk about may well be necessary for meaning, content and intentionality, but they cannot be sufficient (because of the Normativity Problem). As we have seen, these phenomena include – exactly as the SSSM insists – an irreducibly social element. To treat innate, causal mechanisms as intrinsically meaningful and contentful is simply bad philosophy. Thus, when Tooby and Cosmides (1992, 113) talk about innate causal mechanisms that instantiate, among other peculiar things, “a social-inference module, [...] a semantic-inference module, [...] a theory of mind module”, we can see that they have a lot of philosophical housecleaning to do if they wish to avoid making fundamental category mistakes (cf. Kusch 1999, 351). When it comes to the social sciences, the proper task for the natural scientific approaches is to study the *biological preconditions* of the SSSM, not, *pace* Tooby

and Cosmides, replace the SSSM with the ICM or indeed with any other naïve oversimplification.

The fourth and final theme from the first two Parts of this study that I want to bring into conclusion on the basis of the present Part is closely related to the fundamental category mistakes committed by Tooby and Cosmides. What I have in mind is my somewhat ambivalent position regarding Bruno Verbeek's (2002) analysis of co-operation discussed in Part II. On the one hand, I sided with Verbeek in arguing that accounts insisting on mere individualistic instrumental rationality cannot resolve social dilemma situations (such as the Prisoners' Dilemma) in a collectively optimal way. Similarly, I appealed to Verbeek's penetrating argumentation when I defended the position that the attempts to analyse collectively optimal co-operation in terms of individual utility transformation rules cannot succeed. However, when it came to the solution Verbeek himself advocated, I dismissed it rather quickly.

Recall that Verbeek argues that since individual-mode rationality cannot deliver the kind of social co-operation we observe every day, we must assume that the explanations of social action must refer to both individual-mode rationalisations and non-rational, purely causal dispositions (which Verbeek calls co-operative virtues) that simply cause us to co-operate, despite our individual-mode rational inclination to defect. Verbeek's idea is, in short, that since "[c]ollective beneficial instrumental action is indeed individually irrational" (Barnes 2000, 57), co-operative action must be based on non-intentional, causal dispositions evolution has equipped us with.

In Part II my main argument against Verbeek's line of thought was simply that it hardly corresponds to our experience of social action. As Barnes puts this point, "[c]ollective action is scarcely well-described as irrational, since it may be exquisitely calculated and highly effective instrumental action, but it cannot be rationalised by reference either to altruistic or self-interested individual goals" (Barnes 2000, 57). I concluded that while Verbeek explicitly acknowledges the second part of Barnes' description of collective action, *i.e.*, its essential individual irrationality, Verbeek's straightforward causalism cannot account for the obvious intentionality and (collective) rationality of collective action. Hence, I argued, both our everyday experiences and scientific studies of social action support the theory of collective intentionality more than Verbeek's peculiar mix of rational considerations and blind causal dispositions that sometimes cause us to act against our rationality calculations.

Obviously, the arguments of the present Part reveal the philosophical profundity of my objection to Verbeek in a way that was not possible to see in Part II. I have ar-

gued that one cannot explain *actions* (including co-operative social actions) in terms of *causal dispositions*, for actions are essentially (logically, conceptually) tied to *reasons*, and these notions have no echo in the realm of causal dispositions. The intentional (rational) explanations of actions must be conceptually separated from causal explanations of behaviour. In particular, the attempts to combine these two incommensurable categories in the way Verbeek (and Tooby & Cosmides – or the Standard View for that matter) try to do simply will not, and cannot, work. The problem with Verbeek’s solution is not only that it contradicts our experience; rather, it is based on a massive conceptual (indeed, philosophical) confusion.

However, when we keep our philosophical categories clear it is possible to see the lasting achievements of Verbeek’s reasoning. In addition to offering perceptive criticisms of the prospects of individual-mode rationality, Verbeek sees correctly that human co-operation presupposes pre-intentional (or sub-personal), essentially social, causal dispositions. I have made the very same point above in terms of the indispensability of collective agency (III.4.2). I think Verbeek’s co-operative virtues could well play the same role as Kusch’s three devices or my own social dispositions. Thus, Verbeek’s reasoning is largely correct; his philosophical carelessness in mixing together rational, intentional explanations and non-rational, causal explanations simply makes him to fail to grasp the acceptable core of his own reasoning.

Indeed, one of the main morals of this Part of my study is that we must keep our philosophical categories clear. Hence, it is only appropriate to let G. H. von Wright, whose work is a great example of such clarity, to have the last word: “Natural science can be characterized as a study of phenomena under the ‘reign’ of natural [causal] law. Human science again is primarily a study of phenomena under the ‘reign’ of social institutions and [normative] rules.” (Von Wright 1976b, 415.)<sup>182</sup>

<sup>182</sup> Thus, this agreement with von Wright dissociates me from Kusch (2003, 342), who argues that von Wright’s (1989a, 41) statement that “philosophy of action [...] must terminate in a philosophy of society” shows that von Wright is still committed to the methodologically individualistic view that the psychological, including intentional actions, is conceptually prior to the social. However, on the very same page von Wright explains that in his view Wittgensteinian considerations imply that “the conceptualization of behavior as (intentional) *action* presupposes a community of institutions and practices” and *this* is why “philosophy of action must terminate in a philosophy of society”. Thus, *pace* Kusch, it seems clear to me that also for von Wright “[t]he philosophy of society provides the foundations for the philosophy of action, not *vice versa*” (Kusch 2003, 342). Hence, although the present dissertation by and large agrees with Kusch’s (1999; 2002; 2003, 341 ff.) socialisation of philosophy of action, I do not see this as a move foreign to von Wright’s philosophy of action. Of course Kusch is right in saying that von Wright (1971) does not explicitly discuss the constitutive role of the social behind actions – largely because von Wright’s *explicit* appreciation of “the strong influence of Wittgenstein’s last writings” came “[t]oo late to leave an imprint on *Explanation and Understanding* [von Wright 1971]”, although the “influence had been latently there since [...] 1947” (von Wright 1989a, 40). What I in this dissertation call von Wright’s view seeks to make explicit also the “latent” features – as I have explained, the animating interest behind my work is that of systematic reconstruction and analysis, not of historical exegesis. Moreover, von

## CHAPTER III.6: CONCLUSION

In this Part I set out to defend the framework of intentional action and agency assumed in the earlier parts of my dissertation. In III.1 I started with von Wright's Logical Connection Argument against the view that the framework is a *causal* framework. I argued that – contrary to the view accepted almost universally in contemporary analytic philosophy – Davidson's theory of action cannot deliver a refutation of the Logical Connection Argument or support the Standard View.

III.2 challenged another piece of contemporary orthodoxy. I argued that the Multiple Realisability Argument cannot deliver what it is all too often assumed to deliver, *i.e.*, *sui generis* mental causation in a causally closed physical world. I argued that to hold on to the causal interpretation of the intentional framework requires one to compromise physicalism in the way that many philosophers – although not all – find unacceptable and that the Multiple Realisability Argument was meant to circumvent.

III.3 approached the nature of the framework of intentional agency directly by examining the nature of intentionality. This quest took us to the Wittgensteinian problem of rule-following, which tipped the scales in favour of the non-causal, normative view. The framework of intentional agency is a collectively constructed and reproduced *normative* framework within which actions are *understood*.

However, while this kind of hermeneutic humanism is often seen as an enemy of naturalism, III.4 argued that the conditions of the existence of the normative framework can be satisfied by developing a non-intentional or pre-intentional complement to the theory of social facts, social practices and co-operative social action that was defended in Parts I and II. While explaining how there can be an irreducible space of normative reasons in our physical world, this theory itself operates with purely naturalistic causal notions. III.4 defended irreducible, interpretive humanism *within* general naturalism.

By constructing a naturalistic theory of the (normative) framework of intentional agency, III.4 turned the present Part into my final argument for the central thesis of my dissertation: The unavoidability of naturalised methodological holism in the social sciences. Meaningful action must be explained and understood in terms of norms implicit

---

Wright himself (*e.g.*, 1989b, 806 ff.) appears to think that in his writings he may actually have failed to express his own view properly. Hence the importance of the kind of systematic reconstruction – even at the cost of historical adequacy – this dissertation aims at. Stoutland, whose position I largely agree with, says that he defends a view that he attributes also to von Wright “in terms that he himself [von Wright] might not have used but that bring it into more explicit contact with current work in the philosophy of action”. I could say the same concerning the approach of the present study.

in social practices. However, this strong *methodological holism* was argued to be perfectly *naturalistic* in the sense that it (i) rejects all appeals to unnatural entities, be they Platonist principles, group-minds or wonder tissue and, moreover, (ii) builds on biological dispositions – thereby showing how irreducible, non-causal moral sciences fit into the same picture with causal natural sciences. Finally, III.5 explicated the non-causal view of intentional action and its explanation and the constructivist theory of collective intentionality that follow from naturalised methodological holism.



APPENDIX:

DURKHEIM'S GOD or

THE SOCIOPHILOSOPHY<sup>183</sup> MANIFESTO

---

<sup>183</sup> By the label "sociophilosophy" (adopted from Kusch 1997) I mean a philosophical approach which holds that certain central notions of philosophy, such as the intentional mind, meaning and knowledge, are essentially *social* notions.

## INTRODUCTION

In this Appendix I want to explain how I think the view defended in this dissertation (Part III in particular) is to be located in the field of philosophy. To keep this Appendix appropriately short I will have to paint with a broad brush indeed. The combination of a limited space and a vast topic is far from ideal. Consequently, rather than as the last word on these issues, the arguments offered in this Appendix might serve as indications of the paths further research could take. However, I firmly believe that the task I undertake here is important, not only in order to guide further research, but also – and perhaps primarily – because only this kind of general discussion will guarantee that the philosophical nature of the standpoint I defend is seen in its proper light. In particular, I fear that if I do not explicate exactly what kind of general position I am committed to, my views can all too easily be disregarded as a form of rather banal social constructivism, naïve antirealism and unacceptable cultural relativism (which indeed was the accusation of Niiniluoto in I.3), whereas in fact nothing could be further from truth.<sup>184</sup>

---

<sup>184</sup> Accusations of anti-naturalistic constructivism and relativist anti-realism are precisely what I tend to hear when I present arguments this study consists of at philosophical conferences. While I am sympathetic to some views that accept these labels (*e.g.*, Kusch 2002), I do not recognise the picture people often claim to follow from my arguments – including the relativism that, for example, Halfpenny (2001, 378) claims to follow so straightforwardly from Wittgensteinian views that he can present it as an uncontested feature of Wittgensteinian philosophy in a handbook of social theory. Hence the importance of including this Appendix in my dissertation.

## A.1: WHEN THE COMMUNITY IS MISTAKEN

### A.1.1 SOCIAL RELATIVISM AND TALKING *ABOUT* THE WORLD

The motivation for the issues I want to raise in this Appendix comes from my rejection of the naïve communitarian view, in which I identified two major problems (III.3.3). First, it was seen to render the linguistic community infallible. When the community agrees on a judgement that something is the case in the world, it cannot make a mistake – regardless of what it happens to judge to be the case. Second, the naïve communitarian view was argued to re-introduce the Regress Argument and hence to fail as a solution to the rule-following problem. Crucially, my main reason for rejecting the view was the second problem and *not* the fact that it implies objectionable relativism. Consequently, I have argued that my own social solution to the rule-following problem avoids the Regress Argument. The question is, then, whether my view nonetheless commits the first sin of the naïve communitarian view, that of implying questionable relativism.

At first glance it seems that the answer must be in the positive. After all, I argued that meaning of a judgement is constituted by the practice of applying the judgement in a community, *i.e.*, that we replace truth-conditional semantics with the Kripkean view that builds on normative assertability conditions which are constituted by tacit collective acceptance in social practices.<sup>185</sup> This sounds very much like social relativism. However, I want to argue that the essentially *anti-individualistic* collective agency view that grounds my social solution to the problem of rule-following – and which in effect distinguishes the view I defend from the naïve communitarian view with its *summative* (and thus essentially *individualistic*) conception of the social – is also the factor that sets my view clearly apart from the relativism of the naïve communitarian view.<sup>186</sup>

In short, according to the naïve communitarian view the individual must compare her application of a term to the collectively accepted way of applying the term, which gives the truth conditions for her application. In other words, the community is

---

<sup>185</sup> Hence, although I did claim in III.3.2 that my acceptance of Condition (2) dissociates my view from the most provocatively relativistic formulations of the meaning finitism of the Edinburgh School sociologists of science, it could be argued that it is my interpretation of Condition (3) – my conviction that the normativity of meaning must be due to a social aspect – that commits me to social relativism. Meanings may be determined, but they are determined largely by social agreement in practice.

