The London School of Economics
and Political Science

Interpersonal Comparisons of Utility. The Epistemological Problem

Mauro Rossi

A thesis submitted to the Department of Philosophy, Logic and Scientific Method of the London School of Economics for the degree of Doctor of Philosophy

London, August 2008

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

To my family

Declaration

I certify that the thesis I have presented for examination for the PhD degree of the London School of Economics and Political Science is solely my own work other than where I have clearly indicated that it is the work of others (in which case the extent of any work carried out jointly by me and any other person is clearly identified in it).

Abstract

My doctorate thesis investigates a particularly controversial issue in both philosophy of economics and philosophy of mind, namely, the problem of interpersonal utility comparisons (IUCs henceforth).

As I take utility to be a numerical representation of the intensity of individual preferences, IUCs are judgments about how different people's preferences compare in terms of strength. As factual judgments, IUCs appear to be either underdetermined by the empirical evidence or indeterminate. This casts doubt on whether or not we can have (scientific) knowledge of, or, at least, (scientifically) justified beliefs about, how different people's preferences compare in terms of strength.

In general, IUCs can be justified if the assumption of interpersonal similarity, in one of its forms, can be vindicated. I consider two strategies, which attempt to vindicate this assumption by means of, respectively, an inference to the best explanation type of argument and a nativist argument. I argue that both strategies fail.

These results suggest that preferences may be interpersonally incomparable with respect to the dimension of strength. I consider four 'possibility' arguments addressing this challenge. I argue that, although some of them may solve the conceptual problem concerning the interpersonal comparability of preference strengths, they all fail to solve the epistemological problem of IUCs.

Nevertheless, I argue that a 'modest' transcendental argument shows that IUCs can, at least, be justified, provided that we embrace a coherentist view about the structure of epistemic justification.

Acknowledgements

First and foremost, I would like to thank my doctoral supervisor – Professor Richard Bradley – for his constant encouragement, his detailed and eye-opening comments and his precious academic advices.

I am also grateful to the Department of Philosophy, Logic and Scientific Method as well as the AHRC for their generous support.

I am especially indebted towards Armin Schulz and Stuart Yasgur for many inspiring and illuminating discussions and for their invaluable friendship. I have learnt very much from them, both as philosophers and as persons.

I would like to thank my other very good friends Lefteris Farmakis, Andrea Filtri, Damien Lanfrey, Matteo Morganti, Brunello Rosa and Sheldon Steed, for many interesting and stimulating discussions about both philosophical and non-philosophical topics. Furthermore, I am grateful to all the people from whom I have learnt useful lessons during these years.

Last but not least, I would like to express my deepest gratitude towards my parents – Fabio and Lisa – and my sisters – Elena and Rita, for the faith, the patience and the unlimited love which they have given me during these years. Without their support and trust this work could not have been completed.

Table of Contents

Title 1
Dedication 2
Declaration 3
Abstract 4
Acknowledgements 5
Table of Contents 6
List of Figures 8

INTRODUCTION 9

CHAPTER 1 – The problem of interpersonal comparisons of utility 19
1. Introduction 19
2. The problem of IUCs in the 'standard picture' 20
3. IUCs and empirical meaningfulness 29
4. IUCs, incomparability, incommensurability 34
5. IUCs, knowledge and justification 39
6. Conclusion 47

CHAPTER 2 – Inferences to the best explanation 48
1. Introduction 48
2. Third-person approaches 49
3. First-person approaches 57
4. Troubles for Harsanyi's first-person approach 62
5. Against pragmatic solutions 69
6. Conclusion 75

CHAPTER 3 – The argument from nativism 76
1. Introduction 76
2. The problem of meaning 78
3. The problem of mindreading 81
4. Mindreading and the problem of IUCs 86
5. The conditions for scientific justification 88

6. Discussion 97
7. Conclusion 108

CHAPTER 4 – Three 'possibility' arguments 110
1. Introduction 110
2. Broome's strategy 111
3. Can functionalism rescue IUCs? 117
4. Bradley's strategy 122
5. Conclusion 129

CHAPTER 5 – Transcendental arguments 131
1. Introduction 131
2. The background assumption of rationality 133
3. Davidson's strategy 135
4. A 'strong' transcendental argument 139
5. The Principle of Charity and the Principle of Similarity 141
6. Objections 145
7. A 'modest' transcendental argument 148
8. The epistemological problem of IUCs reconsidered 152
9. Conclusion 155

CONCLUSION 157

BIBLIOGRAPHY 165

List of Figures

Figure 1 82
Figure 2 84

INTRODUCTION

It is a commonplace that, in everyday life, we ascribe all sort of mental states to other people: sensations, like pain and hunger; emotions, like fear and love; and propositional attitudes, like desires, preferences and beliefs. Examples vary from the trivial to the more complex. I see Nancy drinking a glass of water and I ascribe to her the belief that the glass contains water and a desire to quench her thirst. In a different circumstance, I see Nancy waving on the street and I ascribe to her the intention of either saying hello or signalling a turn, depending on the information that I possess about the surrounding environment and about her personal history. The capacity of "understanding the mind", or, as it is often referred to, the capacity of mindreading, typically serves a variety of purposes, which include the prediction, explanation, and interpretation of other individuals' behaviour.

In everyday life, not only do we ascribe mental states, but we also compare them. We compare mental states with respect to a variety of dimensions: their type, their intensity and, in the case of propositional attitudes, their content. We compare both our own and other people's mental states, that is, we make both intra-personal and inter-personal comparisons. A remarkable fact is that we typically make interpersonal comparisons (ICs, for short) of mental states with relatively little difficulty. Moreover, we often do not find inter-personal comparisons of mental states more difficult than intra-personal comparisons, that is, of comparisons involving our own mental states[1].

Here is an example offered by Richard Jeffrey of an everyday situation where the comparison of two individuals' mental states has some relevance:

"Shall we open the can of New England clam chowder or the can of tomato soup, for the children's lunch? Adam prefers the chowder; his sister Eve prefers the other. Their preferences conflict. But it is acknowledged between them that Adam finds tomatoes really repulsive, and loves clams, whereas Eve can take clam chowder or leave it alone, but is moderately fond of tomato soup. They agree to have the chowder."[2]

As this example shows, ICs of mental states are often made for normative purposes, e.g. decisions involving the distribution of goods. However, in everyday practice, we also make

[1] See DAVIDSON, D. [1986], reprinted in DAVIDSON, D. [2004], p. 59.
[2] JEFFREY, R. [1974], reprinted in JEFFREY, R. [1992], p. 182.

ICs of mental states for evaluative purposes and, arguably, for explaining other people's behaviour.

The ease with which we make ICs of mental states in everyday life contrasts with the difficulties that such comparisons pose at the theoretical level. In particular, what presents the most challenging puzzles is the comparison of the intensity of different people's mental states. Consider how Stanley Jevons and Lionel Robbins, respectively, describe the problem:

"The susceptibility of one mind may, for what we know, be a thousand times greater than that of another. But, provided that the susceptibility was different in a like ratio in all directions, we should never be able to discover the difference. Every mind is thus inscrutable to every other mind, and no common denominator of feeling seems to be possible."[3]

"[S]uppose that we differed about the satisfaction derived by A from an income of £1,000, and the satisfaction derived by B from an income of twice that magnitude. Asking them would provide no solution. Supposing they differed. A might urge that he had more satisfaction than B at the margin. While B might urge that, on the contrary, he had more satisfaction than A. We do not need to be slavish behaviourists to realise that there is no scientific evidence. *There is no means of testing the magnitude of A's satisfaction as compared with B's.* If we tested the state of their blood-streams, that would be a test of blood, not satisfaction. Introspection does not enable A to measure what is going on in B's mind, nor B to measure what is going on in A's. There is no way of comparing the satisfactions of different people."[4]

It is not surprising that these complaints come from two economists. Since its beginning as a modern science, economics has assigned a central place to mental states of various sorts and represented them numerically through a utility function. Since, in the course of the centuries, economists have taken different mental states as objects of their analysis, some confusion has arisen about the meaning of the utility notion. From a historical point of view, we can broadly distinguish three ways in which the notion of utility has been used[5]. First, there is the traditional use of utility as synonym of happiness, which, in turn, is

[3] JEVONS, S. [1911], p. 14.
[4] ROBBINS, L. [1932], pp. 139-140. [Emphasis in the original]
[5] See COOTER, R. and P., RAPPOPORT [1984]. It is worth emphasising that the proposed distinction is concerned with a historical, rather than conceptual, reconstruction of the meaning of utility.

defined as the net result of pleasures minus pains. Second, there is the use of utility as subjective feeling of satisfaction. Finally, there is the use of utility as a representation of individual desires or preferences[6]. A striking fact is that the interpersonal comparison of different people's utilities remains problematic under each of these interpretations. For instance, while Jevons' remarks target the traditional meaning of utility, Robbins' remarks target (sometimes inconsistently) the second interpretation.

In this thesis, I shall consider utility in the latter sense. More specifically, I shall take utility to be a numerical representation of the intensity of individual preferences[7]. The choice of preferences as objects of my analysis responds to a specific motivation. Although the difficulties in comparing mental states affect all the fields where mental states play a relevant role, the failure to give a plausible theoretical systematization to the problem of ICs of preference strength has particularly far-reaching consequences for several areas within, or connected to, contemporary economic analysis. More specifically, the problem of comparing different people's preference strengths is particularly important in three, inter-related, fields: traditional welfare economics, social choice theory and ethics.

Traditional welfare economics tries to rank alternative states of affairs on the basis of people's preferences towards it. If we cannot compare different individuals' preferences in terms of strength, welfare economics is unable to give recommendations in cases where changing the state of affairs increases the utility of one or more individuals at the price of diminishing the utility of at least one other individual in society. In other words, welfare economics is unable to settle distributive conflicts[8].

Social choice theory offers another clear example of the importance of the problem. Arrow's seminal work on preference aggregation shows that there is no way to aggregate individual preferences in order to obtain a social ranking of alternative states of affairs, which satisfies few, very mild, conditions: collective rationality, unrestricted domain, weak Pareto principle, independence of irrelevant alternatives and non-dictatorship[9]. As Sen and many others have proved, however, Arrow's impossibility result can be turned into a possibility result if we relax the condition of independence of irrelevant alternatives by

---

[6] According to COOTER, R. and P., RAPPOPORT [1984], there is another use, namely, the one of utility as referring to what is objectively 'useful' in terms of need satisfaction. Cooter's and Rappoport's interpretation is questioned by LITTLE, I. M. D. [1985] and HENNIPMAN, P. [1988]. For replies see, respectively, COOTER, R. and P., RAPPOPORT [1985] and RAPPOPORT, P. [1988]. It is worth noticing that, although utility does not refer here to any mental state, interpersonal utility comparisons remain a problem also in this case.

[7] I shall utility to be a *mere* representation of the intensity of individual preferences. This implies that I will not take utility to be any sort of emotion, feeling, or propositional attitude distinct from, or even identical to, preferences.

[8] For extensive surveys of welfare economics, see MISHAN, E. J. [1960] and CHIPMAN, J. and J., MOORE [1978].

[9] See ARROW, K. [1963].

introducing ICs of preference strength[10]. The literature has, for the most part, focused on the implications of allowing ICs of different kinds and with alternative informational bases[11]. However, considerably less work has been done on other more foundational issues concerning ICs of preference strength.

Finally, the problem is particularly important in ethics and applied ethics. The meaningfulness of various ethical doctrines crucially depends on the very possibility of making ICs of preference strength. For instance, such comparisons play a crucial role within the preference satisfaction theory of well-being. If, as it is usually maintained, the degree to which an individual's life goes well is given by the intensity of his preference for the option that the world realises, then, if we cannot compare different individuals' preferences in terms of strength, it follows that we cannot compare different people's degrees of well-being either.

Let us taken for granted that ICs of preferences pose a particularly serious theoretical problem. What exactly is the nature of the problem? We can identify a set of distinct, although not always independent, questions about ICs, which can sometimes be confused and conflated[12].

1. *The semantic question*: what is the meaning of IC judgments?
2. *The measurement question*: how can we measure the mental states to be interpersonally compared?
3. *The descriptive question*: what are the key features of our everyday practice of making ICs?
4. *The explanatory question*: how can we explain our capacity for making ICs?
5. *The metaphysical question*: is there a fact of the matter about ICs?
6. *The epistemological question*: can we have (scientific) knowledge of, or, at least, (scientifically) justified beliefs about, ICs?
7. *The normative question*: how should we make ICs (for different purposes)?

---

[10] See SEN, A. [1970], specially pp. 35-36.

[11] See SEN, A. [1970], [1973], [1977], HAMMOND, P. [1976], D'ASPREMONT, C. and GEVERS [1977], MISKIN, E. [1978], ROBERTS, K. [1980a,b], [1995] and SUZUMURA, K. [1996] among the others. There are several proposals concerning the informational basis to adopt. On the non-utilitarian side, different authors advocate the adoption of primary goods, resources, rights, opportunities, capabilities, basic needs, as relevant objects of comparison. On the utilitarian side, different authors suggest defining utility in terms of welfare, preferences, interests, happiness, desire satisfaction and so on. In particular, see RAWLS, J. [1971], [1982] for primary goods, DWORKIN, R. [1981a,b], [2000], for a resource-based approach, NOZICK, R. [1974] for rights, ARNESON, R. J. [1989] and ROEMER, J. E. [1998] for opportunities, SEN, A. [1985], [1993] for capabilities, GRIFFIN, J. [1986] for desire satisfaction, DAVIDSON, D. [1986] for interests, NG, Y.-K. [1996], [1997] for happiness, while, for the literature concerning basic needs, see SEN, A. [1999], p. 359.