<sup>186</sup> As will become clear below, my views here are strongly influenced by Brandom (1994 & 2000) and Esfeld (2001), with whom I am largely in agreement. Brandom and Esfeld, however, do not discuss in terms of collective agency and collective intentionality. Thus, despite the shared starting points, they would not necessarily appreciate the way in which I discuss these issues.

seen as something external to the individual, *i.e.*, as part of the circumstances an individual must take into account when performing actions (including the actions of applying concepts). This view is certain to lead to Kripke's sceptical conclusion. In contrast, throughout this study I have emphasised that my picture of sociality is utterly different. Social practices are not something external to individuals – and indeed individuals qua agents are not prior to social practices. Rather, it is part of the constitutive essence of individuals that they, via collective agency in the case of bedrock practices (constructing the framework of agency) and via collective we-mode intentionality in the case of full-blown practices (within the framework of agency), *participate in* and indeed *form* the practices.

This is a critical dissimilarity. Recall that social practices are essentially and irreducibly *normative* (from the point of view of the participants, not metaphysically) in nature. The speech acts of individuals normatively commit them to certain other propositions, and exclude commitments to other propositions (this is what makes the speech acts contentful performances in the first place). In the naïve communitarian view the agents are interested in harmonising their understanding of the contents of such commitments with the collective understanding of their contents, and this depends only *on what the majority view happens to be*. In my picture, in contrast, individuals are *responsible* participants collectively interested in getting the commitments *right*, and this depends *on the way the world is* (see Esfeld 2001, 65 for a similar line of thought).<sup>187</sup>

This idea is at the core of Brandom's (1994) inferential role semantics as well. Brandom thinks, correctly to my mind, that the rule-following considerations show that an acceptable semantics cannot be *representationalist* (or truth-conditional) in nature. The Cartesian project of treating *representational content* as a primitive notion, and then worrying about the *descriptive accuracy* of different representations, is doomed to fail. Rather, we must start with the normative notions of commitment and justification, and explain representation on the basis of them. Representation and truth conditions are important, but not semantically primitive. However, also the inferentialist conception of meaning must explain the realist requirement that often our language is about the world.

---

<sup>187</sup> Note that here I put the points about moral philosophy I made towards the end of Part II into ontological and semantic use. Recall that I mentioned that the denial of individualism inherent in the theory of collective intentionality does not amount to saying that individuals should conform to some timeless Kantian principles; rather, the point was to take seriously the Kantian doctrine of treating all individuals as ends in themselves as opposed to the rational choice notion of strategic action where other agents are simply part of the circumstances (*i.e.*, mere means) one has to take into account when striving for individual-mode goals. Similarly here the rejection of naïve communitarianism emphasises social practices constituted by *egalitarian* participants who are allowed and even required to assess the others.

My claim is that this can be done only if we accept the social practice theory of meaning I have defended in this study.

As I have explained, at the primitive level semantics is about normative statuses. To possess a concept is to master the ultimately practical skills of applying the concept only in appropriate circumstances and inferring other commitments that follow from such applications (cf. Esfeld 2001, 55), and this is possible only within a social bedrock practice. The (representational) content of a concept is determined by the normative, inferential relations the concept enters into and not *vice versa*. In other words, content is constituted by a position in a normative web. This is the aspect of *meaning holism* in the present theory. The norms guiding the proprieties of application and inference and thus constituting the normative web are, ultimately, implicit in social practices. This is the aspect of *social holism* in the present view.

Crucially, far from bringing in unacceptable relativism, it is precisely the social aspect of semantics that makes representation and realism possible. To keep score of the commitments of others we must keep score both of their *judgings* (the undertaking of a commitment) and what they have *judged* (the content of the commitment). And this latter aspect requires, at least in the case of judgements involving natural kind terms, that we keep track of concrete relations in the world (*articulated* in the inferentialist sense as well).<sup>188</sup>

For another person to use my *judging* as a premise (or a reason) in her own inferences, she must know what I have *judged* to be the case. It is the social perspective constitutive of normativity that also introduces the representational aspect of meaning. By endorsing a contentful sentence I commit myself to the consequences of the endorsement, even if I do not know or understand them. By making a judgement I undertake a public commitment with external and objective (in the technical sense of Part I) consequences. The social aspect of meaning constitutes a huge positive freedom: the social aspect enables me to talk about things that I do not understand properly, such as quantum physics or the proposed constitution of the European Union. I undertake a

---

<sup>188</sup> For example, if I commit myself to abstain from eating anything that has a heart, then *the way the world is* implies that I am committed to not to eat birds, whether *I* know it or not. This is why the others who keep score of my commitments must introduce the representational aspect to know what exactly I am committed to. Similarly, if I am entitled to the status of being a PhD candidate at the London School of Economics, then it follows that I am also a PhD candidate at the University of London (since the LSE is a college of the university), whether *I* know it or not. In this case, however, it is enough that the scorekeepers keep track of our collectively accepted norms; they do not have to bother with concrete relations. Social kind terms are essentially self-referential; their assertability conditions depend on our practices, not on the way the physical world is – social relativism is of course true of the social world (Kusch 1999). Or as Brandom (1994, 53) puts this, “[w]hatever the Kwakiutl treat as an appropriate greeting gesture for their tribe [...] is one; it makes no sense to suppose that they could be collectively wrong about this sort of thing”. Natural kind terms, in contrast, are answerable to our practices *and* the world.

commitment, and the linguistic division of labour made possible by a social theory of meaning (cf. Putnam 1975) allows me to let the others to find out what exactly I am committed to. The possibility of committing oneself objectively to unknown consequences is possible only within a community – an important positive freedom assigned to an individual by the socially holistic picture.<sup>189</sup>

In other words, representation is based on our interest in finding out the real *normative status* – what a person really *is* committed to – as opposed to mere *normative attitudes* – what anyone, including the person undertaking the commitment, *takes* the person to be committed to. As Esfeld emphasises, “the meaning of a belief *p* is not determined by those other beliefs that a person actually has, but by those other beliefs to which she is committed and entitled to and to which entitlement is precluded by endorsing *p*” (Esfeld 2001, 66). Moreover, as Esfeld also points out, this is precisely the point of the Wittgensteinian insistence on avoiding all forms of psychologism or mentalism in semantics. The social aspect of semantics makes it possible for our beliefs and assertions to be answerable to the world and not only to other beliefs and assertions. Far from destroying the possibility of realism, social holism makes room for realism worth wanting.

Indeed, the point of both the social solution to the Normativity Problem in Part III and my discussion of the possibility of objective social (normative) facts in Part I is precisely that the distinction between a normative status (objective) and normative attitudes (subjective), which grounds representation, is available only within a social practice. This holds, because to determine the normative statuses, as opposed to mere normative attitudes, requires that the participants of a linguistic community distinguish between the *de dicto* (judgings, or commitments *acknowledged* by the speaker) and *de re* (judgeds, or commitments actually *undertaken* by the speaker) ways of assigning contents to their interlocutors.<sup>190</sup> The introduction of this distinction – made possible by the social aspect of semantics – is the introduction of the perspective required for full appreciation of the representational aspect of intentional states and sentences.

---

<sup>189</sup> The fact that the fundamental sociality of agents brings in positive freedoms unavailable for an isolated individual is, of course, one more major point of connection between my anti-individualism in the philosophy of social science, mind, language and action on the one hand and anti-individualistic accounts in social and political philosophy on the other.

<sup>190</sup> The easiest way to explain the difference between *de dicto* and *de re* reading of a content is by means of an example. To Europeanise Brandom’s (1994, 500 & 2000, 170) example, consider the statement “The President of the European Commission will be a woman by the year 2020”. Read *de dicto*, this means that the sentence “The President of the European Commission is a woman” will be true by the year 2020. Read *de re*, the statement means that the present President, José Manuel Barroso, will be a woman by the year 2020.

In III.3.3 we saw Boghossian (1989) to complain that the naïve communitarian view cannot cope satisfactorily with cases where a community, when observing sufficiently horsey looking cows on a dark night, believes *incorrectly* that there are horses in front of them. The present view, in contrast, can accommodate this situation easily. Each member of the community believes that there are horses in front of them. When they keep score of the commitments of others they ascribe these commitments (beliefs) in the *de re* sense: So and so believes *of* those objects (*de re*) in front of us *that* they are horses (*de dicto*).

Of course, in Boghossian's example every ascriber believes *of* the cows (a *de re* identification of a commitment) *that* they are horses (a *de dicto* identification of a commitment). Crucially, however, the present view – precisely because it includes a social aspect – leaves conceptual room for the community to be mistaken. Since the individuals are egalitarian participants in communal practices (and not disparate individuals external to the practice as in the naïve communitarian view), it is an open possibility that one of them (or indeed a new member joining the community) will find out and convince the others that what the whole community believed (*i.e.*, *that* they are horses) *of* these objects, *i.e.*, cows, is in fact wrong.<sup>191</sup>

Conceptual contents are *essentially perspectival* in the sense that they are always specified from some point of view. But this perspectivalism is taken explicitly into account in the *social* picture by distinguishing between the *de dicto* (normative attitude) and *de re* (normative status) ascriptions of contents (see Brandom 2000, 177), and hence an explicit acknowledgement of perspectivalism is essential to participants' understanding of contents – they know that each perspective may hide certain aspects of the situation that *ought* to be taken into account. Social practices based on collective agency or collective we-mode intentionality are *open* to criticism in this sense. The naïve communitarian view – not to mention any explicitly individualistic view! – cannot accommodate this crucial aspect.<sup>192</sup>

---

<sup>191</sup> This, of course, is simply another way of highlighting the open-ended nature of semantic practices so crucial to meaning finitism.

<sup>192</sup> Although Brandom and Esfeld do not discuss in terms of collective notions, I believe that they are after the same point when they insist that social practices should not be seen as “I-we practices”, where an individual is subordinate to an external practice fixing static truth conditions, but as “I-thou practices”, where the dynamic interaction of egalitarian individuals is what the practice consists in. Moreover, I have always found it rather puzzling that in the context of the philosophy of language an appeal to a social practice as the guarantor of representational accuracy is often seen as somehow dubious, whereas in the context of the philosophy of science it has been commonplace since Charles S. Peirce's days to think that open peer criticism of the *judgings* of others is essential for the collective quest for truth – and that hence the true possessor scientific knowledge is the *scientific community*.

Thus, the social view I defend can accommodate both (i) cases where there are truths about which the community is completely ignorant and (ii) cases where a proposition universally accepted in the community is false. In contrast, the naïve communitarian view rejects at least (ii) and, perhaps, also (i). As Brandom (1994, 602-603) argues, unacceptable cultural relativism requires that we reject both (i) and (ii), for that would imply that  $p$  is true iff the community accepts that  $p$ . This is not an implication of my view, and hence I conclude that my version of the social solution to the problem of rule-following does not imply the kind of social relativism that plagues the naïve communitarian solution. In other words, the present social view includes an explicit acknowledgment that brute natural facts are *independent facts* (I.1.4).<sup>193</sup>

#### A.1.2 THE BRANDOM – KUSCH DISSENSION: A DISSOLUTION

I have created my position above largely on the basis of Brandom and Kusch. Although my dissertation concentrates on issues somewhat different from their core interests, I have implied that the views of these two authors are highly compatible with my view – and with one another. However, Kusch (2002, 256 ff.) argues that there is a fundamental difference between his view and that of Brandom. So is the present view completely misguided in emphasising the similarities between Brandom and Kusch?

First of all, Brandom builds his view of social practices on “I-thou relations” rather than “I-we relations” (recall Footnote 179). Kusch (2002, 256) thinks that by not grounding objectivity on I-we relations Brandom rejects the kind of intersubjective sociality that is ultimately required for normativity. Kusch (2002, 258-259) appears to be saying that to prefer I-thou relations to I-we relations in this context is to return to unacceptable individualism. While I agree completely with Kusch’s rejection of individualism, I think that Brandom’s rejection of I-we understanding of practices is *not* an acceptance of individualism but a rejection of the naïve communitarian view, *i.e.*, the idea that the community and its practices instantiating norms are essentially external to an individual.