[12] This set of questions is similar, although not identical, to the one proposed by Davies and Stone in the context of the problem of mindreading. See DAVIES, M. and T., STONE [1996], pp. 119-120.

For the purpose of this thesis, the metaphysical and the epistemological questions are the most important ones. The former question concerns whether or not there are any facts that would make IC judgments true. The latter concerns whether or not we can have epistemic access to these (alleged) facts. As a working hypothesis, I will initially presuppose an affirmative answer to the metaphysical question. This means that I shall take ICs to be factual judgments. My focus will be on the question of whether or not, and to what extent, we can have (scientific) knowledge of, or, at least, (scientifically) justified beliefs about, ICs of preference strength. However, in the course of my analysis, I shall reconsider the metaphysical question more closely and try to offer a more direct justification for my initial assumption. Since economists typically represent individual preferences by a (family of) utility function(s), I shall equally refer (with the qualifications to be seen in chapter 1) to the problem of comparing preference strengths as the problem of interpersonal utility comparisons (IUCs, for short).

In order to answer these questions, I shall briefly touch on the issue of the nature of mental states, in general, and preferences, in particular. Orthodox economics oscillates between behaviourism and dispositionalism. According to the former doctrine – pioneered by Samuelson in his "revealed preference approach"[13] – preferences are nothing but instances of observable choice behaviour. According to the latter doctrine, preferences are dispositions to cause observable choice behaviour. Traditionally, the problem of IUCs has been discussed with respect to these two ways of conceiving the nature of preferences. However, at least since the '70s, some philosophers of mind have suggested adopting a different, functionalist, account of the nature of mental states[14]. More recently, other philosophers have advanced an alternative, experientialist, characterisation of the nature of mental states[15]. According to functionalism, mental states are individuated with respect to the role that they occupy in the individuals' mind, in relation to inputs, other mental states and behavioural outputs. Instead, according to experientialism, mental states are individuated with respect to the family of conscious experiences that individuals undergo. One of the goals of my thesis is to see how these different conceptions affect the conclusions concerning the epistemological problem of ICs of preference strength.

I shall also indirectly consider some of the other question about ICs listed above. In order to understand whether or not we can form justified beliefs about how different people's preferences compare in terms of strength, I shall consider the explanatory question

[13] See SAMUELSON, P. A. [1947].
[14] LEWIS, D. [1972] is the *locus classicus*.
[15] See, in particular, GOLDMAN, A. [1993].

of how we *make* ICs of preference strength. Typically, the activity of making such comparisons is conceived as a two-step process. In the first step, preferences are ascribed to other individuals. In the second step, preferences are compared with respect to their intensity. One suggestion is that the problem of comparing preferences is just a particular case of the more general problem of ascribing mental states. By examining the latter, we can better understand what conditions need to be satisfied in order for our beliefs about how different people's preferences compare in terms of strength to be justified.

Two qualifications should be added. The explanatory question of how we ascribe preferences may refer to the methods and processes used to ascribe preferences by either scientific researchers or by ordinary people. In this thesis, I shall consider both cases. With respect to the latter case, moreover, the explanatory question can be addressed at different levels of description, i.e. personal, sub-personal and physical[16]. The personal level of description focuses on the way in which persons, as such, think about and interpret other people's mental and overt behaviour. The sub-personal level of description focuses on the underlying information-processing mechanisms that need to be postulated in order to explain people's mindreading capacity. The physical level of description focuses on the physical structure that realizes the mental architecture as conceived at the functional level. In this thesis, I shall consider the first two levels of analysis only.

Furthermore, I shall briefly examine the measurement question. Indeed, as I shall claim in chapter 1, one of the conditions for having scientifically justified beliefs about ICs is that the compared mental states must be accurately and precisely measurable. The problem of scientific justification is connected to the problem of measurement. Thus, in my thesis, I shall consider the question of how preferences can be measured. Moreover, I shall devote some attention to the contrast between beliefs, which supposedly are both measurable and comparable, and preferences, which supposedly are measurable but not comparable[17].

Coming now to the structure of this thesis, I shall proceed as follows. In chapter 1, I shall present the problem of IUCs. Indeed, despite its importance, the literature is often vague about how to characterize it. As a consequence, the results that are drawn are often unclear. For instance, different authors conclude that IUCs are impossible[18] or meaningless[19] or, at best, that they are not factual, but normative, judgments[20]. These

[16] See DENNETT, D. [1969].
[17] See BRADLEY, R. [2007b] for a recent parallel between the problem of ICs of degrees of belief and the problem of ICs of degrees of preference.
[18] Although the theme of the "impossibility" of IUCs is a common one, in the economic literature, it is hard to identify a paradigmatic statement of such a position. For an early reaction against the impossibility of IUCs, instead, see LITTLE, I. D. M. [1957], chapter IV.
[19] See ARROW, K. [1963], p. 9.
[20] See ROBBINS, L. [1932], p. 139.

claims seem to conflate some of the distinct issues about ICs that we have listed above: the explanatory issue of whether or not we can make ICs at all, the semantic issue of whether or not IC judgments are meaningful and the metaphysical issue about the nature of ICs, respectively. In chapter 1, my goal is to contrast the 'standard' way of presenting the problem of IUCs with the one that I favour, according to which the problem is whether or not can we have (scientific) knowledge of, or, at least, (scientifically) justified beliefs about how different people's preferences compare in terms of strength.

In the next two chapters I shall look more closely at the issue of whether or nor we can have (scientifically) justified ICs of preference strength. In chapter 2, I shall consider some accounts offered in the economic literature, which are based on an inference to the best explanation kind of argument. The general idea is that a theory, or an assumption, is justified if it offers, or contributes to offering, the best explanation of a certain phenomenon. In turn, the criteria for individuating the best explanation typically make reference to pragmatic considerations, such as explanatory power, simplicity, or parsimony. In the case of IUCs, the argument is that we are justified in assuming that different people's utilities are co-scaled insofar as this provides the best explanation of their behaviour in terms of the pragmatic virtues seen above. Contrary to a common intuition, however, I shall argue that the assumption that different people's utilities are co-scaled does not add anything to the explanation of individual behaviour nor makes a theory including it either more parsimonious or simpler than a theory that does not include it. Therefore, this strategy fails to successfully address the issue of justification.

An interesting feature of the economic literature is that it often attempts to ground the solutions given to the problem of IUCs on the explanation of how ordinary people supposedly make ICs of preference strength in everyday life[21]. Since this explanatory problem concerns mental states (i.e. preferences) and one of their properties (i.e. strength) in particular, one would expect the existence of both a large literature in philosophy of mind addressing the issue and a particularly strong interdisciplinary exchange between economics and philosophy of mind. Instead, and quite surprisingly, neither expectation is actually met. On the one hand, economists offer only casual remarks about how ordinary people make ICs, which lack both empirical and conceptual support. On the other hand, philosophers of mind have almost completely ignored this explanatory problem. One significant exception is constituted by Alvin Goldman, who has attempted to bring the

[21] See, amongst the others, HARSANYI, J. [1955] and [1977], LITTLE, I. D. M. [1957], JEFFREY, R. [1974], LIST, C. [2003].

problem of IUCs in line with current debates in philosophy of mind and epistemology[22]. In chapter 3, I will start from his work and pursue two goals. First, I shall try to show how philosophy of mind can contribute to the debate by extending Goldman's analysis. Indeed, Goldman focuses mainly on ICs of happiness and adopts a very specific approach to mental ascription, i.e. Simulation Theory. By contrast, I shall focus on ICs of preference strength and consider both Simulation Theory and the other main approach to mental ascription, i.e. Theory Theory. Second, I shall assess whether or not philosophy of mind can help us find a successful solution to the problem of IUCs. I shall devote a special interest to Goldman's own argument from nativism. According to it, the assumption that different people's utilities are co-scaled is justified if the assumption that ICs of preference strength are performed through innate mechanisms that are either hyper-similar across individuals or very closely representative of the workings of other individuals' mind-systems is sound. I shall argue that, when the notion of innate cognitive capacity or mechanism is properly spelt out, this argument reduces to an inference to the best explanation kind of argument. Therefore, this strategy too fails to successfully address the issue of justification.

The failure of the previous arguments increases the pressure brought by the sceptical challenge. The idea is that, perhaps, the alleged impossibility of having justified IUCs stems from the *incomparability* of preferences with respect to the dimension of strength. As a consequence, several authors resort to more radical 'in principle' solutions to the problem of IUCs. These solutions are based on 'possibility' arguments. Their primary goal is to show that different people's preference strengths are indeed comparable. Their secondary goal is to show that it is possible, in principle but not by means of empirical or pragmatic considerations only, to have (scientific) knowledge of, or, at least, (scientifically) justified beliefs about how different people's preferences compare in terms of strength. In chapter 4, I shall consider three 'possibility' arguments. Although these arguments are made in the context of a more economic-oriented analysis, they significantly borrow conceptual tools from both metaphysics and philosophy of mind. The first argument is based on Broome's work on personal goodness. It claims that, if individual preferences are independent from personal identity, then it is conceptually possible to construe a universal preference scale, provided that each individual can live at least another individual's pair of lives and that all individuals' lives are connected in a suitable way. If this is the case, different people's preferences are interpersonally comparable in terms of strength. The second and third arguments are based on a functionalist understanding of the nature of preferences. Both arguments claim that it is conceptually possible to identify two points with respect to which

[22] See GOLDMAN, A. [1995a].

different people's preferences play the same causal role. If this is the case, functionalism allows us to conclude that preferences are interpersonally comparable in terms of strength. The former argument claims that these points are given, respectively, by the most preferred and by the least preferred prospects in the individual's lifetime preference ranking. The latter argument – offered by Bradley – claims that one point is given by the ethically neutral prospect, while the other is given by the total desirability of all prospects. However, I shall argue that, although some of these arguments may solve the conceptual problem concerning the comparability of different people's preference strengths, they all fail to solve the epistemological problem of IUCs on the grounds that they do not show that the relevant causal mechanisms determining individual preferences are really the same across individuals.

Once again, the assessment of the more economic-oriented analysis invites a corresponding analysis of the same issues from a more philosophy-oriented perspective. In chapter 5, I shall examine another 'possibility' argument that is typically made in the context of the explanation of people's mindreading capacity at the personal level of description. In its original version, this argument proceeds from the premises that we interpret each other correctly and that the interpersonal comparability of mental states is necessarily required by the very task of interpretation to the conclusion that preferences are indeed interpersonally comparable in terms of strength. As such, it takes the form of a 'strong' transcendental argument. In line with Stroud's critique, I shall argue that this position shows, at best, that, necessarily, interpretation requires *taking* mental states to be comparable. However, it does not show that they *really* are interpersonally comparable. Nevertheless, we may still reach results of anti-sceptical significance by employing a transcendental argument of a more 'modest' form. The goal is to demonstrate only that, necessarily, interpretation requires one to *take*, or believe that, different people's preferences are interpersonally comparable from the start. I shall argue that, if a 'modest' transcendental argument is defensible, then, if it is combined with coherentism, it shows that ICs of preference strength can, at least, be (scientifically) justified.

To summarise, the main strategies examined in this thesis fail to show that that we can have (scientific) knowledge of how different people's preferences compare in terms of strength. An interesting exception is offered by a 'modest' transcendental strategy, which shows that ICs of preference strength can, at least, be (scientifically) justified, *if* one embraces coherentism about epistemic justification. Since the success of this strategy is conditional on the acceptance of a very specific and not uncontroversial thesis, it does not reach the status of conclusiveness that one could hope for. As a consequence, one may read

this thesis in a disjunctive fashion: either it provides a positive argument for the possibility of having (scientifically) justified IUCs, if coherentism is true, or it provides an argument by elimination, to the effect that none of the existing solutions allow for the possibility of having (scientifically) justified IUCs.

CHAPTER 1

The problem of interpersonal comparisons of utility

1. Introduction

The orthodox view in economics and philosophy is that IUCs pose remarkable theoretical difficulties. However, a precise definition of the problem is somehow lacking in the literature. As we have seen in the introduction, part of the confusion stems from the failure to distinguish the existence of a set of distinct and relatively independent questions that can be asked about IUCs. For instance, the problem of IUCs is often characterised by means of expressions of the following sort: interpersonal utilities are not on the same scale; interpersonal utilities have no factual basis; interpersonal utilities are incomparable; IUCs are empirically meaningless; etc. At first sight, however, these expressions are not logically equivalent. In this chapter my goal is to present the approach that I will adopt in this thesis and to illustrate how it relates to alternative ways of formulating the problem of IUCs. This will provide a map of how the relevant notions listed above are connected to each other.

I shall proceed as follows. In section 2, I shall present the 'standard picture' of the problem of IUCs and its main features. I shall argue that this framework limits a more thorough understanding of some relevant issues. First of all, it does not adequately distinguish the metaphysical and the epistemological questions about IUCs. Secondly, it does not clarify the relationship between the problem of comparing the intensity of different people's preferences and the problem of comparing their utilities. Thirdly, it neglects current debates in philosophy of mind and epistemology. A better formulation must be able to take all these issues into account.

In section 3, I shall discuss the first limitation by considering the approach originally followed by Waldner and, more recently, by List, according to which the problem of IUCs is the problem of whether or not IUCs are empirically meaningful[1]. In section 4, I shall discuss the second limitation by considering an alternative characterisation, according to which the problem of IUCs is the problem of whether or not different people's utilities are commensurable. Finally, in section 5, I shall discuss the third limitation by considering the approach that I will adopt in this thesis, which is based on the treatment of the problem of

[1] See WALDNER, I. [1972] and LIST, C. [2003].

IUCs given by Goldman in his "Simulation and Interpersonal Utility"[2]. Broadly speaking, the main focus of my analysis will be on the epistemological question of whether or not we can have knowledge of, or justified beliefs about, how different individuals' preferences compare in terms of strength. More narrowly, I shall consider whether or not we can have *scientific* knowledge of, or *scientifically* justified beliefs about, how different people's preferences compare in terms of strength.

2. The problem of IUCs in the 'standard picture'

Generally speaking, we can say that the problem of IUCs is the problem of comparing different people's utilities. The first difficulty that one encounters concerns the meaning of the word 'utility'. As we have seen in the introduction, 'utility' is a technical notion, whose meaning has changed in the course of the years, as a consequence of the changes in the theories in which it has been embedded[3]. For clarity, here I shall define it as the numerical value of a function, i.e. the utility function, which represents an individual's preferences. Thereby, we can define the problem of IUCs as the problem of comparing different people's preferences, as numerically represented through a (family of) utility function(s).

The 'standard picture' characterises the theoretical framework in which the problem of IUCs arises as a sequence of four steps[4], dealing with:

(1) the determination of individual preferences;
(2) their representation through a (family of) utility function(s);
(3) the interpersonal comparison of utilities;
(4) the formulation of the judgment of interest.

Let us examine each step in detail.

2.1 The determination of preferences

[2] GOLDMAN, A. [1995a]. Here, I shall attempt to improve Goldman's account in two ways. On the one hand, I shall elaborate and expand his presentation by analyzing how the conditions for knowledge apply to the problem of IUCs and by discussing in more details the idea of scientific justification in the case of IUCs. On the other hand, I shall illustrate how this approach is related to alternative ways of formulating the problem of IUCs existing in the literature.

[3] See also STIGLER, G. [1950a,b], COOTER, R. and P., RAPPOPORT [1984] and BROOME, J. [1999] for a historical reconstruction.

[4] DAVIDSON, D. [1986], reprinted in DAVIDSON, D. [2004], and FLEURBAEY, M. and J., HAMMOND [2004] offer similar, although not identical, reconstructions.

The first step in the 'standard picture' is concerned with the individuation of preferences. This task is not independent from issues concerning (a) the nature, (b) the domain and (c) the properties of preferences.

Orthodox economics oscillates between the adoption of either a behaviourist or a dispositionalist account of the nature of preferences. According to the former doctrine – prominent in the early days of the revealed preference approach – preferences are nothing but instances of observable choice behaviour. Thus, for instance, an individual's preference for taking, rather than not taking, his umbrella is nothing but the individual's act of taking the umbrella in the corresponding choice situation. However, behaviourism is a highly problematic theory of the nature of mental states. The main objection is that it excludes the possibility of having preferences in the absence of occurring choice situations. In other words, it excludes the possibility of having hypothetical preferences[5]. Suppose that, in the previous example, the individual is not presented with the choice of taking the umbrella. We may be tempted to say that he still prefers to perform this action, even if this does not currently become manifest in overt behaviour. Thus, it appears that the nature of mental states cannot be entirely defined in behaviourist terms.

According to the latter doctrine, preferences are dispositions to cause observable choice behaviour. Dispositions manifest themselves only if the relevant conditions are satisfied. In our example, the individual's act of taking the umbrella shows that the individual has a categorical state that disposes him to take the umbrella in suitable circumstances. Given that observable choice behaviour is not a necessary condition for having preferences, a dispositional account can take into account the possibility of having hypothetical preferences, while, at the same time, preserving a moderate behaviourist account of their nature.

The 'standard picture' typically embraces a dispositional account of the nature of preferences. In set-theoretic terms, it conceives preferences as binary relations R, that is, relations between two items. The items included in the preference domain vary according to different decision theories. More specifically, preferences may range over either acts, or propositions, or prospects[6]. The argument in this thesis does not depend on any specific ontological choice. However, for clarity, I shall take preferences to range over prospects. Prospects are mutually exclusive vectors of possible outcomes, together with a probability distribution over these outcomes. We can think of outcomes as states of affairs or possible worlds. At the extreme, each outcome is a complete history, or a particular world. More

[5] See PETTIT, P. [2006], especially p. 133.
[6] For acts, see SAVAGE, L. [1954]. For propositions, see JEFFREY, R. [1983]. For prospects, or lotteries, see VON NEUMANN, J. and O., MORGENSTERN, [1944].

commonly, however, it is a set including all the possible worlds of a certain type. Formally, let x be a prospect included in the preference domain A. We can write $x = <O_1, \ldots, O_k, \ldots, O_n>$, where O_k is an outcome that can occur with a fixed probability p_k, for $k = 1, \ldots, n$. A pure prospect represents the case of an outcome whose occurrence is certain, i.e. $x = <O_k>$.

In the 'standard picture', the individuation of preferences is typically governed by two sets of axioms, namely, choice axioms and preference axioms. The former axioms fix the conditions for inferring the existence of preferences from observed or hypothetical choice behaviour, in accordance with the general dispositional account of preferences. As the accurate individuation of an individual's preferences must proceed holistically, further constraints are imposed on the structure on the individual's preferences. More specifically, preference axioms postulate that each individual has complete and transitive preferences.

2.2 *The representation of preferences*

The second step in the 'standard picture' is concerned with the numerical representation of preferences. Before discussing how this task is performed in the case under consideration, I shall illustrate some of the basic elements of measurement theory[7]. In general, measurement consists in the assignment of numbers that preserve certain empirical relations. More precisely, measurement starts with an empirical relational structure Σ and a numerical relational structure N. Then, it seeks a mapping f from the empirical relational structure Σ to the numerical relational structure N, which preserves all the relevant relations and operations in Σ. The mapping f is called a homomorphism. The triple (Σ, N, f) is called a scale. Two items are measured on the same scale if and only if they are measured with respect to an identical triple (Σ, N, f). For simplicity, however, I shall refer to f alone as a scale of measurement. Thereby, I shall say that two items are co-scaled if and only if they are assigned numbers through the same function f.

The first basic problem of measurement is the representation problem. The goal is to find a set of (necessary and) sufficient conditions for the existence of a homomorphism f from Σ to N. If the conditions are stated in axiomatic form, then the representation problem consists in finding a set of axioms that is (necessary and) sufficient to establish a representation theorem. In turn, the representation theorem asserts that, if an empirical relational structure Σ satisfies these axioms, there exists a homomorphism f into a particular

[7] See KRANTZ, D.H., LUCE, R.D., SUPPES, P. & TVERSKY, A. [1971] and ROBERTS, F. S. [1979], for a more detailed illustration.

numerical structure N, that is, there exists a function mapping the empirical relational structure into the numerical relational structure.

The second basic problem is the uniqueness problem. Given the same empirical relational structure Σ and numerical relational structure N, it may be possible to find more than one function preserving the same relations and operations in Σ. The uniqueness problem consists in specifying how unique the homomorphism from Σ to N is. More precisely, let f and g be homomorphisms from Σ to N. If λ is a function that transforms f into g, by preserving all the information carried by f, then we can say that λ is an admissible transformation of scale. The uniqueness theorem specifies the class of admissible transformations λs that yields homomorphisms from the empirical relational structure Σ into the numerical relational structure N[8].

The class of admissible transformations defines the type of measurement scale. Although there are infinite scale-types, four of them are particularly important: ordinal, interval, ratio, and absolute scale of measurement. An ordinal scale is unique up to a monotone increasing transformation. An interval scale is unique up to a positive affine transformation, of the form $\lambda(x) = \alpha x + \beta$, for some $\alpha > 0$ and $\beta \in \mathbb{R}$. A ratio scale is unique up to a similarity/linear transformation of the form $\lambda(x) = \alpha x$, for some $\alpha > 0$. An absolute scale is absolutely unique. Interval, ratio and absolute scales are cardinal scales of measurement. In an interval scale, both the zero point and the unit are arbitrarily fixed. In a ratio scale, the zero is 'natural' but the unit is fixed arbitrarily. Finally, in an absolute scale, both the zero and the unit are 'natural'. The class of admissible transformations defines the meaningfulness of statements involving a numerical scale of measurement. The standard criterion for meaningfulness is invariance under the class of transformations up to which the numerical representation under consideration is unique. Following Roberts, we can say that "a statement involving numerical scales is meaningful if and only if its truth (or falsity) remains unchanged under all admissible transformations of all the scales involved"[9].

Let us go back to preferences. It is worth emphasising that the representation problem arises at two different stages. The first stage concerns the representation of the agent's observed or hypothetical choice behaviour in terms of preference relations with certain properties. The second stage concerns the representation of the agent's preferences by a numerical function. If the agent's choices satisfy the weak axioms of revealed preferences, they can be represented in terms of preferences with ordering properties, i.e. preferences forming an ordering of options. On the other hand, if the agent's preferences satisfy the

[8] See KRANTZ, D.H., LUCE, R.D., SUPPES, P. & TVERSKY, A. [1971].
[9] ROBERTS, F. S. [1979], p. 71.

conditions of completeness and transitivity, they can be represented though an ordinal utility function u, unique up to a monotone increasing transformation.

If we want to represent preferences on a cardinal scale, the corresponding set of axioms must be richer, since such a representation contains more information than the ordinal one. The common view is that the evidence only suffices to obtain a representation of preferences on an interval scale, while it is insufficient to represent preferences on both a ratio and an absolute scale. At the same time, it is possible to represent preferences on an interval scale in more than one way. This means that there is more than one set of axioms that is sufficient for the cardinalization of preferences. The most common suggestion – and the one that I shall mostly refer to in this thesis – is to use a von Neumann-Morgenstern (vNM) utility function[10]. In the vNM framework, the representation of preferences on an interval scale is supposed to capture not only the order of preferences, but also the degree, or intensity, of the individual's preferences for the options in the preference domain. Choice behaviour remains the relevant evidence, although data are gathered both from situations of certainty and from situations of uncertainty. Together with the ordering axioms, then, the set of axioms includes an Archimedean and an independence axiom. If individual preferences satisfy those conditions, they can be represented through a vNM utility function u, unique up to a positive affine transformation. Measurement leads to the formation of profiles of utility functions, that is, of n-tuples of $\{u_i\}$, for any individual $i = 1, \ldots, n$.

2.3 The comparison of different people's utilities

The third step in the 'standard picture' is concerned with the comparison of different people's utilities. There are two main kinds of IUCs, namely, ICs of utility levels and ICs of utility differences. For any two individuals i and j, and for any four options $x, y, w, z \in A$, ICs of utility levels are judgments of the form: $u_i(x) \geq u_j(y)$, while ICs of utility differences are judgments of the form: $u_i(x) - u_i(y) / u_j(w) - u_j(z) = \lambda$, for some $\lambda \in \mathbb{R}$. In addition to these, List has recently drawn attention to a third kind of IUCs, namely, ICs with respect to an interpersonally significant zero-line. If a significant zero-line exists, an individual i can have utility less than / equal to / greater than a utility level of zero. Formally, this means that $sign(u_i(x)) = \delta$, where $\delta \square \{-1, 0, 1\}$. In turn, this sign-function

[10] See VON NEUMANN, J. and O., MORGENSTERN, [1944]. An alternative suggestion is to use the 'just noticeable difference' method. Another suggestion is based on 'probabilistic choice' models. For references, see HAMMOND, P. [1991].

allows us to make ICs of utility levels between individuals with utility, respectively, less than / equal to / greater than the interpersonally significant zero-line[11].

In the last step, IUCs are used to formulate a judgment of interest. If the judgment concerns a decision involving two (or more) individuals, a decision rule typically establishes the relevant kind of comparison to be made. Roughly, we can distinguish two main purposes for which IUCs can be made: explanatory and normative. On the one hand, IUCs are supposed to help explain features of an individual's behaviour by establishing a comparison both with similar features and with the determinants of another individual's behaviour. The judgment of interest is an explanatory one. The idea is that, by ascribing comparable degrees of preference, one can make sense of why different individuals show different behaviours. On the other hand, IUCs are supposed to help reaching decisions based on the individuals' preferences. The judgment of interest is a normative one. The idea is that, when a choice affects other people and the individual's preferences are the variables on which the outcome is based, decision-making requires making IUCs.

According to the 'standard picture', the problem of IUCs arises at the third stage. Choice behaviour is not sufficient to determine whether or not preferences represented by the same utility values, but belonging to different individuals, have really the same intensity. The problem is that, even when different individuals show the same choice behaviour, the evidence is not sufficient to establish whether or not the function representing their preferences is really the same. In other words, on the basis of choice behaviour alone, it is not possible to claim that different people's utilities are co-scaled. To see why, let us consider an example[12].

Suppose there are two individuals, *i* and *j*, and four options $x, y, w, z \in A$. Individual *i* ranks the options in the following way: $xRyRwRz$. On the other hand, individual *j* ranks the options in the following way: $wRzRxRy$. Suppose we measure their preferences on a (interval) zero-one scale, such that we assign the value 1 to the most preferred option and the value 0 to the worst option. Then, we can assign a value that represents the intensity of their preferences for the other options, relative to the best and the worst in each individual's ranking, in the standard vNM way. Suppose we get that $u_i(y) = u_j(x) = 0.6$. We also get that $u_i(x) - u_i(y) = u_j(w) - u_j(x) = 0.4$. Can we conclude that individual *i* prefers option *y* with the same strength with which individual *j* prefers option *x*? In other words, is choice behaviour sufficient to determine the interpersonal comparison of individuals *i*'s and *j*'s utility levels? Or else, can we conclude that the difference in strength of individual *i*'s preference for

[11] See LIST, C. [2001] for a more detailed introduction to this kind of IUCs.