---

<sup>193</sup> One might find it surprising that I do not discuss at all the standard form of cultural relativism, *i.e.*, the claim that the practices of different collectives may well be incommensurable, making it impossible to understand foreign cultures. However, in my picture social practices are ultimately based on the innate dispositions we all share as humans, and hence understanding foreign practices, while often tremendously difficult, is not *a priori* impossible (cf. Esfeld 2001, 147 & Collin 1997, Chapter 2). It seems to me that even the infamous arch-relativist Peter Winch (1958, especially 181) shares this view – as does Wittgenstein: “The common behaviour of mankind is the system of reference by means of which we interpret an unknown language” (Wittgenstein 1953, §206).



As I have argued above, the naïve communitarian view is in fact committed to an ultimately individualistic view of sociality. Thus, I think that by replacing the I-we account of sociality with his I-thou account of sociality Brandom is in fact making room precisely to the kind of non-summative account of sociality (I.1.5) that social bedrock practices require (III.4.2). Michael Esfeld puts this nicely in his exposition of Brandom's philosophy of mind and language: "Setting social holism out in terms of open-ended and symmetric I-thou relations [...] prevents us from having to identify at some stage what is correct with social facts in the sense of what a community takes to be correct" (Esfeld, 200, 132). The I-we view emphasises truth-conditions, and the I-thou view assertability conditions (III.3.3). However, I agree with Kusch that Brandom's terminology *sounds* misleadingly individualistic, and thus I have favoured the terminology of collective agency and collective we-mode intentionality – which, just as Brandom's rejection of I-we accounts, is based on the rejection of the idea that the individual is conceptually prior to and independent of the social. In my view Kusch and Brandom are in agreement here.<sup>194</sup>

Second, Kusch (2002, 256-259) appears to criticise Brandom's (1994, 331) insistence that the nature of our discursive practices is partly up to the mind-independent world. I address the metaphysics behind this issue in detail in A.2 below, but I want to partly clarify this issue already here.

As we saw, in Brandom's (and my) view it would be incorrect if the whole community started to apply the word "cow" to horses. Kusch, on the other hand, holds that the whole point of the rule-following considerations, on which also Brandom is building, is that independent objects themselves cannot determine how they are to be classified. Thus, if the community dictates that the word "cow" is to be applied to these things (horses), then this is what we ought to do.

Despite the seemingly contradictory positions, I think also here Brandom and Kusch are largely in agreement. Kusch is thinking about the problem of how to follow a particular rule, namely how the word "cow" is to be used in classifying objects. It seems obvious that the community cannot be wrong about this. Kusch is concentrating on the *socially* holistic aspect in the inferentialist picture. In contrast, when Brandom makes the point about the relevance of the mind-independent world to our discursive practices, he is thinking about a situation where we have a full-blown language such that the basic unit of application is not a word but a *judgement* and, moreover, where we have a num-

---

<sup>194</sup> Kusch (2002, 256 & 258) also points out certain inconsistencies in Brandom's terminology (especially in his use of the term "fact"). Although Kusch is in my view correct here, this issue is not crucial to my study.

ber of normatively interconnected judgements such that the interconnections constitute the space of reasons that assigns meanings (conceptual contents and not mere extensions) to the judgements. In other words, when making his point Brandom is concentrating on the aspect of *Quinean meaning holism* in the inferentialist view.

Brandom's idea is that *given* the *other* normative commitments of the community, to judge horses to be cows is a mistake by the standards of the community, even if no member of the community realises the mistake. If the community wants to stick to the judgement, some other commitments and judgements must eventually go (Quine 1951). Brandom's rejection of summative accounts of social statuses and truth-conditional semantics implies that social practices that ground meanings are always open to criticism. A member of a community can always argue against a universally accepted view that given their other commitments, the community should actually change its view about something. This is the core of both Brandom's replacement of inflexible I-we practices with open-ended I-thou practices *and* Kusch's finitistic emphasis on concrete applications of terms and judgements being always subject to reconsiderations and assessments.

This brings me to the third and last apparent disagreement between Kusch and Brandom that I wish to dissolve. Kusch (2002, 259) argues against Brandom's dismissal of the I-we views of objectivity that in Kusch's we-based view there is no external standard in need of interpretation, but the communal view is constructed and maintained in concrete, local interactions, where essentially social individuals harmonise their inclinations. Thus, the community view is not fixed but subject to continuing, gradual change. I have no quarrel with this: it is indeed what I have argued for too. But as we saw above, it is also precisely what Brandom is after with his replacement of the I-we picture with the I-thou picture. Similarly, when Kusch (2002, 259) emphasises that although the community can be mistaken, the *judgement* that this is the case must nonetheless be made from within some other community (or a later time-slice of the same community), this is in my view exactly what Brandom is after as well when he emphasises that (i) all judgements are perspectival, but (ii) the I-thou view of sociality implies an explicit acknowledgement that each perspective is open to criticism from some other perspective (A.1.1). I see no deep disagreement here.<sup>195</sup>

---

<sup>195</sup> In his 2006 book, Kusch (2006, 195-201) reformulates his criticism of Brandom so that it is not a criticism of Brandom's I-Thou position anymore but a demonstration that Brandom is mistaken in thinking that Kripke (1982) is committed to the I-We view in Brandom's sense. I have no quarrel with this. Indeed, I have argued that the view I have built on the basis of Brandom, Kripke, Kusch and Wittgenstein (and others) does not fall prey to Brandom's criticism and that even if Brandom and Kusch disagree on whether Kripke's view is acceptable in this matter, they do agree to large extent (explained in detail

## A.2: DURKHEIM'S GOD: CONCRETE PRACTICES AND DIRECT REALISM

The social view I defend does not imply social relativism. In this Section I want to take this claim further and suggest that in fact *realism* worthy of the name actually presupposes the kind of social picture in which assertability conditions replace truth-conditions. The problem of realism is of course a huge issue in philosophy and I cannot hope to discuss it exhaustively in a short Section like the present one, but I think it is nonetheless important to indicate what kind of view regarding realism is implied by the views I defend in this dissertation.

Let us start with the observation that, as I show above, contrary to, e.g., Rorty's (1979) relativism, the present social practice view does not deny the importance of representation. What is denied is only the Cartesian conviction that representation is a primitive semantic notion. *Realism*, however, is usually understood as essentially *representational realism*, which is the thesis that (i) there *is* a physical world ontologically independent of us and that (ii) we have *epistemic access* to the physical world *via representations* (cf. Esfeld 2001, 112). The anti-representationalist social practice view, in contrast, implies *direct realism* that accepts the ontological thesis (i) but rejects the representationalist thesis (ii), namely the idea of epistemic intermediaries (representations) between us and the world.<sup>196</sup> Or so I argue.

Note that my argumentation has taken me to a broadly speaking Sellarsian (1963) position. I have argued that the notions of agency, action, mental content, knowledge, belief and indeed intentionality belong to the logical space of reasons,

---

above) what the acceptable view is (be it Kripke's view or not). However, even though Kusch's 2006 book does not imply such a strong disagreement between Brandom and Kusch regarding the issue at stake (although it does imply one concerning the correct reading of Kripke), I believe the present dissolution is nonetheless an important philosophical exercise in making the present view as clear as possible.

<sup>196</sup> Accordingly, I think the social practice view I defend is the best way to make sense of the criticisms of contemporary analytic philosophy presented within the so-called *phenomenological tradition* in continental philosophy. Phenomenologists claim that analytic philosophy is dominated by the Cartesian tradition of assuming that when we are conscious we are primarily aware of our own ideas (*i.e.*, representations) and not the world itself. The phenomenological tradition wants to replace the Cartesian representationalism with direct realism in my sense by claiming that the mind is essentially intentional, *i.e.*, about real objects and not simply a possessor of representations. Phenomenologists think that it is the Cartesian underpinnings of analytic philosophy that have led to the well-known embarrassments of analytic philosophy, such as the mind-body problem, the problem of realism, the problem of other minds, solipsism and so on. Just as the phenomenological tradition, also I think that the way out of these essentially Cartesian problems is the rejection of the dual thesis of representationalism and individualism. As I have argued throughout this study, we should not build on disparate (including epistemic, representational autonomy) individuals, but acknowledge the priority of collective agency, social practices and collective we-mode intentionality (or the priority of intersubjectivity over subjectivity, as phenomenologists like to put this point).

which is essentially *normative* in nature and made possible by social bedrock practices. In contrast, when we do natural science and give empirical descriptions of causal processes we are operating with the logical space of *non-normative* causation. The logical space of reasons is governed by norms and rules, whereas the logical space of causation is governed by laws of nature. Moreover, the very heart of my argumentation in defence of the intentionalist programme and the possibility of *sui generis* human science is that the logical space of reasons, although compatible with ontological (materialistic) monism, is *not* reducible to the logical space of causation. It is not reducible, because the relations “that constitute the logical space of nature [causation] [...] are *different in kind* from the normative relations that constitute the logical space of reasons” (McDowell 1994, xv).

Thus, as we have seen, phenomena belonging to the logical space of reasons *have no echo* (Davidson 1974a, 231) in the logical space of causation, and thus *short of changing the subject* (Davidson 1974a, 230) we cannot move from one logical space to another in our explanations and discussions. This fundamental discontinuity of the two spaces is crucial for irreducible intentionality, mental realism and free agency, for we cannot move from the space of reasons to the space of causation. However, the very same discontinuity seems likewise to block a move from the space of causation to the space of reasons, since an “‘empirical description’ cannot amount to placing something in the logical space of reasons” (McDowell 1994, xv). The misunderstanding that this could be done is what Sellars famously calls *the Myth of the Given*, which is based on a serious philosophical failure to see the unbridgeable gap between the two logical spaces that are essentially different in kind. In III.3.3 we saw Kripke (1982) formulating the unacceptability of the Myth of the Given in terms of the rule-following problem forcing us to reject the fact-based view of semantics (the idea that a non-normative fact could ground normative meanings). It seems that my view is indeed committed to this.

The obvious question, then, is to ask what comes of realism in the sense of the claim that we have epistemic access to the physical world that belongs to the space of causation? Realism seems to require that there is no fundamental gap between the two spaces, or between the mind and the world. The view I have defended is at risk of closing our mind into a Cartesian realm of ideas (the logical space of reasons – or, to the extent that Kant is read as a representationalist, the Kantian phenomenal world), which may be nothing like the realm of physical things in themselves (the logical space of causation – or the Kantian noumenal world). Thus, the fundamental problem of realism is that intentionality, contentful propositions and mental contents belong essentially to the

Sellarsian logical space of normative reasons, which is governed by norms of rationality and appropriate inferences, but the objects that our thoughts are about belong to the Sellarsian logical space of non-normative causation, governed by the causal laws of nature. There seems to be no way of having epistemic access to the space of causation, for epistemic relations belong exclusively to the space of reasons.

In Part III we have repeatedly seen that attempts to force these two fundamentally different realms following essentially different logics into one picture is bound to fail. However, this is exactly what representational realism aims to do by claiming that representations bridge the gap between the two spaces. Thus, representational realism repeats the same fundamental category mistake that was so fatal for the Standard View, dispositional solution to the problem of rule-following, evolutionary psychology, the co-operative virtue theory of co-operation etc. All these lines of thought are guilty of accepting the Myth of the Given, namely the view that purely causal processes (including sense-data) could somehow ground normative (epistemic) features that determine representational content. In contemporary philosophy representations have replaced the Cartesian pineal gland as the *deus ex machina*. Both notions, however, simply beg the question.

Sellars' point is that sense-data, if there are such at all, belong to the logical space of causation. *Contentful* experience, on the other hand, is normative all the way down, and thus has no echo in the realm of causation. To be contentful is to have a position in the game of giving and asking for reasons, and this game is constituted by normative proprieties. To think that non-normative causal constraints could play a role in such a normative game amounts to an uncritical acceptance of the Myth of the Given, a fundamental category mistake on a par with the naturalistic fallacy in ethics.