[12] I shall offer an example in terms of an interval scale of measurement. The problem remains, *mutatis mutandis*, if we measure preferences on an ordinal scale.

option x over y is the same as the difference in strength of individual j's preference for option w over x? In other words, is choice behaviour sufficient to determine the interpersonal comparison of individuals i's and j's utility differences?

The answer is negative in both cases. Even if we identically normalize the scales used for measuring different individuals' preferences, the evidence is not sufficient to determine whether the resulting utility values represent the same interpersonal preference strengths. The measurement is relative to the best and the worst options in each individual's preference ranking. However, choice behaviour does not imply anything about how different people's preferences for their best (worst) option compare in terms of strength. As it is typically put, choice behavioural evidence is consistent with the case in which i prefers the most preferred option with intensity ten times greater than j.

2.4 The analogy with temperature

The problem described in the previous section is not just that the individuals' preferences can be represented by more than one utility function, in accordance with the admissible transformations specified by the uniqueness theorem. Rather, the problem is that the evidence is not sufficient to determine, in the first instance, the admissible transformations that should be applied in order to co-scale different individuals' utilities. In other words, it is not only the case that the evidence is insufficient for showing that $u_i = u_j = u$. Rather, it also the case that the evidence is insufficient for individuating the admissible transformations λ_i and λ_j, which would allow us to co-scale individuals i's and j's utilities.

It is easy to understand this point if we consider the following analogy with the measurement of temperature. Consider a domain of objects T. For any four objects $x, y, w, z \in$ T, we can establish a temperature ranking on the basis of the empirical relation 'warmer than'. If this relation satisfies the axioms that are relevant for measurement, it can be represented though a cardinal scale that measures degrees of warmth, or, more commonly, the temperature of each of the objects in the domain. As it is well known, there are many interval scales of measurement that can be used to represent the warmth relation. For instance, we can use either a Celsius or a Fahrenheit scale. Suppose we measure the temperature of x on a Celsius scale C and the temperature of y on a Fahrenheit scale F. Suppose also that the numerical value representing the temperature of x and y is the same, e.g. 20°. We cannot conclude from this that both x and y are equally warm. As a matter of fact, the numerical value representing their temperature is relative to different scales of measurement. Therefore, it is more correct to say that the temperature of x is 20° C and the

temperature of y is 20° F. Since the scales are different, it is (at least) not obvious that the temperature of x and y is really the same.

We can think about the problem of IUCs in a similar way. When we get $u_i(y) = u_j(x) = 0.6$, we cannot conclude that the intensity of individual i's and j's preferences is really the same. After all, the scale of measurement representing i's preferences might as well be different from the scale of measurement representing j's preferences. For instance u_i could be a Celsius-like utility function, whereas u_j could be a Fahrenheit-like utility function. As a consequence, the fact that the utility value in correspondence of option x and y is the same is not sufficient to conclude that their preferences have identical strength.

The difference between the comparison of the temperature of different objects and IUCs is that in the former case, but not in the latter, the empirical evidence is sufficient to determine a function λ that transforms one scale into the other while preserving the same information. In other words, in the case of temperature, we can determine the admissible transformation that allows us to measure the temperature of different objects on the same scale. More precisely, for any object $x \in T$, we can convert the measurement from Celsius to Fahrenheit degrees (and viceversa) by means of the following formula: $C(x) = (F(x)-32)/1.8$. What makes the determination of the admissible transformation λ possible, and, more generally, the very determination of whether an object is measured on a Celsius or a Fahrenheit scale, is the existence of two common points with respect to which the temperature of every object can be compared, namely, the water's freezing point and the water's boiling point. The former is at 0° C, while it is at 32° F; the latter is at 100° C, while it is at 212° F.

According to the 'standard picture', choice behavioural evidence is insufficient to determine whether or not any such common point exists, with respect to which the intensity of different people's preferences can be measured and compared. As a consequence, not only it is not possible to determine the class of admissible transformations that would co-scale different people's utilities, but it is also not possible to determine exactly whether the utility function representing an individual's preferences is really the same as the utility function representing another individual's preferences or a different one.

2.5 Limitations

Let us take stock. The 'standard picture' characterises the problem of IUCs in terms of two features. First, it describes the problem by suggesting that, although the measurement is relative to the same type of measurement scale, different people's utilities may not be co-

scaled. This means that it is not possible to determine whether or not $u_i = u_j = u$. Second, it identifies the source of the problem in the insufficiency of choice behavioural evidence for co-scaling different people's utilities. This means that IUCs are underdetermined by choice behavioural evidence, in a sense that will be clarified below.

The worry is that this characterisation offers too narrow a view of the problem of IUCs. To begin with, the 'standard picture' identifies the potential sources of the problem of IUCs too narrowly. Insofar as choice behaviour is the only admissible evidence, the only clear reason why IUCs are problematic is that they are underdetermined by choice behavioural evidence. This seems to suggest that the problem is epistemological and due to the limited evidence available[13]. Yet, as I shall illustrate in section 3, IUCs may not only be underdetermined by further empirical evidence, but also indeterminate. If the latter is the case, the problem of IUCs is a metaphysical one.

Moreover, the 'standard picture' does not clarify how the problem of comparing utilities is related to the problem of comparing preference strengths. Since we have taken utility to be a numerical representation of preferences, one may think that these problems are identical. However, as I shall try to show in section 4 by introducing the notions of comparability and commensurability, the problems differ in some respects that are worth being considered.

Finally, the 'standard picture' insulates the problem of IUCs from current debates in philosophy of mind and epistemology. By ignoring contemporary philosophy of mind, it ignores recent advances concerning the question of the nature of mental states of mental states, the question of their meaning and the question of how we ascribe preference strengths to different individuals. Alternative accounts are likely to shape the problem of IUCs in different ways.

By ignoring contemporary epistemology, the 'standard picture' limits the scope of the inquiry. In fact, either it neglects the epistemological question of whether or not we can have (scientific) knowledge of, or (scientifically) justified, IUCs; or it implicitly assumes that underdetermination by choice behavioural evidence entails the impossibility of having (scientific) knowledge and (scientific) justification. However, as I shall illustrate in section 5, this may not be the case. Although choice behaviour is insufficient to determine IUCs, it does not follow that there are no other considerations that can give us (scientific) knowledge or, at least, (scientifically) justified ICs of preference strength.

[13] On the other hand, if the problem of IUCs is characterised in terms of underdetermination *simpliciter*, rather than underdetermination by choice behavioural evidence, then the 'standard picture' has stronger conceptual resources than suggested here. Indeed, as underdetermination can be defined relative to different, i.e. non-empirical, bases, the 'standard picture' may avoid at least the first of the limitations illustrated in this sub-section.

3. IUCs and empirical meaningfulness

The first limitation of the 'standard picture' concerns the identification of the source of the problem of IUCs. By refusing to go beyond choice behavioural evidence, the 'standard picture' precludes a more thorough analysis of the nature of the problem. An alternative consists in formulating the problem of IUCs as the problem of whether or not IUCs are empirically meaningful, as List has recently done, in the wake of Waldner's more dated analysis[14]. According to List, IUCs are empirically meaningful if and only if they are determined by empirical evidence. Instead, they are empirically meaningless if and only if they are either underdetermined by all the possible empirical evidence or indeterminate. As we shall see below, by distinguishing between underdetermination and indeterminacy, this approach highlights the distinction between the epistemological and the metaphysical sides of the problem of IUCs. I shall proceed by considering the notion of underdetermination and indeterminacy in general[15] and, then, by applying such a general analysis to the specific case of IUCs.

3.1 Underdetermination by the empirical evidence

Let us consider a theory T_1 and a set of empirical observations E_{T1}, describing observable phenomena. I shall say that, if E_{T1} plays a role in the determination of T_1, E_{T1} offers an empirical basis for the theory T_1. There is a variety of other considerations that may play a role in the derivation of T_1 in addition, or in substitution, to empirical observations. If they are of a non-empirical kind (e.g. pragmatic, moral, metaphysical, etc.), these considerations offer a non-empirical basis for the theory T_1.

A theory T_1 is empirically adequate with respect to E_{T1} if and only if T_1 implies E_{T1}, that is, if and only if all the observations can be deduced from the theory[16]. Furthermore, a theory T_1 is determined by a set of observations E_{T1} if and only E_{T1} implies T_1, that is, if and only if the theory can be deduced from the observations. Following List, we can then

[14] See LIST, C. [2003] and WALDNER, I. [1972].

[15] This section follows rather closely the illustration of underdetermination and indeterminacy given by GIBSON, R. [1986], PEIJNENBURG, J. and R., HÜNNEMAN, [2001], LIST, C. [2003], LEPORE, E. and K., LUDWIG [2005].

[16] It is intended that the relation between theory and empirical observations is always mediated by auxiliary, e.g. methodological, assumptions.

say that a theory T_1 is empirically meaningful if and only if it is determined by some set of observations E_{T1}[17].

The problem with this characterisation of empirical meaningfulness in terms of deductive inferences is that it is too strong. In fact, it implies that a theory T_1 derived on the basis of probabilistic inferences counts as empirically meaningless. It would thus be preferable to adopt a weaker characterisation, according to which a theory T_1 is empirically meaningful if and only if it is inferred by some set of observations E_{T1}, where the inference is either deductive or probabilistic. On the other hand, I regard the subsequent analysis to be largely independent from this distinction. Thus, for simplicity, in what follows I shall still refer to List's characterisation. The reader has simply to keep in mind that the same or similar considerations apply *mutatis mutandis* to the weaker characterisation.

Let us now consider a second theory T_2 and a set of empirical observations E_{T2}. Suppose E_{T2} is identical to E_{T1}. In other words, let $E_{T1} = E_{T2} = E$. If T_2 is empirically adequate with respect to E_{T2}, then we can say that T_1 and T_2 are empirically equivalent with respect to E. More generally, we can say that two theories are empirically equivalent if and only if the same observable facts figure amongst their implications. Typically, there is an infinite number of empirically equivalent theories, implying the same observable facts. Suppose that two empirically equivalent theories T_1 and T_2 make incompatible, or mutually inconsistent, claims. Clearly, since they imply the same observable facts, their incompatibility stems from their assumptions about entities and relations postulated to account for unobservable facts. Let us call theoretical terms those terms, in a theory, that refer to theoretical entities. Thus, if T_1 and T_2 make incompatible claims, they have incompatible conceptions of unobservable facts, that is, they make incompatible assumptions about theoretical terms and theoretical relations[18].

We can now define underdetermination as follows. A theory T_1 is underdetermined by a set of empirical observations E if and only if it is empirically adequate with respect to E, but it is not determined by E. The last condition is satisfied if there exists a theory T_2, which is both empirically adequate with respect to E and incompatible with T_1. It is worth noticing that underdetermination is relative to a specific set of observations E. This means that a broader set of empirical observations E^+ may be sufficient to show that one of the incompatible theories is empirically meaningful with respect to E^+ and the other is not. This

[17] See LIST, C. [2003], p. 232.
[18] See PEIJNENBURG, J. and R., HÜNNEMAN, [2001], p. 23.

can happen if, although both T_1 and T_2 are consistent with E, E^+ determines T_1, but not T_2[19].

3.2 Indeterminacy

It may happen that two theories T_1 and T_2 are underdetermined with respect to the set of all the possible available evidence, E^{max}. No matter how many observations we collect, T_1 and T_2 remain consistent with them and yet incompatible. This poses an interesting problem. T_1 and T_2 make incompatible claims as a consequence of positing alternative theoretical terms and relations. If, although referring to unobservable phenomena, such theoretical entities and relations do indeed exist, we can say that there is a fact of the matter as to which theory is the correct one. On the other hand, if neither the theoretical entities nor the relations postulated by the theories exist, then there is no fact of the matter as to which theory is the correct one. This is the case of indeterminacy[20].

More precisely, a theory T_1 is indeterminate if and only if it is underdetermined with respect to all the possible evidence E^{max} – that is, there exists an incompatible theory T_2 that is also empirically adequate with respect to E^{max} – *and* there is no fact of the matter as regard to which theory is the correct one. Following List, we can then say that a theory T_1 is empirically meaningless if and only if it is either underdetermined by a set of observations E^{max} or indeterminate[21].

Another way of defining indeterminacy is the following[22]. Let us assume that a term or a relation is purely theoretical if its content is exhausted by the role that it has for keeping track of observable facts. Then, we can say that a theory T_1 is indeterminate if and only if it is underdetermined with respect to all the possible evidence E^{max} and the theoretical terms and relations that the theory postulates are purely theoretical[23].

The content of a purely theoretical term or relation is exhausted by its role in keeping track of observable facts. This means that purely theoretical terms or relations do not refer or exist independently of their function within the theory. In other words, there is no fact of the matter about them. If the incompatible claims that two theories T_1 and T_2 stems from the assumptions made about theoretical terms and relations, then we have indeterminacy. There is no fact of the matter as regard to which theory is the correct one, because there is

[19] See LEPORE, E. and K., LUDWIG [2005], pp. 223-224.
[20] See PEIJNENBURG, J. and R., HÜNNEMAN, [2001], p. 23.
[21] See LIST, C. [2003], p. 232.
[22] The two definitions do not seem to be strictly equivalent. This poses the problem of how they are related. For simplicity, here I shall ignore this complication.
[23] See LEPORE, E. and K., LUDWIG [2005], pp. 224-225.

nothing more in the content of the conflicting theoretical terms and relations that the theories postulate than what is required for accounting for observable facts.