Thus, the unacceptability of the Myth of the Given seems to re-establish Cartesian dualism. Individuals qua intentional agents (knowers, minds, souls) and not qua physical bodies belong to the normative realm of ideas and representations, whereas the physical world belongs to the non-normative realm of causation. To resist the representationalist reading of Kant that leads to Cartesian scepticism, according to which we can never know the world as it really is but only our own representations that in fact belong to the space of reasons, seems to require that we indeed postulate a Cartesian benevolent god to guarantee the truth of our representations, or his Leibnizian colleague to establish a harmony between the two realms.

Of course we can also try to sidestep the problem by rejecting one or the other of the two realms. We can refuse to admit that there exists a normative space of reasons

and representations, but then we are back with the kind of eliminativism I have already rejected.<sup>197</sup> Alternatively, we could follow the objective idealism of Berkeley and, perhaps, Hegel and deny the existence of the physical world of causation, and claim that what is real must be rational. This would amount to the view that the world is somehow normative and conceptual in itself, and hence we can have epistemic access to it without accepting the Myth of the Given. If the world itself is not conceptual, the reasoning goes, the world cannot set rational (epistemic) constraints on our beliefs. But we can have beliefs about the world and, hence, the world must be intrinsically conceptual and normative.

To a modern thinker this sounds like a completely unacceptable position, but it seems to me that at least McDowell (1994) is willing to bite the bullet and accept the idea as a necessary postulate (Kusch 2002 shares this reading of McDowell). We know the world precisely to the extent that it is essentially normative and conceptual, *i.e.*, a possible object of beliefs and other propositional attitudes residing in the space of reasons. In coherence with his quietism discussed in III.3.3, McDowell thinks that we simply have to accept this undeniably astonishing conclusion. Thus, one of McDowell's central claims is that we should not think that the world in itself must be equated with the logical space of causation (McDowell 1994, xx). Arguably this animating thought of objective idealism is alive also in the thought of, for example, Millikan and Dretske, who we have seen to insist on the anti-naturalist view that there are intrinsically and independently (of us) normative functions out there in the physical world, and hence we can have epistemic access to the world.

I, on the other hand, want to keep following Wittgensteinian naturalism and hold that the objective and irreducible normative space of reasons is simply bootstrapped into existence within our social bedrock practices in the way I have described. Recall how Wittgenstein opens his *Philosophical Investigations*:

---

<sup>197</sup> Indeed, as McDowell puts this with his distinctive style, more often than not the failure to acknowledge the space of reasons does not even merit the status of a philosophical view – it is rather a form of “unreflective scientism: not a principled avoidance of unprofitable philosophy, but a way of thinking that does not explicitly appreciate what threatens to lead to it. Perhaps people who think like this should be congratulated on their immunity, but it ought not to be mistaken for an intellectual achievement.” (McDowell 1994, 89.)

Let us imagine a language [...] between a builder A and an assistant B. A is building with building-stones: there are blocks, pillars, slabs and beams. B has to pass the stones, and that in the order in which A needs them. For this purpose they use a language consisting of the words 'block', 'pillar', 'slab', 'beam'. A calls them out; – B brings the stone which he has learnt to bring at such-and-such a call. – Conceive this as a complete primitive language.  
(Wittgenstein 1953, §2.)

Contents and meanings reside, ultimately, in our practices. But these practices are not only purely theoretical inferential practices. Rather, they are wide and concrete in the sense that they *include* physical objects (cf. Esfeld 2001, 150 ff.). Some commitments we undertake do not require us only to produce certain noises, but to perform certain *actions involving physical objects*, such as bringing the object to which the word “block” is appropriately applied within the practice.<sup>198</sup>

Wittgenstein's view is *modest* or *naturalised* idealism. Physical objects *are* indeed normative in nature, but *only* to the extent they are assigned a role in our social practices. More precisely, objects can have a normative status within our practices – certain pieces of paper are money, which means that their bearer is allowed and required to perform certain actions *involving* those pieces. In this sense physical objects *are* conceptual, for to be conceptual is to have a position or role in a language game, and this clearly is the case with blocks, pillars, slabs and beams in Wittgenstein's example. Physical objects can be assigned a normative status. To the extent that we do so, the objects can enter the normative space of reasons, since the status boils down to normative rules of appropriate action (Part I).

So this is how Wittgenstein points towards direct realism. Semantics does not build on the notion of representation, but on the notion of practice. Practices ground meanings, and those practices involve individuals qua physical bodies and other physical objects such as blocks and pillars. Crucially, however, *within* such practices it is possible to be assigned an objective normative status, including the status of an individual qua an intentional agent or a block qua a possible object of beliefs and knowledge. Such social statuses consist of proprieties, implicit in the practice, the totality of which constitute the *human form of life* featuring not only causal relations but also intentional agents, knowers and objects of knowledge; representers and things to be represented.<sup>199</sup>

---

<sup>198</sup> Needless to say, in another sense Wittgenstein's example is not a good example of human language with conceptual contents in my sense. Non-conceptual animals could play this game, for it involves no judgements and inferences. As I have explained, in my view the basic semantic unit is a judgement that can function as a premise or conclusion in an inference, and sub-sentential units, such as concepts, are to be explained in terms of their role in judgements (cf. Brandom 1994, 360 ff.).

<sup>199</sup> In other words, I find myself again in broad agreement with the phenomenological tradition, which holds that when we are doing phenomenology, *i.e.*, when we treat individuals qua intentional agents, we

The social practice view gives us an ontologically naturalistic but nonetheless non-reductive theory of intentional agents with a *direct* epistemic access to the world. There are no representations or indeed any other epistemic intermediaries between us and the world. Rather, the very objects of the world themselves may have a role in our epistemic practices by being assigned a normative status – just like the pieces of paper in my pocket are not something behind money. They simply are – or count as – money in our practices. Similarly, certain objects simply play certain roles in our epistemic practices; we need to assume *neither* that they are conceptual or contentful independently of our practices (just as the pieces of paper in my pocket are not money independently of our practices) *nor* that it is not the object themselves that play the roles or have the normative statuses (just as it is the pieces of paper that are – count as – money).<sup>200</sup>

At the risk of repeating some of the things I have just explained, I believe it is nonetheless important to clarify the position I have arrived at in the context of Davidson's (1983) defence of a coherence theory of truth and, in particular, knowledge. Davidson's starting point is Quine's (1951) famous criticism of the two dogmas of empiricism. The first dogma Quine rejects is the claim that we can separate analytic statements (the truth of which depends solely on the meaning of words) from synthetic claims (the truth of which depends on two factors, meaning and the world) a statement by statement.

What is crucial in the present context, however, is Quine's rejection of the second dogma, the thesis of meaning atomism, which claims that an individual statement can have a meaning in isolation from other statements. Quine replaces this with his famous thesis of meaning holism, which we saw (in III.1.3) Davidson to formulate as the claim that "the content of a propositional attitude derives from its place in the pattern" (Davidson 1970, 221) such that "it is misleading to speak of the empirical content of an individual statement" (Quine 1951, Section 6). As Quine also points out, the rejection of the dogma of meaning atomism means that we cannot hold to the first dogma either, for,

---

*bracket* the natural world (individuals qua physical objects; qua inhabitants of Sellars' logical space of causation) and operate within the *Lebenswelt* (the human form of life, *i.e.*, individuals qua agents and objects qua possible objects of conceptual experience; in short, Sellars' normative space of reasons) constructed in our intersubjective practices. Consequently, the often-heard claim that the phenomenologists' bracketing of the external world would amount to some sort of internal realism where we do not have epistemic access to the world as such is a gross misapprehension, for that would amount to a return to representationalism, the rejection of which is, as I mentioned above, at the very heart of the phenomenological tradition.

<sup>200</sup> I believe this explains also Stoutland's perplexing claim that our material environment includes "intentional, and not merely physical phenomena" (Stoutland 1988, 44) in the sense that physical objects can be "observed as intentional phenomena" (Stoutland 1988, 55; recall Footnote 97).



as I have argued, a single statement is meaningful only when it occupies a normatively defined role in the game of giving and asking for reasons.

Add to this Davidson's conviction – which I have also argued for in this dissertation – that the principles of rationality constituting the pattern from which meanings are derived (*i.e.*, Sellars' space of reasons) “have no echo” (Davidson 1974a, 231) in the empirical world (Sellars' logical space of causation), and it follows directly “that nothing can count as a reason for holding a belief except another belief” (Davidson 1983, 141). Thus, when Quine claims that “our statements about the external world face the tribunal of sense experience not individually but only as a corporate body” (Quine 1951, Section 5) such that “[t]he unit of empirical significance is the whole of science” (Quine 1951, Section 5), in Davidson's view Quine still accepts the *third* dogma of empiricism: The dualism of conceptual scheme and empirical content (Davidson 1974b, 189). As Davidson explains, this third dogma is “the last” dogma of empiricism, “for if we give it up it is not clear that there is anything distinctive left to call empiricism” (Davidson 1974b, 189).

The third dogma is of course nothing but Sellars' Myth of the Given, namely that non-conceptual causal processes could somehow set a rational constraint on our beliefs, be it individually or “as a corporate body”. Davidson's (1983) point is that truth and knowledge are notions that belong to the space of reasons. Hence, they cannot be constrained by natural processes that operate within the space of causation and, Davidson (1983) concludes, *coherentism* as the claim that our knowledge and beliefs are not constrained by anything in the natural world follows directly.

In other words, I think Davidson is right in thinking that when Quine talks about “the tribunal of sense experience”, while conceiving sense experience as “the stimulation of [...] sensory receptors” (Quine 1969, 75), *i.e.*, as a *causal* process, the experience *cannot* set a *rational* (no doubt a minimal requirement of a “tribunal”) constraint on our statements about the world (cf. McDowell 1994, 132-133). McDowell's (1994) peculiar form of objective idealism is motivated precisely by this dilemma: The choice between Davidson's coherentism and Quine's dogmatic Myth of the Given seems to be all we have. But Davidson's view, according to which nothing outside the sphere of thought can *rationally* constrain us and hence the mind-independent world is irrelevant to our *beliefs* about it, is simply an intolerable version of anti-realism. After all, “[i]f our activity in empirical thought and judgement is to be recognizable as bearing on reality at all, there must be [rational] external constraint” (McDowell 1994, 9). However, the other horn of the dilemma, Quine's Myth of the Given, is but an unintelligible dogma.

To avoid the dilemma, McDowell thinks, requires that we admit that somehow the world itself belongs to the space of reasons, *i.e.*, is rational or conceptual in nature: “The constraint [on our beliefs about the world] comes from outside *thinking*, but not from outside what is *thinkable*. When we trace justification [of the contents of our beliefs about the world] back, the last thing we come to is still a thinkable content; not something more ultimate than that, a bare pointing to a bit of the Given.” (McDowell 1994, 28-29.)

In short, McDowell’s (1994) argument is this: Either there are reasons for empirical judgements or there are not. If there are no reasons for empirical judgements, judgements cannot be rational and we cannot be responsible for them and thus, strictly speaking, they are not contentful or meaningful at all. If there are reasons for empirical judgements, they are either conceptual or they are not. If they are not conceptual, they cannot rationally constrain our empirical judgements on the pain of falling back to the unintelligible Myth of the Given. Hence they must be conceptual; but then they seem to belong to the space of reasons and not to the empirical world. This “threatens to make what was meant to be empirical thinking degenerate [...] into a frictionless spinning in a void” (McDowell 1994, 66), committing us to Davidson’s intolerable coherentism.<sup>201</sup> All options are bad. McDowell thinks that the only tolerable possibility we have is to admit that there are reasons for empirical judgements, these reasons are conceptual but they do not belong merely to the realm of thoughts because also the mind-independent world contains conceptual contents, namely *facts*. In a sense, McDowell is arguing for a return to the world-view of the ancient Greeks, who thought that the *cosmos* is essentially and intrinsically rational, or to Hegel’s idealism, according to which what is real is also rational.

---

<sup>201</sup> Davidson tries to get out of the problems of this position by arguing for the principle of charity, *i.e.*, that we can assume that most of our beliefs are true. But, as McDowell (1994, 68) correctly says, this comes too late: his coherentism has already endangered the whole idea of empirical content – it is not clear that in his position it makes sense to talk about contents, not to mention their being true or false. More recently (see the essays in Davidson 2001), Davidson has aimed to restore empirical objectivity in terms of “triangulation”, *i.e.*, in terms of autonomous subjects who are engaged in mutual interpretation. Again, McDowell rejects this, and rightly so: “if subjects are already in place, it is too late to set about catering for the constitution of the concept of objectivity. We must take subjectivity and the concept of objectivity to emerge together, out of initiation into the space of reasons.” (McDowell 1994, 186.) Similarly, since Davidson’s theory of language builds on *interpretation*, Davidson is bound to either beg the question by assuming the interpreter to possess mysterious intrinsic intentionality or render intentionality and meaning impossible, for no account building essentially on interpretation can bypass the Regress Argument (see III.3.2 above and Williams 2000).