The previous definitions of indeterminacy raise the issue of when we can say that there is a fact of the matter or, alternatively, that the theoretical terms and relations postulated by incompatible theories are not purely theoretical. There are (at least) two readings of indeterminacy: an epistemological reading and an ontological one[24]. According to the former, our epistemology, that is, the evidence and methods through which we acquire knowledge, fixes our ontology, that is, what there is in the world. If this is the case, indeterminacy collapses into underdetermination by all the possible empirical evidence. If two theories are empirically equivalent with respect to E^{max}, then, necessarily, there is no case in which we can appeal to a fact of the matter to establish which theory is the correct one. Indeed, no further evidence is available to establish what the fact of the matter is. What there is cannot be established autonomously from our epistemology.

According to the second reading, instead, our ontology is relatively autonomous, although not necessarily completely independent, from our epistemology. Although indeterminacy implies underdetermination by all the possible empirical evidence, it does not collapse into that. Indeed, indeterminacy and underdetermination are on a par epistemologically, but not ontologically. The available empirical evidence is what makes a theory empirically justified – at least according to an evidentialist theory of epistemic justification – while the existence of a fact of the matter is what makes a theory true. Thus, the same set of observations can make two empirically equivalent, but alternative, theories equally justified, from an epistemic point of view. However, under an ontological reading, if there is a fact of the matter, only one of the two theories can be true; by contrast, if there is no fact of the matter, no issue of truth arises. As a consequence, it is possible to distinguish cases of radical underdetermination, in which the question of which theory is correct is meaningful, because there is a fact of the matter that can make one theory true and the other false; and cases of indeterminacy, in which the question of which theory is correct is meaningless, because there is no fact of the matter that can make either theory true.

Finally, there are different ways of understanding the nature of the fact of the matter – that is, what counts as fact of the matter – within an ontological understanding of indeterminacy. For instance, Quine understands the expression in a physicalistic way. Fact of the matter refers to physical facts. This implies adopting a specific ontological stance, according to which our ontology is entirely physicalistic. Ultimately, it is physical facts that

[24] My illustration of indeterminacy follows rather closely GIBSON, R. [1986].

make a theory true or false. Therefore, the additional condition for indeterminacy that there is no fact of the matter is equivalent to the condition that there are no further physical facts that can adjudicate between two empirically equivalent theories[25]. Here, however, I shall adopt an ontological reading of indeterminacy that does not take any specific ontological stance. Thus, I shall remain neutral about the nature of the fact of the matter. More precisely, I shall take fact of the matter to indicate whatever fact (physical or otherwise) can make a theory or a statement true.

In the light of this analysis, I shall distinguish the following expressions. I shall say that a theory, or a statement, has a factual basis if and only if there is a fact of the matter, that is, there are facts that can make the theory, or the statement, true. By contrast, I shall say that a theory, or a statement, has an empirical basis if and only if there is a set of empirical observations that can make the theory, or the statement, epistemically justified. It is worth emphasizing the nature of the distinction. Factual basis is an ontological notion, while empirical basis is an epistemological notion.

3.3 IUCs, underdetermination and indeterminacy

Let us now go back to IUCs. On the basis of our previous definitions, we can say that IUCs are empirically meaningful, with respect to an empirical basis E, such that $E \leq E^{max}$, if and only if they are determined by E, that is, if they are determined by a set of observations E. On the other hand, IUCs are empirically meaningless if and only if they are either (i) underdetermined by E^{max}; or (ii) indeterminate, that is, when they are both underdetermined by all the possible empirical evidence E^{max} and there is no fact of the matter about interpersonal preference strengths.

Let us consider again the example seen above, where $u_i(y) = u_j(x) = 0.6$ and $u_i(x) - u_i(y) = u_j(w) - u_j(x) = 0.4$. In the 'standard picture', choice behaviour is the only relevant evidence for the ascription of individual preferences. However, choice behaviour is not sufficient to determine either the interpersonal comparison of utility levels or the interpersonal comparison of utility differences in the example under consideration. The problem is that choice behavioural evidence can be consistently accounted for by two incompatible theories: a theory T_1, which maintains that different people's utilities are co-scaled and, thereby, concluding that *i* and *j* have the same preference strengths, in correspondence of the options with the same numerical values; and a theory T_2, which maintains that different people's utilities are not co-scaled and, thereby, concluding that *i*

[25] See GIBSON, R. [1986], pp.146-153.

and *j* have not the same preference strengths, in correspondence of the options with the same numerical values. Both T_1 and T_2 fit the same observable phenomena about each individual's behaviour. Their incompatibility derives from their assumptions about unobservable facts and relations, i.e. each individual's preferences and the way in which they relate to other individual's preferences in terms of strength. As such, IUCs are underdetermined by choice behavioural evidence.

This approach to the problem of IUCs emphasises two important things. The first is that, by restricting the set of admissible evidence to choice behaviour, the 'standard picture' adopts a particularly narrow empirical basis. As underdetermination is always relative to a body of evidence, the possibility remains open that a broader empirical basis may determine IUCs. The second is that the 'standard picture' hides the distinction between the metaphysical and the epistemological questions about IUCs. Either it makes it appear that the problem of IUCs can only be epistemological or it makes it appear that the underdetermination of IUCs by choice behaviour entails the claim that IUCs have no factual basis. In the light of the previous analysis, however, we can say that IUCs have no factual basis if and only if there is no fact of the matter about how different people's preferences compare in terms of strength. Since underdetermination is an epistemological notion, the underdetermination of IUCs by choice behaviour does not imply that there is no fact of the matter about how different people's preferences compare in terms of strength. As such, it does not entail the claim that IUCs have no factual basis. This is the case only if IUCs are indeterminate.

Finally, it is worth noticing that the notion of empirical meaningfulness is not equivalent to the notion of meaningfulness considered in section 2. IUCs are meaningful if they are invariant under the class of admissible transformations. The crucial point is that meaningfulness does not imply any restriction on the basis used to make IUCs. They can be formed by using an empirical basis or a non-empirical basis or a combination of both. By contrast, IUCs are empirically meaningful if and only if they are determined by the empirical evidence *alone*. This means that, in order to be empirically meaningful, IUCs must be invariant under a certain set of admissible transformations (that is, they must be meaningful) *and* formed on the basis of empirical evidence only.

4. IUCs, incomparability and incommensurability

The second limitation of the 'standard picture' is that it does not clarify how the problem of comparing different people's preference strengths and the problem of comparing

different people's utilities are related. This becomes clear if we formulate the problem of IUCs with respect to the notions of incomparability and incommensurability[26]. By doing this, we can also shed light on one way to present the problem of IUCs that is sometimes employed in the literature. Once again, I shall firstly illustrate these notions in general and then apply them to the case of IUCs.

4.1 Incomparability and Incommensurability

Following Chang's analysis, I shall say, as a first approximation, that two items are incomparable if no positive comparative judgment between them can be made[27]. More specifically, incomparability is relative to four common elements:

(1) a domain of objects D;
(2) a property Φ;
(3) a basis B;
(4) a set of positive comparative relations.

Let us consider each element in turn. First, incomparability is a relation between two or more objects in a domain of interest. For instance, we can say that a career as a clarinettist is incomparable with a career as a lawyer. The objects in the domain may belong to the same ontological category (e.g. only careers, only states of affairs, only persons, etc.) or to different ontological categories (e.g. persons and states of affairs). Second, incomparability is relative to a property Φ[28]. A property Φ is any respect in terms of which the objects included in the domain of interest can be compared. For instance, a clarinettist may be incomparable with a lawyer in terms of talent. Third, incomparability is relative to a basis. A basis is a set of considerations that can be used to compare the objects in the domain of interest. For instance, comparisons can be made with respect to empirical considerations, or moral considerations, or a combination of both, etc. Fourth, incomparability is relative to a set of positive comparative relations. A set of positive comparative relations includes any positive relation that can be made to establish a comparison between the items in the domain of interest. The issue of what are the relevant comparative relations is crucial in the literature. According to the Trichotomy Thesis, there are only three comparative relations,

[26] My illustration is based on, and develops, suggestions contained in various papers included in CHANG, R. [1997a].
[27] Cfr. CHANG, R. [1997b], p.2.
[28] Cfr. WIGGINS, D. [1997], pp. 53-54.

namely, 'more than', 'less than' and 'equal then'. For instance, with respect to goodness, the relative comparative relations are 'better', 'worse', or 'equally good'. However, many authors have recently suggested that there are other comparative relations, such as the parity relation or the rough equality relation[29].

We can now define incomparability more precisely. We can say that two items are relatively incomparable with respect to a property Φ, a basis B and a set of positive comparative relations, if it is not the case that either of the comparative relations holds between them. For instance, if no comparative relation can be established between the goodness of two different careers, these alternatives are value incomparable, or incomparable in terms of goodness. If two items are incomparable with respect to any property that they have in common, any basis and any set of positive comparative relations, then they are absolutely incomparable.

The notion of 'incommensurability' suggests the lack of a common measure. It shares with incomparability the four elements seen above, but it is also relative to an additional element, namely:

(5) a type of measurement scale.

Incommensurability is relative to a scale of measurement because two items in the domain of interest may be incommensurate, with respect to any property Φ, relative to one type of scale but not relative to another. For instance, the talent of two artists may be incommensurate, relative to an interval scale, but not relative to an ordinal scale of measurement[30].

We can say that two items are relatively incommensurable with respect to a property Φ, a basis B, a set of positive comparative relations, and a scale of measurement *f* of Φ-ness, if it is not the case that either of the comparative relations holds between them, with respect to their measures of Φ. For instance, if no numerical comparative relation can be established between the goodness of two different careers, then these alternatives are value incommensurable. If two items are incommensurate with respect to any property that they have in common, any basis, any scale of measurement and any set of positive comparative relations, then they are absolutely incommensurable.

We can distinguish two special cases. The first is the case of ontic incomparability (incommensurability). Two items are ontically incomparable (incommensurable) if and

[29] See CHANG, R. [1997b]. See also RABINOWICZ, W. [2004].
[30] See CHANG, R. [1997b], p.2.

only if, as a matter of fact, no comparative relation holds between the items in the domain of interest. The second is the case of epistemic incomparability (incommensurability). Two items are epistemically incomparable (incommensurable) if and only if it is impossible to know what ontic comparability relation, if any, holds between the items in the domain of interest. In the former case, the problem is metaphysical; in the latter case, the problem is epistemological.

4.2 What is the relationship between incomparability and incommensurability?

Let us assume, for simplicity, that the set of positive comparative relations is fixed. There are some interesting cases to consider. First, suppose that incommensurability and incomparability are assessed with respect to a different pair of property and basis. Clearly, incommensurability with respect to a specific pair of property and basis does not imply incomparability with respect to any alternative pair. That is, two items may be incommensurate relative to a property Φ_1 and a basis B_1, but they may be comparable relative to a property Φ_2 and a basis B_2. For instance, two careers may be incommensurate in terms of their goodness, relative to both empirical and moral considerations, but they may be comparable in terms of money, relative to empirical data only.

Second, suppose that incommensurability and incomparability are assessed with respect to the same property Φ. Once again, incommensurability with respect to a specific property does not imply incomparability with respect to the same property when comparability is relative to a basis different from the one used for commensurability. That is, two items may be incommensurate relative to a property Φ and a basis B_1, but they may be comparable relative to the same property Φ and a different basis B_2.

Third, suppose that incommensurability and incomparability are assessed with respect to the same pair of property Φ and basis B. In this case, incomparability seems to entail incommensurability. Likewise, incommensurability seems to entail incomparability. However, the latter case is true only when two items are incommensurable with respect to all types of measurement scale. In fact, the same basis may be insufficient for cardinal commensurability, but sufficient for comparability. In what follows, however, I shall ignore this complication.

Finally, it is worth noticing that incommensurability between two items does not necessarily entail absence of a common scale between all the items in the domain of interest. Some items in the domain of interest may be locally incommensurate and yet there may be no doubt that the items in the domain of interest are co-scaled, in the sense

that the function *f*, which assigns numbers to represent information about a property Φ, is the same for all items. For instance, local incommensurability may occur because the available evidence is insufficient for either of the relevant comparative relations to hold between the items, with respect to *some* measures of Φ only. On the other hand, if the available evidence is insufficient for either of the relevant comparative relations to hold between the items, with respect to *all* measures of Φ, then we have complete incommensurability between the items.

4.3 Comparing preferences and comparing utilities

As the previous sub-sections show, comparability is a relation-theoretic notion, whereas commensurability is a measure-theoretic notion. In the light of this analysis, one way to understand how the problem of comparing different people's preferences and the problem of comparing different people's utilities are related to each other is the following. We can say that the former is a particular case of the problem of incomparability; whereas the latter is a particular case of the problem of incommensurability. To explain why, let us proceed by examining how we can map the elements seen above into the case under consideration.

The first element for comparability is the domain of interest D. In the case under consideration, the domain is constituted by different individuals' preferences. The second element for comparability is a property Φ. Different individuals' preferences may be interpersonally compared with respect to different properties, e.g. their content, the goodness of their content, their strength, etc. Here the problem under consideration is the problem of comparing different individual's preferences with respect to their strength. The third element is a determining basis. As seen above, in the 'standard picture', economists typically ascribe preferences and related properties to agents on the basis of choice behavioural evidence only. The last element is a set of comparative relations. The 'standard view' adopts the Trichotomy Thesis, according to which different individuals' preferences are interpersonally compared, with respect to their strength, only in terms of the relations 'more than', 'less then' and 'equal then'. For simplicity, in this thesis, I shall follow this stance.

Relative to these elements, the problem in the 'standard picture' is whether or not different people's preferences are comparable with respect to their property of strength, choice behavioural evidence and the Trichotomy Thesis. We saw that commensurability is relative also to a type of measurement scale. Relative to this additional element, the problem in the 'standard picture' is whether or not different people's preferences are

commensurable with respect to their property of strength, choice behavioural evidence, the Trichotomy Thesis, and a utility representation.

This approach highlights three important things. First, as suggested in the previous section, it stresses the fact that choice behavioural evidence provides an empirical, but remarkably narrow, basis for comparison. Although preferences are incomparable (incommensurable) with respect to choice behaviour, they may turn out to be comparable (commensurable) with respect to either a broader empirical basis or to a non-empirical basis (i.e. moral, pragmatic, metaphysical, etc.) or a combination of both.

Second, it helps us understand the relationship between the problem of comparing preferences and the problem of comparing utilities. As incomparability and incommensurability can be assessed with respect to different pairs of property and basis, it may occur that different people's preferences are comparable with respect to a pair of property and basis, but they are not commensurable with respect to a different pair. For instance, preferences may be comparable with respect to the goodness of their object, relative to a moral basis, but may be incommensurable with respect to strength, relative to an empirical basis. Or else, they may be comparable, but incommensurable, with respect to the same property, if the basis is different. For instance, preferences may be comparable with respect to strength, relative to a moral basis, but they may be incommensurable with respect to strength, relative to an empirical basis. Finally, when incomparability and incommensurability are assessed with respect to the same property and basis, it may occur that different people's preferences are comparable, but not commensurable. For instance, this may happen if the relevant basis is sufficient for ordinal comparisons, but not for comparisons on a cardinal scale of measurement.

Third, this approach takes into account the difference between the epistemological and the metaphysical questions about IUCs. The former is the question of whether or not different people's preferences are epistemically comparable (commensurable) in terms of strength. The latter is the question of whether or not different people's preferences are ontologically comparable (commensurable) in terms of strength.

5. IUCs, knowledge and justification

The third limitation of the 'standard picture' is that it insulates the problem of IUCs from current debates in philosophy of mind and epistemology. By ignoring current philosophy of mind, it ignores recent theories addressing the question of the nature of preferences, the question of their meaning and the question of how preferences are ascribed to other

individuals. I shall disregard these issues now and postpone their discussion to chapters 3, 4 and 5, where I shall examine some alternative solutions that explicitly draw from recent advances in philosophy of mind. By ignoring current epistemology, the 'standard picture' insulates the problem of IUCs from issues concerning (scientific) knowledge and (scientific) justification. This gap is of particular relevance for the present purpose. As I stated in the introduction, the epistemological questions about IUCs is the main focus of this thesis. In order to illustrate it in more detail, I shall begin by introducing the relevant notions in general and then show how they apply to the case under consideration.

5.1 Knowledge

The standard analysis defines knowledge as justified true belief (JTB). Accordingly, an agent S has knowledge of a theory, or a statement, T if and only if S has justified true beliefs that T. Thus, the JTB account of knowledge fixes three conditions as individually necessary and jointly sufficient for knowledge. The first condition for an agent S to have knowledge of a theory, or a statement, T, is the belief condition. It requires that the agent S believes that T. The second condition is the truth condition. It requires that T is true. The third condition is the justification condition. It requires that the agent S is justified in believing that T. As such, knowledge implies justified beliefs, but not viceversa.

The debate concerning the justification condition is at the heart of theories of knowledge. Relevant issues are both whether or not the justification condition is necessary and, if it is, how it should be formulated. As far as the former issue is concerned, the standard argument in favour of the justification condition is that it is needed in order to rule out cases of epistemic luck. These are cases in which beliefs turn out to be true by mere accident or luck. Thus, having true beliefs is not sufficient for having knowledge, insofar as epistemic luck is possible. According to JTB, a true belief must be justified in order to count as a genuine instance of knowledge[31].

[31] The JTB account of knowledge has been challenged by two counterexamples presented by Gettier in 1963. See GETTIER, E. [1963]. Essentially, both of them show that having justified true beliefs is not sufficient for knowledge. More specifically, they show that the justification condition, in its original formulations, is not by itself sufficient to ensure that, in certain circumstances, an agent's beliefs are not true by mere luck. The so-called "Gettier problem", then, consists in specifying how the analysis of knowledge should be modified in order to be immune from these counterexamples. Although the "Gettier problem" has a crucial relevance for the analysis of knowledge, it does not have any specific bearing on the problem with which this thesis is mainly concerned, namely, the problem of IUCs. The opposite is true for the issues related to the justification condition. Therefore, in what follows, I shall proceed by assuming that the JTB account of knowledge is basically correct and illustrate the main accounts of the justification condition offered in the literature. The caveat is that, in order to turn true beliefs into knowledge, these accounts need to be suitably refined so to "degettierize" justified true beliefs.

5.2 *The justification condition*

The answer to the question of when a belief is justified requires specifying four things: (1) what justification means; (2) what makes a belief justified; (3) what is the basis of justification; and (4) what is the structure of justification.

The literature distinguishes two main notions of justification, namely, a deontological and a non-deontological one. Typically, according to the former, an agent S is justified in believing a theory, or a statement, T if and only if, in believing T, S is not violating any epistemic obligations. According to the latter, S is justified in believing a theory, or a statement, T if and only if S believes the theory, or statement, T on a basis that properly 'probabilifies' S's belief that T.

Amongst deontological conceptions, what makes a belief justified is the fulfilment of one's epistemic duties or obligations. The question of what these obligations are is substantial. Usually, they belong to the class of actions that contribute to the achievement of the main epistemic goals. In turn, these are identified with the achievement of a body of beliefs that has the optimal truth-falsity ratio.

Amongst non-deontological theories, there are two main approaches to the issue of what makes a belief justified: evidentialism and reliabilism[32]. According to evidentialists, justification comes from evidence. This means that an agent S is justified in believing a theory, or a statement, T if and only if S's evidence for T supports his belief that T. In other words, S is justified in believing that T if and only if S believes that T on the basis of the possession of adequate evidence[33]. According to reliabilists, instead, justification comes from the reliability of the process whereby the belief originates. This means that an agent S is justified in believing a theory, or a statement, T if and only if S's belief that T results from a reliable process. In turn, a process is reliable if it tends to produce true beliefs[34].

Second, the basis of justification can be either internal or external. The key idea of internalism is that what makes an agent's belief justified is internal to the agent, while externalism is simply the denial of internalism[35]. Finally, there are two views about the structure of justification: foundationalism and coherentism. According to foundationalists,

---

[32] Reliabilism can be conceived both as a theory of knowledge and as a theory of justification. For simplicity, here I shall ignore the former case.

[33] For a paradigmatic statement of this position, see FELDMAN, R. and E., CONEE [1985].

[34] For a paradigmatic statement of this position, see GOLDMAN, A. [1979].

[35] Internalism may come in two forms. On the one hand, accessibility internalism claims that justification for the agent's beliefs is internal because it is always directly recognizable by the agent. This means that that the agent is always in a position to know whether or not his beliefs are justified. On the other hand, mentalist internalism claims that the justification for the agent's beliefs is internal because what makes a belief justified is a mental state of the agent.

justification is structured like a building. Beliefs belonging to the superstructure are justified by other beliefs at the foundations. The latter beliefs are basic in the sense that their justification does not derive inferentially from other justified beliefs. Coherentism is the denial of foundationalism. It maintains that justification is structured like a web. There are no basic beliefs. Rather, each belief is justified in terms of other beliefs and justification is simply a function of the relationship between various beliefs. While internalism is typically associated with evidentialism and externalism with reliabilism, foundationalism and coherentism may be equally associated with either evidentialism or reliabilism.

5.3 Scientific justification

Scientific knowledge differs from garden-variety forms of knowledge by further constraining the satisfaction of the justification condition. Although scientific justification is a necessary condition for having scientific knowledge, it is not entailed by the more general justification condition. Consider the case of a prophet, who acquires good evidence through God's revelation. According to evidentialism, he is justified and yet his belief is not scientifically justified. Similarly, consider the case of a clairvoyant who forms a belief in a perfectly reliable way. According to reliabilism, he is justified and yet his belief is not scientifically justified. Whatever account of epistemic justification one favours, an additional condition needs to be met for scientific justification: the belief needs to be validated, or justified, in accordance with scientific standards.

One of the marks of science is that it leads to intersubjective agreement. Yet, not all cases of intersubjective agreement conform to scientific standards. Intersubjective agreement must be reached in conformity to specific requirements. Once again we can distinguish a deontological notion of scientific justification and a non-deontological notion[36]. According to the former, a belief about a theory, or a statement, T is scientifically justified if and only if it formed without violating the epistemic obligations accepted by the scientific community. According to the latter, a belief about a theory, or a statement, T is scientifically justified if and only if it is supported by a scientifically acceptable basis that properly 'probabilifies' it[37].

Within this field, we can distinguish an evidentialist and a reliabilist version of scientific justification. In the evidentialist framework, a belief about a theory, or a statement, T is scientifically justified if it is supported by evidence that is (a) public, (b) replicable; (c)

[36] See ADAM, M. [2007].

[37] Since the literature has found the deontological conception wanting, in what follows I shall mainly concentrated on the alternative, non-deontological, conception of scientific justification.

such as to lead to accurate and precise measurements of the relevant variables. In the reliabilist framework, a belief about a theory, or statement, T is scientifically justified if it is formed through methods and techniques that are (a) reliable and – crucially – *known* to be reliable on the basis of scientific evidence[38]; (b) replicable; and (c) such as to lead to accurate and precise measurements of the relevant variables.

Alternatively, I shall say that, in the evidentialist framework, a belief about a theory, or a statement, T is scientifically justified when it is inferred by evidence that satisfies the previous conditions, where the inference can be either deductive or probabilistic. Likewise, I shall say that, in the reliabilist framework, a belief about a theory, or a statement, T is scientifically justified when the methods and techniques whereby it is acquired satisfy the previous conditions. In both cases scientific justification may involve the use of non-empirical principles. In such circumstances, it is required that the adoption of these principles can be justified by appeal to reasons acceptable by the scientific community.

5.4 The epistemological problem of IUCs

One of the crucial epistemological questions about other people's mental states is whether or not we can have knowledge of, or, at least justified beliefs about, them. This question naturally descends from the problem of other people's minds in traditional philosophy of mind. Following Goldman's suggestion[39], we can think of the problem of comparing mental states as a particular case of the problem of mental ascription. As such, we can define the problem of ICs of preference strength as the problem of whether or not we can have knowledge of, or justified beliefs about, how different people's preferences compare in terms of strength. Moreover, we can define the problem of IUCs as the problem of whether or not we can have scientific knowledge of, or scientifically justified beliefs about, how different people's preferences compare in terms of strength. The latter definition follows from the fact that one of the conditions for scientific justification is that the relevant variables must be accurately and precisely measurable and that, as we have seen above, preferences are numerically represented through a utility function. From now on, therefore, when I speak about the problem of IUCs, I shall refer to it as the problem concerning the scientific knowledge or justification of ICs of preference strength.

As one of the necessary conditions for having knowledge is the truth condition, the question arises as to how we can conceive it in the case of IUCs. Roughly speaking, truth is

[38] See GOLDMAN, A. [1995a].
[39] See GOLDMAN, A. [1995a].

correspondence to what is the case. Earlier, I implicitly equated the notion of 'what is the case' with the notion of 'fact of the matter', where the latter is interpreted in an ontological sense. While the available evidence or the reliability of the relevant processes is what makes a theory, or a statement, T justified, the correspondence to the fact of the matter is what makes a theory, or a statement, T true. In the case of IUCs, correspondence to the fact of the matter is what makes IUCs true. In other words, the fact of the matter is the truth-maker of a belief about different people's preference strengths. Thus, the possibility of having knowledge of IUCs presupposes the existence of a fact of the matter about IUCs. That is, it presupposes an affirmative answer to the metaphysical question about IUCs. In what follows, I shall proceed by taking this presupposition for granted. Later – in the wake of the failure of the two main strategies offered to solve the epistemological problem of IUCs in the literature – I will turn again to this presupposition and consider whether it can be justified.

When can we say that there is a fact of the matter about preference strengths? If we adopt a physicalistic understanding of the nature of factuality, then we can say that IUCs have a factual basis if and only if it is possible, either in practice or in principle, to reduce preferences and their properties to neurophysiological states and properties. By contrast, we can say that there is no fact of the matter about preference strengths if and only if the reduction of preferences and their properties to neurophysiological states and properties is excluded in principle. On the other hand, if we adopt a more neutral understanding of the nature of factuality, then we can say that IUCs have a factual basis if and only if there are facts (physical or otherwise) that can make IUCs true; and that IUCs have no factual basis otherwise.

5.5 Relationships

I want to conclude by trying to map the relationship between the various notions introduced in this chapter. To begin with, the idea of empirical meaningfulness is clearly related to the evidentialist idea of scientific justification. According to List's characterisation, IUCs are empirically meaningful if and only if they are determined by the empirical evidence. No non-empirical consideration, apart from strictly methodological assumptions, is additionally required. Thus, if we hold an evidentialist position in terms of justification, then, if the subject S forms empirically meaningful IUCs on the basis of the evidence E, his belief is scientifically justified by the empirical evidence E.

At first sight the empirical meaninglessness of IUCs seems to threaten possibility of having scientifically justified beliefs about how different people's preferences compare in terms of strength. Once again, this is clearly the case if we adopt an evidentialist theory of epistemic justification. The argument is straightforward. If all the possible empirical evidence is insufficient to determine IUCs, then, if the empirical evidence is what makes IUCs scientifically justified, it follows that IUCs cannot be scientifically justified.