In this study, however, I have argued for another way<sup>202</sup> of avoiding the Dilemma between anti-realism and the Myth of the Given (and I believe that at least Brandom (1994 & 2000) and Esfeld (2001) are largely in agreement with what I say). As so often in this dissertation, the key to an acceptable solution is the rejection of methodological individualism and truth-conditional semantics. We must reject the Cartesian setting of an asocial agent struggling to gain an epistemic access to the world. Rather, what is primitive and fundamental is a social bedrock practice within which there can be intentional agents and objects of empirical knowledge in the first place. As I have explained, the starting point here is the view that we are simply disposed to *causally* react to certain input by a certain output much as Quinean naturalism suggests. Indeed, precisely this was needed for solving the Infinity Problem.

Now for McDowell this is not enough. He thinks (correctly) that *if all we have* is such a dispositional theory of empirical experience, we cannot be rationally responsible for our empirical observations. In the Quinean picture empirical observations are mere causal reactions that happen to us, and thus they are not normatively or conceptually (indeed, logically) to propositions and actions – and thus, according to Quine’s own meaning holism, they cannot be *contentful* at all. However, just as in the case of rule-following the dispositional move is only the first part of the solution, similarly also here the causal disposition only serves to bring the new, empirical aspect into our language-game (hence Sellars and Brandom call these language-entry moves). As I have argued, within social practices a causal reaction may *count as* undertaking a normative commitment, if the others accept it as such. The content of the commitment is determined by the norms (implicit in social practices) governing the inferential relations that define the other commitments such undertaking of a commitment brings with it. Accordingly, the rational responsibility comes from the *continuing acceptance of the commitment* when we figure out what the commitment *means* (when we see its place in the normative web of our other commitments).

Similarly, in the context of rule-following considerations the dispositional aspect of my solution explains only the *forming* of a commitment (a *judging*); the causal story does not give us *contents* of commitments (what is *judged*), since, as I have explained, that requires a *social practice* of assessing the *de dicto* *judgings* of others in the *de re* sense of what is *judged*. The social aspect is required for assessing whether the performed language-entry move was justified in the circumstances. As I have said already

---

<sup>202</sup> Besides as an alternative to McDowell, my theory could also be seen as a *naturalisation* of McDowell’s line of thought, for I explain how physical objects can receive a normative status (as opposed to being intrinsically normative).

so many times, a causal reaction (a disposition) is not all there is to content of empirical statements (because it runs to the Normativity Problem), but it is the starting point (because it is needed for solving the Infinity Problem). And the social aspect that resolves the Normativity Problem is also the aspect that is required for representational (empirical) content.

Thus, although I do build on causal dispositions, I think McDowell's criticism of Quine does not apply to the view I defend:

The only connection he [Quine] countenances between experience and the acceptance of statements is a brutally causal linkage that subjects are conditioned into when they learn a language. It is not that it is right to revise one's belief system thus and so in the light of such-and-such an experience, but just that that revision is what would probably happen if one's experience took that course. Quine conceives experiences so that they can only be outside the space of reasons, the order of justification.  
(McDowell 1994, 133.)

The key phrase is: The *only* connection. I think the *causal* link really is like this, namely that "thus and so [...] is what would probably happen if one's experience took that course". Such a dispositional link is required for solving the Infinity Problem. But I add a *social* aspect to solve the Normativity Problem, and it is the social aspect, *missing from Quine*, that grounds the idea that "it is right to revise one's belief system thus and so in the light of such-and-such experience". Only when the causal reaction of an individual is embedded in a social practice it gains a content. Experiences qua contentful belong to a social practice, and within a social practice they are integrated into the space of reasons and the order of justification. As I have emphasised in Part III, acceptable holism must combine Quinean holism with social holism.

Hence, in order to overcome the unacceptable options McDowell discusses, we must reject the Cartesian idea of taking the notion of an autonomous subject capable of representations as primitive. An individual cannot by herself have contentful thoughts and, consequently, an epistemic access to the world. She needs something greater than herself to make these marvels, indispensable for the human form of life, possible. Surprisingly, this is where my line of thought meets Descartes' reasoning. What I have in mind is Descartes' philosophical motivation for theism, which ultimately lies in the inescapable fact that the capacities of an individual are dreadfully limited. I have no quarrel with this. But I reject Descartes' God; nor have I metaphysical taste for Leibniz's. My God is that of Durkheim. As Kusch explains:

Descartes obviously did not have [...] concept of social institution, and therefore had to anchor stability of meaning elsewhere. As is well known, in Descartes' system it is God who makes meanings stable, truths persistent, and reality objective [...] yet there *is* a way to bridge the gap between Cartesius and us. The bridge is Durkheim's suggestion that in their gods societies celebrate themselves and their achievements. In celebrating his God, Cartesius was celebrating collectively sustained use [...].

(Kusch 1999, 363.)

And what a powerful god Durkheim's God is! Outside social practices there are no agents, no meanings, no mental contents, no representation and, of course, no possibility of realism.

There really is no serious alternative to sociophilosophy in the sense of naturalised methodological holism.

## REFERENCES

Abell, Peter

- 2000 "Sociological Theory and Rational Choice Theory" in Bryan S. Turner (ed.), *The Blackwell Companion to Social Theory*, Oxford, Blackwell, 223-44.

Anderson, Elizabeth

- 2001 "Unstrapping the Straitjacket of 'Preference': A Comment on Amartya Sen's Contributions to Philosophy and Economics", *Economics and Philosophy* 17, 21-38.

Anscombe, G. E. M.

- 1959 *Intention*, Oxford, Blackwell.

Axelrod, Robert A.

- 1984 *The Evolution of Cooperation*, New York, Basic Books.

Baker, G. P. & P. M. S. Hacker

- 1984 *Scepticism, Rules and Language*, Oxford, Basil Blackwell.

Barnes, Barry

- 1983 "Social Life as Bootstrapped Induction", *Sociology* 17(4), 524-545.

- 1995 *The Elements of Social Theory*, Princeton, Princeton University Press.

- 2000 *Understanding Agency: Social Theory and Responsible Action*, London, Sage Publications.

- 2001 "Practice as Collective Action" in Theodore R. Schatzki, Karin Knorr Cetina & Eike von Savigny (eds.), *The Practice Turn in Contemporary Theory*, London, Routledge, 17-28.

- 2002 "Searle on Social Reality: Process Is Prior to Product" in Günther Grewendorf & Georg Meggle (eds.), *Speech Acts, Mind, and Social Reality: Discussions with John R. Searle*, Dordrecht, Kluwer, 247-257.

Barnes, Barry & David Bloor

- 1982 "Relativism, Rationalism and the Sociology of Knowledge" in Martin Hollis & Steven Lukes (eds.), *Rationality and Relativism*, Oxford, Basil Blackwell, 21-47.

Barnes, Barry; David Bloor & John Henry

- 1996 *Scientific Knowledge: A Sociological Analysis*, London, Athlone.

Bhargava, Rajeev

- 1992 *Individualism in Social Science: Forms and Limits of a Methodology*, Oxford, Clarendon Press.

Bilgrami, Akeel

- 1992 *Belief and Meaning: The Unity and Locality of Mental Content*, Oxford, Blackwell.

Binmore, Ken

- 1994 *Game Theory and the Social Contract Volume I: Playing Fair*, Cambridge MA, The MIT Press.
- 1999 "Why Experiment in Economics?", *The Economic Journal* 109, F16-F24.

Binmore, Ken; Avner Shaked & John Sutton

- 1985 "Testing Noncooperative Bargaining Theory: A Preliminary Study", *The American Economic Review* 75(5), 1178-1180.

Black, Max

- 1989 "Some Remarks about 'Practical Reasoning'" in Schilpp & Hahn (1989), 405-416.

Blackburn, Simon

- 1984 "The Individual Strikes Back" in Miller & Wright (eds.) 2002, 28-44. Originally in *Synthese* 58, 1984, 281-301.

Bloor, David

- 1983 *Wittgenstein: A Social Theory of Knowledge*, London, Macmillan Education.
- 1996 "Idealism and the Sociology of Knowledge", *Edinburgh Working Papers in Sociology* 5, 1-24.
- 1997 *Wittgenstein, Rules and Institutions*, London, Routledge.
- 2001 "Wittgenstein and the Priority of Practice" in Theodore R. Schatzki, Karin Knorr Cetina & Eike von Savigny (eds.), *The Practice Turn in Contemporary Theory*, London, Routledge, 95-106.

Boghossian, Paul A.

- 1989 "The Rule-Following Considerations" in Miller & Wright (eds.) 2002, 141-187. Originally in *Mind* 98, 1989, 507-549.

Bowles, Samuel

- 2006 "Group Competition, Reproductive Leveling, and the Evolution of Human Altruism", *Science* 8 (December 2006), 1569-1572.

Brandom, Robert B.

- 1994 *Making it Explicit: Reasoning, Representing, and Discursive Commitments*, Cambridge MA, Harvard University Press.
- 2000 *Articulating Reasons: An Introduction to Inferentialism*, Cambridge MA, Harvard University Press.

- 2002 *Tales of the Mighty Dead: Historical Essays in the Metaphysics of Intentionality*, Cambridge MA, Harvard University Press.
- Bratman, Michael E.
- 1999 *Faces of Intention: Selected Essays on Intention and Agency*, Cambridge, Cambridge University Press.
- Brewer, Marilynn B. & Sherry K. Schneider
- 1990 "Social Identity and Social Dilemmas: A Double-Edged Sword" in Dominic Abrams & Michael A. Hogg (eds.), *Social Identity Theory: Constructive and Critical Advances*, Hemel Hempstead, Harvester Wheatsheaf.
- Brown, Rupert
- 2000 "Social Identity Theory: Past Achievements, Current Problems and Future Challenges", *European Journal of Social Psychology* 30(6), 745-778.
- Burge, Tyler
- 1979 "Individualism and the Mental", *Midwest Studies in Philosophy* 4, 73-121.
- Carroll, Lewis
- 1895 "What the Tortoise Said to Achilles", *Mind* 4(14), 278-280.
- Cartwright, Nancy
- 1999 *The Dappled World: A Study of the Boundaries of Science*, Cambridge, Cambridge University Press.
- Chydenius, Anders
- 1765/1994 *The National Gain* (with an introduction by Georg Schauman and Matti Klinge, translator unknown), *s.l.*, Hanprint. Originally published in 1765 as *Den Nationnale Winsten*.
- Coates, Paul
- 1986 "Kripke's Sceptical paradox: Normativeness and Meaning", *Mind* 95(377), 77-80.
- 1997 "Meaning, Mistake and Miscalculation", *Minds and Machines* 7, 171-197.
- Collin, Finn
- 1997 *Social Reality*, London, Routledge.
- Cowen, Tyler
- 1998 "Do Economists Use Social Mechanisms to Explain?" in Peter Hedström & Richard Swedberg (eds.), *Social Mechanisms – An Analytical Approach to Social Theory*, Cambridge, Cambridge University Press, 125-146.



Crane, Tim

1995 "The Mental Causation Debate", *Proceedings of the Aristotelian Society*, Supplementary Volume 69, 211-236.

Cubitt, Robin P; Chris Starmer & Robert Sugden

2001 "Discovered Preferences and the Experimental Evidence of Violations of Expected Utility Theory", *Journal of Economic Methodology* 8(3), 385-414.

Davidson, Donald

1963 "Actions, Reasons, and Causes" in Davidson 1980, 3-19. Originally in *Journal of Philosophy* 60, 1963.

1967 "Causal Relations" in Davidson 1980, 149-162. Originally in *Journal of Philosophy* 64, 1967.

1970 "Mental Events" in Davidson 1980, 207-227. Originally in Lawrence Foster and J. W. Swanson (eds.), *Experience and Theory*, The University of Massachusetts Press and Duckworth, 1970.

1973a "Freedom to Act" in Davidson 1980, 63-81. Originally in Ted Honderich (ed.), *Essays on Freedom of Action*, London, Routledge & Kegan Paul, 1973.

1973b "Radical Interpretation" in Davidson 1984, 125-139. Originally in *Dialectica* 27, 1973.

1974a "Psychology as Philosophy" in Davidson 1980, 229-244. Originally in S. C. Brown (ed.), *Philosophy of Psychology*, The Macmillan Press and Barnes & Noble Books, 1974.

1974b "On the Very Idea of a Conceptual Scheme" in Davidson 1984, 183-198. Originally in *Proceedings and Addresses of the American Philosophical Association* 47, 1974.

1976 "Hempel on Explaining Action" in Davidson 1980, 261-275. Originally in *Erkenntnis* 10, 1976.

1980 *Essays on Actions and Events*, Oxford, Clarendon Press.

1982 "Paradoxes of Irrationality" in Davidson 2004, 169-187. Originally in R. Wollheim & J. Hopkins (eds), *Philosophical Essays on Freud*, Cambridge, Cambridge University Press, 1982.

1983 "A Coherence Theory of Truth and Knowledge" in Davidson 2001, 137-153. Originally in Dieter Henrich (ed.), *Kant oder Hegel? über Formen der Begründung in der Philosophie: Stuttgarter Hegel-Kongress 1981*, Stuttgart, Klett-Cotta, 1983.

- 1984 *Inquiries into Truth and Interpretation*, Oxford, Clarendon Press.
- 1987 “Problems in the Explanation of Action” in Davidson 2004, 101-116. Originally in P. Pettit, R. Sylvan & J. Norman (eds), *Metaphysics and Morality: Essays in Honour of J. J. C. Smart*, Oxford, Blackwell, 1987.
- 1995 “Could There Be a Science of Rationality?” in Davidson 2004, 117-134. Originally in *International Journal of Philosophical Studies* 3, 1995.
- 2001 *Subjective, Intersubjective, Objective*, Oxford, Clarendon Press.
- 2004 *Problems of Rationality*, Oxford, Clarendon Press.
- Dennett, Daniel C.
- 1991a *Consciousness Explained*, London, Allen Lane.
- 1991b “Real Patterns”, *The Journal of Philosophy* 88(1), 27-51.
- 2003 *Freedom Evolves*, London, Allen Lane.
- Dretske, Fred
- 1988 *Explaining Behavior: Reasons in a World of Causes*, Cambridge MA, the MIT Press.
- 1989 “Reasons and Causes” in James E. Tomberlin (ed.), *Philosophical Perspectives 3: Philosophy of Mind and Action Theory*, Atascadero, Ridgeview, 47-76.
- Drury, John & Steve Reicher
- 2000 “Collective Action and Psychological Change: The Emergence of New Social Identities”, *British Journal of Social Psychology* 39, 579-604.
- Dunbar, Robin I. M.
- 2002 “Brains on Two Legs: Group Size and the Evolution of Intelligence” in Frans B. M. de Waal (ed.), *Tree of Origin: What Primate Behavior Can Tell Us about Human Social Evolution*, Cambridge MA, Harvard University Press, 173-191.
- Dupré, John
- 1993 *The Disorder of Things: Metaphysical Foundations of the Disunity of Science*, Cambridge MA, Harvard University Press.
- 2001 *Human Nature and the Limits of Science*, Oxford, Clarendon Press.
- 2002 *Humans and Other Animals*, Oxford, Clarendon Press.

Durkheim, Emile

1895/1938 *The Rules of Sociological Method* (translated by Sarah A. Solovay & John H. Mueller, edited by George E. G. Catlin), Glencoe, The Free Press, 1938, 3<sup>rd</sup> ed. 1962. Originally published in 1895 as *Les Regles de la Methode Sociologique*.

Elster, Jon

1985 "The Nature and Scope of Rational-Choice Explanation", in Michael Martin & Lee C. McIntyre (eds.), *Readings in the Philosophy of Social Science*, Cambridge MA, The MIT Press, 311-322, 1994. Originally in Ernest LePore & Brian McLaughlin (eds.), *Actions and Events: Perspectives on Donald Davidson*, Oxford, Blackwell, 60-72, 1985.

1989 *Nuts and Bolts for the Social Sciences*, Cambridge, Cambridge University Press.

Esfeld, Michael

1999 "Rule-Following and the Ontology of the Mind", in Uwe Meixner & Peter M. Simons (eds.), *Metaphysics in the Post-Metaphysical Age. Papers of the 22<sup>nd</sup> International Wittgenstein Symposium*, Kirchberg, Austrian Ludwig Wittgenstein Society, 191-196.

2001 *Holism in Philosophy of Mind and Philosophy of Physics*, Dordrecht, Kluwer.

Fehr, Ernst & Herbert Gintis

2007 "Human Motivation and Social Cooperation: Experimental and Analytical Foundations", *Annual Review of Sociology* 33, forthcoming (available at <http://www-unix.oit.umass.edu/~gintis/>).

Fodor, Jerry

1974 "Special Sciences (or: The Disunity of Science as a Working Hypothesis)" in Michael Martin & Lee C. McIntyre (eds.), *Readings in the Philosophy of Social Science*, Cambridge MA, The MIT Press, 1994, 687-699. Originally in *Synthese* 28, 1974, 97-115.

1990 *A Theory of Content and Other Essays*, Cambridge MA, the MIT Press.

Foote, Nelson N.

1951 "Identification as the Basis for a Theory of Motivation", *American Sociological Review* 16(1), 14-21.

Forbes, Graeme

- 1983-4 "Scepticism and Semantic Knowledge" in Miller & Wright (eds.) 2002, 16-27. Originally in *Proceedings of the Aristotelian Society* 1983-4, 223-237.

Frank, Robert H.

- 1988 *Passions within Reason: The Strategic Role of the Emotions*, New York, W. W. Norton & Company.

Frank, Robert H; Thomas Gilovich & Dennis T. Regan

- 1993 "Does Studying Economics Inhibit Cooperation?", *The Journal of Economic Perspectives* 7(2), 159-171.

Føllesdal, Dagfinn

- 1982 "The Status of Rationality Assumptions in Interpretation and the Explanation of Action" in Michael Martin & Lee C. McIntyre (eds.), *Readings in the Philosophy of Social Science*, Cambridge MA, The MIT Press, 299-310, 1994  
Originally in *Dialectica* 36, 1982, 301-316,

Gauthier, David

- 1986 *Morals by Agreement*, Oxford, Clarendon Press.

Giddens, Anthony

- 1984 *The Constitution of Society: Outline of the Theory of Structuration*, Cambridge, Polity Press.

Gilbert, Margaret

- 1989 *On Social Facts*, London, Routledge.
- 1996 *Living Together: Rationality, Sociality, and Obligation*, Lanham MD, Rowman and Littlefield.
- 2000 *Sociality and Responsibility: New Essays in Plural Subject Theory*, Lanham MD, Rowman and Littlefield.
- 2002 "Belief and Acceptance as Features of the Group", *ProtoSociology* 16, 35-69.
- 2003 "The Structure of the Social Atom: Joint Commitment as the Foundation of Human Social Behavior" in Frederick F. Schmitt (ed.), *Socializing Metaphysics: The Nature of Social Reality*, Lanham MD, Rowman and Littlefield, 39-64.
- 2006 *A Theory of Political Obligation: Membership, Commitment, and the Bonds of Society*, Oxford, Oxford University Press.

2007 “Searle and Collective Intentions”, in Savas L. Tsohatzidis (ed.), *Intentional Acts and Institutional Facts: Essays on John Searle’s Social Ontology*, Berlin, Springer, forthcoming.

Glüer, Kathrin & Åsa Maria Wikforss

2006 “Against Content Normativity”, unpublished manuscript, available at <http://people.su.se/~kgl/kathrin.htm>, 1-32.

Glüer, Kathrin & Peter Pagin

1998 “Rules of Meaning and Practical Reasoning”, *Synthese* 117(2), 207-227.

Goodman, Nelson

1973 *Fact, Fiction and Forecast* (3<sup>rd</sup> Edition), Indianapolis, Bobbs-Merrill.

Gross, Neil

2006 “Comment on Searle”, *Anthropological Theory* 6(1), 45-56.

Hacking, Ian

1984 “Five Parables” in *Historical Ontology*, Cambridge MA, Harvard University Press, 2002, 27-50. Originally in Richard Rorty, Jerry Schneewind and Quentin Skinner (eds.), *Philosophy in its Context*, Cambridge, Cambridge University Press, 1984, 103-124.

1986 “Making Up People” in *Historical Ontology*, Cambridge MA, Harvard University Press, 2002, 99-114. Originally in Thomas Heller, Morton Sosna & David Wellberry (eds.), *Reconstructing Individualism*, Stanford, Stanford University Press, 1986, 222-236.

1995a *Rewriting the Soul: Multiple Personality and the Sciences of Memory*, Princeton, Princeton University Press.

1995b “The Looping Effects of Human Kinds” in Dan Sperber, David Premack & Ann James Premack (eds.), *Causal Cognition: A Multidisciplinary Debate*, Oxford, Clarendon Press, 351-383.

1997 “John Searle’s Building Blocks”, *History of the Human Sciences* 10, 83-92.

1999 *The Social Construction of What?*, Cambridge MA, Harvard University Press.

Hakli, Raul

2006 “Group Beliefs and the Distinction Between Belief and Acceptance”, *Cognitive Systems Research* 7, 286-297.

Halfpenny, Peter

2001 “Positivism in the Twentieth Century” in George Ritzer & Barry Smart (eds.), *Handbook of Social Theory*, London, Sage, 371-385.

Hanson, Norwood Russell

- 1958 *Patterns of Discovery: an Inquiry into the Conceptual Foundations of Science*, Cambridge, Cambridge University Press.

Hardin, Russell

- 1982 *Collective Action*, Baltimore, The Johns Hopkins University Press for Resources for the Future.

Hare, Brian & Michael Tomasello

- 2004 "Chimpanzees Are More Skilful in Competitive than in Cooperative Cognitive Tasks", *Animal Behaviour* 68(3), 571-581.

Haugeland, John

- 1990 "The Intentionality All-Stars", *Philosophical Perspectives* 4, 383-427.

Haukioja, Jussi

- 2000 *Rule-Following, Response-Dependence and Realism*, Turku, Turku University Press (Reports from the Department of Philosophy).

Hausman, Daniel M.

- 2000 "Revealed Preference, Belief, and Game Theory", *Economics and Philosophy* 16, 99-115.

Henrich, Joseph; Robert Boyd; Sam Bowles; Colin Camerer; Herbert Gintis; Richard McElreath & Ernst Fehr

- 2001 "In Search of Homo Economicus: Experiments in 15 Small-Scale Societies", *The American Economic Review* 91(2), 73-79.

Hogg, Michael A. & Dominic Abrams

- 1988 *Social Identifications: A Social Psychology of Intergroup Relations and Group Processes*, London, Routledge.

Hollis, Martin

- 1994 *The Philosophy of Social Science: An Introduction*, Cambridge, Cambridge University Press.

- 1996 *Reason in Action: Essays in the Philosophy of Social Science*, Cambridge, Cambridge University Press.

Hornsby, Jennifer

- 1997 "Collectives and Intentionality", *Philosophy and Phenomenological Research* 57, 429-434.

Horwich, Paul

- 1995 "Meaning, Use and Truth", *Mind* 104(414), 355-368.

Jackson, Frank; Philip Pettit and Michael Smith

2004 *Mind, Morality, and Explanation: Selected Collaborations*, Oxford, Clarendon Press.

Kagel, John H. & Alvin E. Roth (eds.)

1995 *The Handbook of Experimental Economics*, Princeton, Princeton University Press.

Kerr, Norbert L. & Ernest S. Park

2001 "Group Performance in Collaborative and Social Dilemma Tasks: Progress and Prospects" in Michael A. Hogg & R. Scott Tindale (eds.), *Blackwell Handbook of Social Psychology: Group Processes*, Oxford, Blackwell, 107-138.

Kim, Jaegwon

1984 "Epiphenomenal and Supervenient Causation" in Kim 1993, 92-108. Originally in *Midwest Studies in Philosophy* 9, 1984, 257-270.