However, it is important to notice that this argument does not entail that it is *never* possible to have scientifically justified IUCs. Indeed, evidentialism is not the only theory of epistemic justification. If we adopt a reliabilist conception, what makes a belief justified is not the empirical evidence, but the reliability of the processes whereby the belief in question is formed. Thus, even if all the possible empirical evidence is insufficient to determine IUCs, these can nonetheless be scientifically justified, *provided* that they are acquired through reliable processes, that is, processes that tend to produce true beliefs. If we have independent grounds to prefer reliabilism to evidentialism as a theory of epistemic justification, we can still have scientifically justified IUCs despite the fact that they are underdetermined by all the possible empirical evidence. On the other hand, if the empirical evidence includes observations about the reliability of the relevant processes, then, if all the possible evidence underdetermines IUCs, it follows that IUCs cannot be scientifically justified even if we adopt a reliabilist theory of justification.

Nevertheless, there are other possibilities to have scientifically justified IUCs. Firstly, if we adopt evidentialism as our theory of epistemic justification, this is the case if IUCs can be probabilistically inferred from the empirical evidence, despite the fact that they cannot be deductively determined by it. Secondly, under both evidentialism and reliabilism, this is the case if there are non-empirical considerations that can break the underdetermination and that accord with appropriate scientific standards. Indeed, even if all the possible empirical evidence is insufficient to determine IUCs, these can nonetheless be determined and, thereby, scientifically justified, *provided* that there are other non-empirical considerations that have recognised evidential value. Together, all these remarks show that empirical meaningfulness is, at best, only a sufficient condition for scientific knowledge or scientific justification.

As seen above, the problem of whether or not IUCs are empirically meaningful is different from the problem of whether or not IUCs have a factual basis. The empirical meaningfulness of IUCs implies the existence of a factual basis, but not viceversa. Even if IUCs have a factual basis, the empirical evidence may be insufficient to determine them. Likewise, the empirical meaninglessness of IUCs does not imply that IUCs have no factual

basis. Even if all the possible empirical evidence E^{max} is not sufficient to determine IUCs and, thereby, renders IUCs empirically meaningless, it may still be the case that there is a fact of the matter concerning IUCs[40]. This is not true only if IUCs are indeterminate.

Both underdetermination by the empirical evidence and indeterminacy are relations between a statement, or a theory, T and an empirical basis E. On the other hand, both (relative) incomparability and incommensurability are relations between the items included in a theory, or a statement, T and a specific basis B. When incomparability and incommensurability are relative to an empirical basis E, they are related to underdetermination by the empirical evidence and indeterminacy. For instance, if incomparability (incommensurability) is due to underdetermination by the empirical evidence only, we have a case of epistemic incomparability (incommensurability). The empirical evidence is insufficient to determine whether or not any of the admissible relations required to establish a comparison between the items in the domain of interest hold. However, underdetermination with respect to an empirical basis E, does not entail epistemic interpersonal incomparability (incommensurability) with respect to a broader empirical basis E^+, or a different non-empirical basis with recognised evidential value, or reliable epistemic processes. On the other hand, incomparability (incommensurability) due to the indeterminacy of the comparative statement is logically equivalent to ontic incomparability (incommensurability) of the items in the domain of interest with respect to the property Φ. There is no fact of the matter about the property Φ that they have in common, so that the question about the truth of the comparison is meaningless.

In the 'standard picture', the problem of IUCs can be seen as the problem concerning the interpersonal commensurability of different people's preferences, relative to their strength, choice behavioural evidence, the Trichotomy Thesis and a suitable scale of measurement. The problem arises because IUCs are underdetermined by choice behavioural evidence. On the basis of the previous analysis, however, we can say that underdetermination by choice behavioural evidence does not entail the epistemic incomparability (incommensurability) of IUCs. Furthermore, the underdetermination of IUCs by choice behavioural evidence does not entail their ontological incomparability (incommensurability).

If we are interested in scientific knowledge, or justification, the problem of IUCs can be seen as the problem concerning the epistemic commensurability of different people's preferences, relative to their strength, a scientifically acceptable basis, the Trichotomy Thesis and a suitable scale of measurement. The reference basis can be either an empirical one or the union of an empirical and a non-empirical basis, provided that its adoption can

[40] *A fortiori*, when the empirical evidence taken into account is less than all the possible empirical evidence.

be justified in accordance with scientific standards. As we have seen above, the necessary presupposition is that, relative to the same elements, different people's preferences *are* metaphysically commensurable.

Although it is possible to characterise the problem of IUCs in this way, in what follows, in order to avoid confusion, I shall ignore the idea of epistemic comparability (commensurability) and use the notion of comparability (commensurability) only in the ontological sense.

6. Conclusion

In this chapter, I focused on the question of how we can formulate the problem of IUCs. I presented the 'standard picture' and discussed three of its limitations. The first is that it does not take into account the possibility that the problem is not only epistemological but also ontological. The second is it does not clarify the relationship between the problem of comparing the intensity of different people's preferences and the problem of IUCs. The third is that it neglects current debates in philosophy of mind and epistemology.

I suggested an alternative formulation of the problem of IUCs. According to it, the problem of IUCs is the problem of whether or not we can have scientific knowledge of, or scientifically justified beliefs about, how different people's preferences compare in terms of strength. This characterization is broad enough to provide a unified framework to discuss the concerns left unanswered by the 'standard picture' and to connect the alternative modes of presenting the problem of IUCs existing in the literature.

CHAPTER 2

Inferences to the best explanation

1. Introduction

In chapter 1, I defined the problem of IUCs as the problem of whether or not we can have scientific knowledge of, or scientifically justified, ICs of preference strength. In the 'standard picture', the problem arises because choice behaviour is sufficient for measuring each individual's preferences but not for determining whether or not different people's preference strengths are really the same when their utilities have the same value. That is, it is not possible to claim that different people's utilities are co-scaled. IUCs are underdetermined by choice behavioural evidence.

This is an obstacle for the possibility of having scientific knowledge of, or scientifically justified, ICs of preference strength. However, this obstacle is not in principle insurmountable. On the one hand, since underdetermination is always relative to a specific body of evidence, it may turn out that IUCs can be determined by gathering further empirical evidence, in addition to choice behaviour. On the other hand, other non-empirical considerations may help break the underdetermination by the empirical evidence and potentially lead to scientifically justified beliefs about how different people's preferences compare in terms of strength.

Both considerations are central features of the most common class of solutions existing in the literature. These solutions are token applications of a more general strategy, which attempts to solve the problem of IUCs by appealing to an inference to the best explanation kind of argument. In general, the idea is that a theory, or an assumption, is justified if it offers, or contributes to offering, the best explanation of a certain phenomenon. In turn, the criteria for individuating the best explanation typically include considerations such as explanatory power, simplicity, or parsimony. In the case of IUCs, the argument is that we are justified in assuming that different people's utilities are co-scaled insofar as this provides the best explanation of the empirical evidence, with respect to one, or more, of the considerations listed above.

In this chapter, my goal is to illustrate and assess this strategy, by focusing on the issue of scientific justification in particular. To begin with, we can distinguish two approaches pursuing the strategy under consideration, namely, a third-person approach and a first-

person approach. The former tries to connect the measurement of utility, conceived as the representation of preferences revealed by choices, to the measurement of more objective empirical proxies. In section 2, I shall illustrate the work done along these lines by the early Harsanyi, by Waldner and by List[1]. The latter approach attempts to reduce inter-personal comparisons to intra-personal comparisons of utility. In section 3, I shall illustrate the version proposed by the later Harsanyi[2]. In section 4, I shall argue that Harsanyi's first-person approach is unsuccessful on multiple grounds. In section 5, I shall argue, more generally, that all the solutions to the problem of IUCs based on an inference to the best explanation type of argument fail to show that IUCs can be scientifically justified. Finally, I shall summarise my findings in section 6.

2. Third-person approaches

2.1 Harsanyi's "Cardinal Welfare, Individualistic Ethics, and Interpersonal Comparisons of Utility"

Let us start from Harsanyi's work. Harsanyi offers his first defence of IUCs[3] four years after Arrow's claim that "the interpersonal comparison of utilities has no meaning"[4], which fixes the orthodox view about the subject for many years. Harsanyi is not interested in defending the possibility of IUCs[5], but rather in clarifying the "*logical basis* of such comparisons"[6]. Against Robbins, he wants to prove that IUCs are not value judgments "but rather factual propositions based on certain principles of inductive logic"[7]. Using the terminology introduced in the first chapter, we can say that Harsanyi wants to show that IUCs have a factual basis and that they can be scientifically justified.

Since my purpose is not exegetical, I shall present Harsanyi's argument in a way that is partly different from the one in the original text. Harsanyi takes utility to be a measure of preference satisfaction and suggests that the evidence for the ascription of preference satisfaction is given by both choice behaviour and (verbal and non verbal) expressions[8]. As

[1] See HARSANYI, J. [1955], WALDNER, I. [1972], LIST, C. [2003].
[2] See HARSANYI, J. [1977] and [1982].
[3] See HARSANYI, J. [1955].
[4] ARROW, K. [1963], p. 9.
[5] According to Harsanyi, this had already been done by Little. See LITTLE, I. D. M. [1957], chapter IV.
[6] See HARSANYI, J. [1955], footnote 20, p. 317. [Emphasis in the original]
[7] See HARSANYI, J. [1955], p. 320.
[8] In his "Cardinal Welfare, Individualistic Ethics, and Interpersonal Comparisons of Utility", Harsanyi distinguishes utility as a measure of people's satisfaction from two indicators of it: "their preferences as revealed by their actual choices, and their (verbal or non verbal) expressions of satisfaction or dissatisfaction in each situation". See HARSANYI, J. [1955], p. 317. My reading is different. Three reasons can be adduced as

for preferences, they coincide with choice behaviour, in a radical behaviouristic manner. By contrast, here I shall take preferences to be behavioural dispositions. Moreover, I shall take utility to be a numerical representation of the intensity of an individual's preferences, rather than a representation of preference satisfaction. With this interpretation at hands, Harsanyi can be seen as extending the evidence admissible for IUCs beyond choice behaviour, so as to include both verbal and non verbal behavioural expressions.

Even if we dispose of such a broader empirical basis, two problems remain according to Harsanyi. The first is the "metaphysical problem". The idea is that, even if we assume complete isomorphy between different individuals, with respect to the admissible evidence, it may still be the case that they have different mental states and, in particular, different preference strengths. As a consequence, even if different individuals' choice behaviour and expressive reactions are actually the same, their utility (functions) may be different.

Although this is a conceptual possibility, in his 1955 paper Harsanyi takes it to be no more than "a metaphysical curiosity"[9]. In order to block any "metaphysical" scepticism, Harsanyi suggests adopting a "principle of unwarranted differentiation", according to which "if two objects or human beings show similar behaviour in *all* their relevant aspects open to observation, the assumption of some unobservable hidden difference between them must be regarded as a completely gratuitous hypothesis and one contrary to sound scientific method"[10]. Harsanyi concludes that, if two individuals show the same choice behaviour and expressive reactions, they have the same (*absolute*) utilities. This inference is sanctioned by the "principle of unwarranted differentiation" and it is justified to the extent that such a principle is justified. According to Harsanyi, justification comes from the fact that the "principle of unwarranted differentiation" conforms to good scientific practice.

The second problem of IUCs is the "psychological problem". According to the early Harsanyi, this is the most serious difficulty in making IUCs, separate and independent from the "metaphysical problem". It arises because in the real world different individuals do not actually show the same choice behaviour and expressive reactions. Harsanyi suggests looking at the variables that actually determine different choice behaviour and expressive

---

justifications. First, Harsanyi's overall work is vitiated by many repeated ambiguities, so that a literal reading is not always the best way to get closer to the spirit of his analysis. Second, Harsanyi gives a purely behavioural characterization of preferences, in terms of choices, which is now completely outdated. In order to discuss the problem of IUCs, it is better to clearly separate choice behaviour as evidence for the ascription of preferences from preferences themselves. Finally, my reading may still be compatible with Harsanyi's idea that utility represents preference satisfaction, provided that we *also* make the assumption that the degree to which an individual's preferences are satisfied is measured by the intensity of his preferences.

[9] See HARSANYI, J. [1955], p. 317.

[10] See HARSANYI, J. [1955], p. 317 [Emphasis in the original]. It is worth noticing that, according to Harsanyi, this is not only a recommendation for scientific investigation, but the very principle that guides ordinary people's practice of third-person mental state ascription.

reactions and, ultimately, causally affect people's preference strengths and satisfaction. The analysis of how preference strength is connected to changes in those variables promises to offer an answer to the question of how different individuals' utilities compare when the relevant conditions are not the same.

According to Harsanyi, this can be done in two ways. The first method consists in re-creating a situation where the isomorphy of the relevant conditions holds in reality. This provides a "direct empirical solution" to the psychological problem[11]. If it is possible to manipulate a representative individual in such a way to subject him to different psychological determinants, then such an individual will be able to make a direct comparison of the utilities associated with the individuals that are actually subject to the same psychological determinants. The representative individual provides the metric for IUCs. The main difficulty for this method is that some of the relevant conditions cannot be changed. This means that there may be no representative individual that can be subject to all the relevant determinants.