1985 "Psychophysical Laws" in Kim 1993, 194-215. Originally in Ernest LePore & Brian McLaughlin (eds.), *Actions and Events: Perspectives on the Philosophy of Donald Davidson*, Oxford, Basil Blackwell, 1985, 369-386.

1989a "Mechanism, Purpose, and Explanatory Exclusion" in Kim 1993, 237-264. Originally in James E. Tomberlin (ed.), *Philosophical Perspectives 3: Philosophy of Mind and Action Theory*, Atascadero, Ridgeview Publishing, 77-108, 1989.

1989b "The Myth of Nonreductive Materialism" in Kim 1993, 265-284. Originally in *Proceedings and Addresses of the American Philosophical Association* 63, 1989, 31-47.

1991 "Dretske on How Reasons Explain Behavior" in Kim 1993, 285-306. Originally in Brian McLaughlin (ed.), *Dretske and His Critics*, Oxford, Basil Blackwell, 1991, 52-72.

1993 *Supervenience and Mind: Selected Philosophical Essays*, Cambridge, Cambridge University Press.

1998 *Mind in a Physical World: An Essay on the Mind-Body Problem and Mental Causation*, Cambridge MA, The MIT Press.

2003 "Philosophy of Mind and Psychology" in Kirk Ludwig (ed.), *Donald Davidson*, Cambridge, Cambridge University Press, 113-136.

2005 *Physicalism, or Something Near Enough*, Princeton, Princeton University Press.

Kitcher, Philip

- 1998 “Psychological Altruism, Evolutionary Origins, and Moral Rules”, *Philosophical Studies* 89, 283-316.

Klandermands, Bert

- 2000 “Identity and Protest: How Group Identification Helps to Overcome Collective Action Dilemmas” in Mark van Vugt, Mark Snyder, Tom R. Tyler & Anders Biel (eds.), *Cooperation in Modern Society: Promoting the Welfare of Communities, States and Organizations*, London, Routledge, 162-183.

Knudsen, Thorbjørn

- 2004 “General Selection Theory and Economic Evolution: The Price Equation and the Replicator / Interactor Distinction”, *Journal of Economic Methodology* 11(2), 147-173.

Kollock, Peter

- 1998 “Transforming Social Dilemmas: Group Identity and Co-operation” in Peter A. Danielson (ed.), *Modeling Rationality, Morality, and Evolution*, Oxford, Oxford University Press, 185-209.

Kripke, Saul A.

- 1982 *Wittgenstein on Rules and Private Language*, Oxford, Basil Blackwell.

Kusch, Martin

- 1997 “The Sociophilosophy of Folk Psychology”, *Studies in History and Philosophy of Science* 28(1), 1-25.
- 1999 *Psychological Knowledge: A Social History and Philosophy*, London, Routledge.
- 2002 *Knowledge by Agreement: The Programme of Communitarian Epistemology*, Oxford, Clarendon Press.
- 2003 “Explanation and Understanding: The Debate over von Wright’s Philosophy of Action Revisited” in Leila Haaparanta & Ilkka Niiniluoto (eds.), *Analytic Philosophy in Finland (Poznan Studies in the Philosophy of Science and the Humanities 80)*, Amsterdam, Rodopi, 327-353.
- 2004 “Rule-Scepticism and the Sociology of Scientific Knowledge: The Bloor-Lynch Debate Revisited”, *Social Studies of Science* 34(4), 571-591.
- 2006 *A Sceptical Guide to Meaning and Rules: Defending Kripke’s Wittgenstein*, Chesham, Acumen.



Larson, James L.

- 1971 *Reason and Experience: The Presentation of Natural Order in the Work of Carl von Linné*, Berkeley, University of California Press.

Lukes, Steven M.

- 1982 "Introduction" in *Emile Durkheim: The Rules of Sociological Method and Selected Texts on Sociology and its Method*, London, Macmillan, 1-27.
- 2007 "Searle versus Durkheim", in Savas L. Tsohatzidis (ed.), *Intentional Acts and Institutional Facts: Essays on John Searle's Social Ontology*, Berlin, Springer, forthcoming.

Malcolm, Norman

- 1989 "Intention and Behavior" in Schilpp & Hahn (1989), 353-376.

McClennen, Edward F.

- 1997 "Pragmatic Rationality and Rules", *Philosophy and Public Affairs* 26(3), 210-258.

McClennen, Edward F. & Scott Shapiro

- 1998 "Rule-Guided Behaviour" in Peter Newman (ed.), *The New Palgrave Dictionary of Economics and the Law* (Vol. 3), London, Macmillan Reference, 363-369.

McDowell, John

- 1994 *Mind and World* (with a new introduction 1996), Cambridge MA, Harvard University Press, 2003.
- 1998a *Mind, Value, and Reality*, Cambridge MA, Harvard University Press.
- 1998b "Having the World in View: Sellars, Kant, and Intentionality", *The Journal of Philosophy* 95(9), 431-491.

Melden, A. I.

- 1960 "Willing" in Alan R. White (ed.), *The Philosophy of Action*, London, Oxford University Press, 1968, 70-78. Originally in *Philosophical Review* 69, 1960, 475-484.

Mele, Alfred R.

- 2003 "Philosophy of Action" in Kirk Ludwig (ed.), *Donald Davidson*, Cambridge, Cambridge University Press, 64-84.

Mellor, D. H.

- 1982 "The Reduction of Society", *Philosophy* 57, 51-75.

Mill, John Stuart

- 1865 *A System of Logic: Ratiocinative and Inductive, Being a Connected View of the Principles of Evidence, and the Methods of Scientific Investigation*, 8<sup>th</sup> Edition 1925, London, Longman, Green, and Co.

Miller, Alexander & Crispin Wright (eds.)

- 2002 *Rule-Following and Meaning*, Chesham, Acumen.

Miller, Seumas

- 2001 *Social Action: A Teleological Account*, New York, Cambridge University Press.
- 2007 "Joint Action: An Individualistic Account" in Savas L. Tsohatzidis (ed.), *Intentional Acts and Institutional Facts: Essays on John Searle's Social Ontology*, Berlin, Springer, forthcoming.

Millikan, Ruth Garrett

- 1990 "Truth Rules, Hoverflies, and the Kripke-Wittgenstein Paradox" in Miller & Wright (eds.) 2002, 209-233. Originally in *The Philosophical Review* 99(3), 1990, 323-353.

Mäki, Uskali

- 1996 "Scientific Realism and Some Peculiarities of Economics" in Robert S. Cohen, Risto Hilpinen & Qiu Renzong (eds.), *Realism and Anti-realism in the Philosophy of Science: Beijing International Conference 1992*, Dordrecht, Kluwer, 427-447.

Nagel, Ernest

- 1961 *The Structure of Science*, New York, Harcourt Brace.

Niiniluoto, Ilkka

- 1999 "Rule-Following, Finitism, and the Law", *Associations* 3(1), 83-90.

Okasha, Samir

- 2001 "Why Won't the Group Selection Controversy Go Away?", *The British Journal for the Philosophy of Science* 52, 25-50.

O'Neill, John (ed.)

- 1973 *Modes of Individualism and Collectivism*, London, Heinemann.

Pagin, Peter

- 2002 "Rule-Following, Compositionality and the Normativity of Meaning" in Dag Prawitz (ed.), *Meaning and Interpretation*, Stockholm, Almqvist & Wiksell. Now at <http://people.su.se/~ppagin/pagineng.htm>, 1-34.

Papineau, David

- 1999 "Normativity and Judgement", *Proceedings of the Aristotelian Society*, suppl. 73, 17-44.
- 2002 *Thinking about Consciousness*, Oxford, Clarendon Press.

Peacocke, Christopher

- 1981 "Rule-Following: The Nature of Wittgenstein's Arguments" in Steven H. Holtzman & Christopher M. Leich (eds.), *Wittgenstein: To Follow a Rule*, London, Routledge & Kegan Paul, 72-95.

Pettit, Philip

- 1990 "The Reality of Rule-Following" in Pettit 2002, 26-48. Originally in *Mind* 99, 1-21, 1990.
- 1993 *The Common Mind: An Essay on Psychology, Society, and Politics*, Oxford, Oxford University Press.
- 1996 "Three Aspects of Rational Explanation" in Pettit 2002, 177-191. Originally in *ProtoSociology* 8-9, 1996, 170-183.
- 2002 *Rules, Reasons, and Norms: Selected Essays*, Oxford, Clarendon Press.
- 2003 "Groups with Minds of Their Own" in Frederick F. Schmitt (ed.), *Socializing Metaphysics: The Nature of Social Reality*, Lanham, Rowman & Littlefield, 167-193.

Pettit, Philip & David Schweikard

- 2006 "Joint Actions and Group Agents", *Philosophy of the Social Sciences* 36(1), 40-66.

Plott, Charles R.

- 1996 "Rational Individual Behaviour in Markets and Social Choice Processes: the Discovered Preference Hypothesis" in Kenneth J. Arrow, Enrico Colombaro, Mark Perlman & Christian Schmidt (eds.), *The Rational Foundations of Economic Behaviour: Proceedings of the IEA Conference held in Turin, Italy*, Houndmills & London, Macmillan Press, 225-254.

Pusey, Anne E.

- 2002 "Of Genes and Apes: Chimpanzee Social Organization and Reproduction" in Frans B. M. de Waal (ed.), *Tree of Origin: What Primate Behavior Can Tell Us about Human Social Evolution*, Cambridge MA, Harvard University Press, 9-37.

## Putnam, Hilary

- 1967 "The Nature of Mental States" in *Mind, Language, and Reality*, Cambridge, Cambridge University Press, 1975, 429-440. Originally as "Psychological Predicates" in W. Capitan & D. Merrill (eds.), *Art, Mind, and Religion*, Pittsburgh, University of Pittsburgh Press, 1967, 37-48.
- 1975 "The Meaning of 'Meaning'" in K. Gunderson (ed.), *Language, Mind and Knowledge: Minnesota Studies in the Philosophy of Science 7*, Minneapolis, University of Minnesota Press.
- 1981 *Reason, Truth and History*, Cambridge, Cambridge University Press.

## Quine, Willard Van Orman

- 1951 "Two Dogmas of Empiricism", *Philosophical Review* 60, 20-43.
- 1969 "Epistemology Naturalized" in *Ontological Relativity and Other Essays*, New York, Columbia University Press.

## Quinton, Anthony

- 1975-76 "Social Objects", *Proceedings of the Aristotelian Society* 75, 1-27.

## Rachels, James

- 1986 *The Elements of Moral Philosophy*, Philadelphia, Temple University Press.

## Rakoczy, Hannes &amp; Michael Tomasello

- 2007 "The Ontogeny of Social Ontology: Steps to Shared Intentionality and Status Functions" in Savas L. Tsohatzidis (ed.), *Intentional Acts and Institutional Facts: Essays on John Searle's Social Ontology*, Berlin, Springer, forthcoming.

## Ridley, Matt

- 1996 *The Origins of Virtue*, London, Viking.

## Rorty, Richard

- 1979 *Philosophy and the Mirror of Nature*, Princeton, Princeton University Press.

## Rosenberg, Alexander

- 1995 *Philosophy of Social Science* (2<sup>nd</sup> Edition), Boulder, Westview Press.

## Ross, Don

- 2002 "Dennettian Behavioural Explanations and the Roles of the Social Sciences" in Andrew Brook & Don Ross (eds.), *Daniel Dennett*, Cambridge, Cambridge University Press, 140-183.

Routledge, Bryan R.

- 1998 "Economics of the Prisoner's Dilemma: A Background" in Peter A. Danielson (ed.), *Modeling Rationality, Morality, and Evolution*, Oxford, Oxford University Press, 92-118.

Rovane, Carol

- 1998 *The Bounds of Agency: An Essay in Revisionary Metaphysics*, Princeton, Princeton University Press.

Ruben, David-Hillel

- 1985 *The Metaphysics of the Social World*, London, Routledge & Kegan Paul.

Saaristo, Antti

- 2003 "On the Objectivity of Social Facts", *ProtoSociology* 18-19, 291-316.
- 2004a "Personhood as a Social Status" in Heikki Ikäheimo, Jussi Kotkavirta, Arto Laitinen and Pessi Lyyra (eds.), *Personhood* (Publications in Philosophy 68), University of Jyväskylä, 192-198.
- 2004b "Merkitys sosiaalisena käytäntönä: kohti ihmistieteiden filosofisia perusteita" ("Meaning as Social Practice: Towards the Philosophical Foundations of the Human Sciences", in Finnish), *Ajatus* 61, 81-114.
- 2006a "Intentional Action and the Limits of Scientific Naturalism: Davidson's Alleged Refutation of the Logical Connection Argument" in Sami Pihlström, Heikki J. Koskinen and Risto Vilkkö (eds.), *Science – A Challenge to Philosophy?*, Frankfurt am Main, Peter Lang, 247-257.
- 2006b "There Is No Escape from Philosophy: Collective Intentionality and Empirical Social Science", *Philosophy of the Social Sciences* 36(1), 40-66.
- 2007 "On the Possibility of Naturalised Anti-Individualism in Social Ontology" in Gurol Irzik (ed.), *Logic and Philosophy of Science (Proceedings of the 21<sup>st</sup> World Congress of Philosophy: Philosophy Facing World Problems, Volume V)*, Ankara, the Philosophical Society of Turkey, forthcoming.

Scheff, Thomas J.

- 1990 *Microsociology: Discourse, Emotion, and Social Structure*, Chicago, Chicago University Press.

Schilpp, Paul Arthur & Lewis Edwin Hahn (eds.)

- 1989 *The Philosophy of Georg Henrik von Wright*, La Salle, Open Court.

Schmid, Hans Bernhard

- 2004 "Personhood and the Structure of Commitment" in Heikki Ikäheimo, Jussi Kotkavirta, Arto Laitinen and Pessi Lyyra (eds.), *Personhood* (Publications in Philosophy 68), University of Jyväskylä, 199-211.
- 2005 "Beyond Self-Goal Choice: Amartya Sen's Analysis of the Structure of Commitment and the Role of Shared desires", *Economics and Philosophy* 21, 51-63.

Searle, John R.

- 1983 *Intentionality: An Essay in the Philosophy of Mind*, Cambridge, Cambridge University press.
- 1990 "Collective Intentions and Actions" in Philip R. Cohen, Jerry Morgan & Martha E. Pollack (eds.), *Intentions in Communication*, Cambridge MA, the MIT Press, 401-415.
- 1995 *The Construction of Social Reality*, London, Penguin.
- 1997 "Responses to Critics of *The Construction of Social Reality*", *Philosophy and Phenomenological Research* 57, 449-458.
- 2006 "Searle versus Durkheim and the Waves of Thought: Reply to Gross", *Anthropological Theory* 6(1), 57-69.
- 2007a "What Is Language: Some Preliminary Remarks" in Savas L. Tsohatzidis (ed.), *John Searle's Philosophy of Language: Force, Meaning, and Mind*, Cambridge, Cambridge University Press, forthcoming.
- 2007b "Social Ontology: The Problem and Steps toward a Solution" in Savas L. Tsohatzidis (ed.), *Intentional Acts and Institutional Facts: Essays on John Searle's Social Ontology*, Berlin, Springer, forthcoming.

Sellars, Wilfrid

- 1963 *Science, Perception and Reality*, London, Routledge & Kegan Paul.

Sen, Amartya K.

- 1977 "Rational Fools: A Critique of the Behavioral Foundations of Economic Theory", *Philosophy and Public Affairs* 6, 317-344.
- 1985 "Goals, Commitment, and Identity", *Journal of Law, Economics, and Organization* 1(2), 341-355.

## Simmel, Georg

1908/1971 "How Is Society Possible?" (translated by Kurt H. Wolff) in *On Individuality and Social Forms*, Chicago, Chicago University Press, 1971, 6-22. Originally published in 1908 as "Exkurs über das Problem: Wie ist Gesellschaft möglich?" in *Soziologie: Untersuchungen über die Formen der Vergesellschaftung* (the translation was first published in Kurt H. Wolff (ed.), *Georg Simmel, 1858-1918: A Collection of Essays, with Translations and a Bibliography*, 1959).

## Singer, Peter

1981 *The Expanding Circle: Ethics and Sociobiology*, Oxford, Oxford University Press.

1994 *How Are We to Live? Ethics in an Age of Self-Interest*, London, Mandarin.

1999 *A Darwinian Left: Politics, Evolution and Cooperation*, London, Weidenfeld & Nicolson.

## Skyrms, Brian

1996 *Evolution of the Social Contract*, Cambridge, Cambridge University Press.

## Smart, J. J. C.

1959 "Sensations and Brain Processes", *Philosophical Review* 68, 141-156.

## Smith, Barry

2003 "John Searle: From Speech Acts to Social Reality" in Barry Smith (ed.), *John Searle*, Cambridge, Cambridge University Press, 1-33.

## Snowdon, Charles T.

2002 "From Primate Communication to Human Language" in Frans B. M. de Waal (ed.), *Tree of Origin: What Primate Behavior Can Tell Us about Human Social Evolution*, Cambridge MA, Harvard University Press, 193-227

## Sober, Elliott

1993 *Philosophy of Biology*, Oxford, Oxford University Press.

1999 "The Multiple Realizability Argument against Reductionism", *Philosophy of Science* 66, 542-564.

## Sober, Elliott &amp; David Sloan Wilson

1998 *Unto Others: The Evolution and Psychology of Unselfish Behavior*, Cambridge MA, Harvard University Press.

Stanford, Craig B.

- 2002 "The Ape's Gift: Meat-eating, Meat-sharing, and Human Evolution" in Frans B. M. de Waal (ed.), *Tree of Origin: What Primate Behavior Can Tell Us about Human Social Evolution*, Cambridge MA, Harvard University Press, 95-117.

Sterelny, Kim

- 2003 *Thought in a Hostile World: The Evolution of Human Cognition*, Oxford, Blackwell.

Sterelny, Kim & Paul E. Griffiths

- 1999 *Sex and Death: An Introduction to Philosophy of Biology*, Chicago, University of Chicago Press.

Stoutland, Frederick

- 1976 "The Causal Theory of Action" in Juha Manninen & Raimo Tuomela (eds.), *Essays on Explanation and Understanding: Studies in the Foundations of Humanities and Social Sciences*, Dordrecht, D. Reidel, 271-304.
- 1986 "Reasons, Causes, and Intentional Explanation", *Analyse & Kritik* 8, 28-55.
- 1988 "On Not Being a Behaviourist" in Lars Hertzberg & Juhani Pietarinen (eds.), *Perspectives on Human Conduct*, Leiden, E.J. Brill, 37-60.
- 2005 "The Problem of Congruence" in Ilkka Niiniluoto & Risto Vilkkko (eds.), *Philosophical Essays in Memoriam Georg Henrik von Wright (Acta Philosophica Fennica 77)*, Helsinki, The Philosophical Society of Finland, 127-150.

Strawson, Peter

- 1962 "Freedom and Resentment", *Proceedings of the British Academy* XLVIII, 1-25.

Sugden, Robert

- 2000 "Team Preferences", *Economics and Philosophy* 16(2), 175-204.
- 2002 "Beyond Sympathy and Empathy: Adam Smith's Concept of Fellow-Feeling", *Economics and Philosophy* 18, 63-87.

Taylor, Charles

- 1985 "Social Theory as Practice" in *Philosophy and the Human Sciences: Philosophical Papers 2*, Cambridge, Cambridge University Press, 91-115.

Taylor, Michael

- 1987 *The Possibility of Cooperation*, Cambridge, Cambridge University Press.



Thomasson, Amie L.

2003 "Foundations for a Social Ontology", *ProtoSociology* 18-19, 269-290.

Tollefsen, Deborah Perron

2004 "Collective Intentionality", *Internet Encyclopedia of Philosophy*, <http://www.iep.utm.edu/>, 19<sup>th</sup> August 2004.

Tomasello, Michael

2004 "Understanding and Sharing Intentions: The Origins of Cultural Cognition", manuscript presented at the *Conference on Collective Intentionality IV*, Siena, Italy, October 2004.

Tomasello, Michael & Luigia Camaioni

1997 "A Comparison of the Gestural Communication of Apes and Human Infants", *Human Development* 40, 7-24.

Tomasello, Michael & Hannes Rakoczy

2003 "What Makes Human Cognition Unique? From Individual to Shared Collective Intentionality", *Mind & Language* 18(2), 121-147.

Tooby, John & Leda Cosmides

1992 "The Psychological Foundations of Culture" in Jerome H. Barkow, Leda Cosmides and John Tooby (eds.), *The Adapted Mind: Evolutionary Psychology and the Generation of Culture*, Oxford, Oxford University Press, 19-136.

Tuomela, Raimo

1976 "Explanation and Understanding of Human Behavior" in Juha Manninen & Raimo Tuomela (eds.), *Essays on Explanation and Understanding: Studies in the Foundations of Humanities and Social Sciences*, Dordrecht, D. Reidel, 183-205.

1977 *Human Action and its Explanation: A Study on the Philosophical Foundations of Psychology*, Dordrecht, Reidel.

1995 *The Importance of Us: A Philosophical Study of Basic Social Notions*, Stanford, Stanford University Press.

1998 "A Defense of Mental Causation", *Philosophical Studies* 90, 1-34.

2000 *Cooperation: A Philosophical Study*, Dordrecht, Kluwer.

2002 *The Philosophy of Social Practices: A Collective Acceptance View*, Cambridge, Cambridge University Press.

2007 *The Philosophy of Sociality: The Shared Point of View*, New York, Oxford University Press, forthcoming.

Velleman, David J.

- 1997 "How to Share an Intention", *Philosophy and Phenomenological Research* LVII(1), 29-50.

Verbeek, Bruno

- 2002 *Instrumental Rationality and Moral Philosophy: An Essay on the Virtues of Cooperation*, Dordrecht, Kluwer.

Weber, Max

- 1922/1947 *The Theory of Social and Economic Organization* (translation edited by Talcott Parsons), New York, The Free Press, 1947. Originally published in 1922 as Part I of *Wirtschaft und Gesellschaft*.

Wikforss, Åsa Maria

- 2001 "Semantic Normativity", *Philosophical Studies* 102, 203-226.

Williams, Meredith

- 1999 *Wittgenstein, Mind and Meaning: Towards a Social Conception of Mind*, London, Routledge.
- 2000 "Wittgenstein and Davidson on the Sociality of Language", *Journal for the Theory of Social Behaviour* 30(3), 299-318.

Wilson, David Sloan

- 2002 *Darwin's Cathedral: Evolution, Religion, and the Nature of Society*, Chicago, the University of Chicago Press.

Winch, Peter

- 1958 *The Idea of a Social Science and its Relation to Philosophy*, London, Routledge & Kegan Paul.

Wittgenstein, Ludwig

- 1953 *Philosophical Investigations* (translated by G.E.M. Anscombe), Oxford, Blackwell.
- 1969 *On Certainty* (edited by G. E. M. Anscombe & G. H. von Wright, translated by Denis Paul & G. E. M. Anscombe), Oxford, Blackwell.

Wright, Crispin

- 1981 "Rule-Following, Objectivity and the Theory of Meaning" in Steven H. Holtzman & Christopher M. Leich (eds.), *Wittgenstein: To Follow a Rule*, London, Routledge & Kegan Paul, 99-117.

von Wright, Georg Henrik

- 1971 *Explanation and Understanding*, Ithaca, Cornell University Press.
- 1972 "On So-Called Practical Inference", *Acta Sociologica* 15(1), 39-53.
- 1976a "Replies" in Juha Manninen & Raimo Tuomela (eds.), *Essays on Explanation and Understanding: Studies in the Foundations of Humanities and Social Sciences*, Dordrecht, D. Reidel, 371-413.
- 1976b "Determinism and the Study of Man" in Juha Manninen & Raimo Tuomela (eds.), *Essays on Explanation and Understanding: Studies in the Foundations of Humanities and Social Sciences*, Dordrecht, D. Reidel, 415-435.
- 1988 "Reflections on Psycho-Physical Parallelism" in Lars Hertzberg & Juhani Pietarinen (eds.), *Perspectives on Human Conduct*, Leiden, E.J. Brill, 22-32.
- 1989a "Intellectual Autobiography of Georg Henrik von Wright" in Schilpp & Hahn (1989), 3-55.
- 1989b "A Reply to my Critics" in Schilpp & Hahn (1989), 733-887.
- 2001 *Mitt liv som jag minns det* (My Life as I Remember it, in Swedish), Stockholm, Bonnier.