The second method provides a less direct empirical solution. It starts by gathering observations about how different individuals' choice behaviour and expressive reactions are correlated to different psychological conditions. On the basis of our knowledge of the laws connecting psychological variables to preferences, we can ascribe degrees of preference to each individual that best explain interpersonal differences in choice behaviour and expressive reactions. In other words, the second method conceives IUCs as inferences to the best explanation, where the explanandum is given by differences in choice behaviour and expressive reactions across individuals and the auxiliary nomological information concerns the relation between psychological variables and preferences.

According to Harsanyi, this method faces two difficulties. The first is that our knowledge of the laws connecting psychological determinants to preferences is precarious. This may not be an irresolvable difficulty. In fact, Harsanyi believes that knowledge of psychological laws can be derived from further empirical research. However, an additional difficulty is that it is possible to gather empirical information only for those variables that are capable of change. Therefore, the second, and more serious, difficulty is that there is no direct empirical evidence to uncover the possible influence of unchangeable variables upon preferences.

The problem is that, when we make IUCs in order to explain different people's alternative behaviour, we can impute preference strengths only on the basis of the observation of changeable variables. Harsanyi's solution consists in postulating that

[11] See HARSANYI, J. [1955], p. 318.

unchangeable variables have no influence whatsoever on preferences and, thereby, do not affect choice behaviour and the expressive reactions to be explained. This is Harsanyi's "principle of unwarranted correlation"[12]. It is an a priori principle and cannot be subject to empirical scrutiny. Once again, its adoption is justified on the grounds that the principle conforms to good scientific practice.

To summarise, Harsanyi's solution to the "metaphysical problem" guarantees that different individuals' utilities are co-scaled, once they are derived from the 'correct' inputs, that is, 'correct' preference strengths. On the other hand, Harsanyi's solution to the "psychological problem" guarantees that the inputs used to derive interpersonal utilities are indeed 'correct'. This is done, on the one hand, by further empirical investigation about the relationship between preferences and psychological determinants and, on the other hand, by imposing a *ceteris paribus* stricture on variables that are not capable of change.

2.2 *Waldner's "The Empirical Meaningfulness of Interpersonal Utility Comparisons"*

Waldner's work on the problem of IUCs brings some improvements to Harsanyi's seminal contribution. To begin with, Waldner is the first to present the problem of IUCs in terms of the notion of empirical meaningfulness, although he does not provide an explicit definition of it. Moreover, he draws a wedge between preference satisfaction and preference strength and takes utility to be a numerical representation of the latter notion only. Finally, he abandons Harsanyi's untenable behaviourism and adopts a broad dispositional account of preferences[13]. As we have seen in the previous chapters, according to the radical behaviourist account – fashionable in the post-Robbins period – preferences are nothing but choice behaviour. This explains why the 'standard picture' takes choice behaviour as the only admissible evidence. A milder form of behaviourism conceives preferences as behavioural dispositions, that is, as mental states independent and separate from their behavioural manifestations. According to a narrow dispositional account, however, preferences are dispositions to cause choice behaviour only, or, which is the same, dispositions towards actions. As a consequence, choice behaviour remains the only type of admissible evidence for the ascription of preferences. By contrast, Waldner adopts a broad dispositional account, according to which preferences are dispositions towards a broad range of behavioural outputs, such as actions, primarily, but also "certain unintentional expressive reactions, facial expressions, thoughts, day-dreams, musings,

[12] See HARSANYI, J. [1955], footnote 27, p. 319.
[13] Strictly speaking, Waldner's analysis is formulated in terms of desires and not preferences. I shall ignore possible complications and consider preferences as relational desires. See WALDNER, I. [1972], p. 90.

etc."[14]. This view of preferences finally opens the possibility of extending the admissible evidence beyond mere choice behaviour in a less ambiguous way than in Harsanyi's 1955 paper.

The first problem for Waldner is that there is no theory establishing, in a precise way, which behavioural outputs are connected to preferences. This problem can be solved by considering all the platitudinous beliefs about preferences that enjoy intersubjective agreement between the folks[15]. Everyday beliefs concerning how preferences and behavioural outputs are connected provide the first approximation towards a more scientific dispositional theory of preferences. This way, behavioural outputs other than choices can be used as additional evidence for IUCs, as they typically are in everyday cases. In order to "bring out the logical problems involved"[16], Waldner focuses not just on rough correlations, but on hypothetical precise laws. In the terminology introduced above, his working hypothesis is that, although rough correlations may suffice to justify everyday ICs of preferences, only well-established and precise laws may provide sufficient evidence for the scientific justification of ICs of preference strength.

Waldner is not interested in all the possible behavioural outputs connected to preferences; rather, he selects only two behavioural expressions, namely, latency of choice, that is, the time delay between the presentation of the option and the actual choice-making, and probability of choice, that is, the probability of choosing one option rather than another. Moreover, he considers laws that connect behavioural proxies only to differences in, and not levels of, preference strengths. Then, for each individual *i*, possible laws of the kind envisaged by Waldner take the following form: $P_i \rightarrow t_i(x/y) = f_i(u_i(x) - u_i(y))$ and $P_i \rightarrow p_i(x/y) = g_i(u_i(x) - u_i(y))$, where P_i is a general type of preference, $t_i(x/y)$ and $p_i(x/y)$ are, respectively, the latency of individual *i*'s choice of *x* rather than *y* and the probability of individual *i*'s choice of *x* rather than *y*, and $u_i(x) - u_i(y)$ is the difference in intensity of individual *i*'s preference for option *x* over option *y*, determined on the basis of choice behavioural evidence only[17].

At first sight, the goal is to discover functional relations *f* and *g*, unique for all individuals, which connect intervals of preference strengths, individuated through choice behavioural evidence, to other proxies, such as latency of choice and probability of choice. One problem is that the representation of preference strengths through a utility function is unique only up to a positive affine transformation. As a consequence, the evidence offered

[14] See WALDNER, I. [1972], p. 95.
[15] This strategy clearly reminds Lewis' version of commonsense functionalism. See LEWIS, D. [1972].
[16] See WALDNER, I. [1972], p. 96.
[17] See WALDNER, I. [1972], p. 97.

by choice behaviour and other behavioural proxies is not sufficient to determine absolute functions f and g connecting preference strengths to latency of choice and probability of choice, respectively, but it can only determine a family of functions preserving the same information. Unless there is an independent way to determine absolute utility functions in the first instance, it is also impossible to determine absolute functions f and g. Clearly, Waldner's conclusion contrasts with Harsanyi's hope of deriving absolute utility functions by collecting further empirical evidence about psychological laws. The role of empirical investigation is different in Waldner's account. It aims at individuating the admissible transformations α and β, unique for all individuals, such that any one of the functions establishing a connection between an individual's preference strengths and, respectively, latency of choice and probability of choice, can be transformed into one of the same functions of another individual. More formally, the goal is to individuate the unique admissible transformations α and β, which preserves the following correlations: $P \rightarrow [t_i(x/y) = t_j(x/y) \leftrightarrow u_i(x) - u_i(y) = \alpha (u_j(x) - u_j(y))]$ and $P \rightarrow [p_i(x/y) = p_j(x/y) \leftrightarrow u_i(x) - u_i(y) = \beta (u_j(x) - u_j(y))]$, for any individuals i and j[18].

This task presents two difficulties. First, for each individual, both latency of choice and probability of choice may depend on variables other than preference strength. This complicates the generalization of the admissible transformations α and β across individuals, that is, the individuation of αs and βs unique for all individuals. At best, the laws connecting preference strengths to latency of choice and probability of choice are only *ceteris paribus* laws. As such, the issue concerns the individuation of the conditions that should be included in the *ceteris paribus* clauses. This was part of Harsanyi's "psychological problem" of IUCs. Unlike Harsanyi, however, Waldner does not resort to any non-empirical regulatory principle, but confides that further empirical research will be able to unpack the relevant conditions.

The second difficulty is that there may be hidden differences in "sensitivity" across individuals which may prevent the individuation of unique admissible transformations α and β, even in the case in which different individuals show equal latency of choice and probability of choice. This is Harsanyi's "metaphysical problem" again. According to Waldner, this claim raises a purely conceptual issue. No further empirical evidence could ever prove such a special "sensitivity", since any further empirical evidence would be used to refine the *ceteris paribus* clauses annexed to the laws under consideration. Like Harsanyi, Waldner suggests handling the problem by resorting to "the requirement to strive

[18] See WALDNER, I. [1972], p. 99.

for simplicity of empirical theories", which is coded in "the principle of not postulating any differences unless there is some reason to do so"[19].

2.3 *List's "Are Interpersonal Comparisons of Utility Indeterminate?"*

Thirty years later, List generalizes, and provides new justifications for, Waldner's approach. List follows Waldner in presenting the problem of IUCs as the problem of whether or not IUCs are empirically meaningful. In addition, he provides a detailed definition of the notion of empirical meaningfulness, which brings the problem of IUCs closer to issues in philosophy of science. Moreover, List adopts a broader conception of utility than Waldner, which covers a whole range of different interpretations. For the present purpose, I shall still take utility to represent preference strengths only. The key property is that utility, so defined, "may *surface observably* in the form of a person's choice behaviour and/or other observable proxies"[20]. This allows for the possibility that further empirical evidence, in addition to choice behaviour, may be considered for IUCs. Finally, List generalizes Waldner's approach in two ways. First, he considers all the potential types of empirical evidence by dividing the proxies into classes, in accordance with the type of scale through which they are measured. Second, he considers the implications that the use of these empirical proxies has not only for ICs of utility differences, but also for ICs of utility levels and IUCs with respect to an interpersonally significant zero-line[21].

List investigates the unwelcome prospect that IUCs are empirically meaningless in the most favourable and ideal case in which one disposes of a particularly rich set of empirical evidence E. Such a set includes "a person's observable facial expression of pleasure or pain, a person's relevant neural activity, in response to the options or in response to switches between options", in addition to Waldner's latency of choice and probability of choice[22]. If the relevant conditions are satisfied, these observable proxies can be measured, for each individual i, through proxy functions f_i , g_i and h_i having different uniqueness properties. Moreover, if the proxies behave in a way that is consistent with individual preferences, then, for each individual i, it may possible to find, respectively, admissible transformations φ_i, ψ_i, $sign_i$ that connect these proxy functions to some profile of utility functions, where utility is taken here to be a representation of individual i's preferences as

[19] See WALDNER, I. [1972], p. 102.
[20] LIST, C. [2003], p. 229. [Emphasis in the original]
[21] See LIST, C. [2003], p. 243.
[22] See LIST, C. [2003], p. 243 and p. 245.

revealed by choice behaviour. The resulting relations look as follows: $f_i(x) = \varphi_i(u_i(x))$, $g_i(x, y) = \psi_i(u_i(x) - u_i(y))$, $h_i(x) = sign_i(u_i(x))$[23].

There are two difficulties. First, the proxy functions *f*, *g* and *h* may present the same problems of measurability, uniqueness and comparability of the utility function *u*. In particular, for any two individuals *i* and *j*, it might not be possible to determine whether their respective proxy functions are co-scaled. Second, it might be the case that the admissible transformations φ, ψ and *sign* are not the same across individuals. Both conditions should be satisfied, in order for IUCs to be empirically meaningful. The idea is that, for any two individuals *i* and *j*, if their proxy functions *f*, *g* and *h* are co-scaled, and if they have identical admissible transformations φ, ψ and *sign* then it is possible to take interpersonally comparable proxy measures as arguments and transform them into interpersonally comparable utility measures, by applying the relevant admissible transformation.

List avoids the first difficulty by stipulation. In particular, he claims that "what makes f_i, g_i and h_i *observable* is that, whatever scale of measurement we choose, this scale is a *common* one for all persons"[24]. However, no similar solution is available for the second difficulty, on pain of begging the question. Even if we dispose of all the admissible empirical evidence, we cannot determine whether different individuals have identical admissible transformations φ, ψ and *sign*.

Two conclusions follow. First, IUCs are underdetermined by the empirical evidence in a particularly robust way. Even if we extend the set of empirical evidence E beyond choice behaviour, IUCs remain underdetermined. Furthermore, since they are underdetermined with respect to all the possible empirical evidence E^{max}, IUCs are potentially indeterminate, if it turns out that there is no fact of the matter about preference strengths. Notice that if we assume that all the relevant empirical evidence is the same across individuals, we are back to Harsanyi's "metaphysical problem", which appears now as a particular case of the more general problem of underdetermination of IUCs by the empirical evidence.

Second, one way to break the underdetermination and determine IUCs consists in assuming "interpersonal sameness of the conversion of utility into the proxy functions"[25]. This is equivalent to assuming that different individuals have identical admissible transformations φ, ψ and *sign*. However, this assumption is non-empirical, since it cannot be sanctioned by the empirical evidence. List offers a pragmatic justification for embracing it. The idea is that, even if the assumption of "interpersonal sameness of the conversion of

[23] See List, C. [2003], pp. 244-245.
[24] List, C. [2003], footnote 16, p. 259. [Emphasis in the original]
[25] List, C. [2003], p. 247.

utility into the proxy functions" cannot be empirically confirmed, in a realist sense, it should still be taken to hold in a pragmatic sense. In analogy with Quine's analysis of translation practice, such an assumption provides the most parsimonious way of accounting for different people's behaviour. In other words, the principle of parsimony provides the justification for the assumption that the admissible transformations φ, ψ and *sign* are identical across individuals[26].

3. First-person approaches

The three approaches seen above have various things in common. First, they are all third-person approaches to IUCs. Typically, they either limit or exclude the role of introspection as a source of evidence for IUCs. Instead, they assign a major role to empirical research in order to gain scientific knowledge of precise laws connecting preferences to other empirical proxies. Second, they all explore the use of non-empirical principles, in addition to the available empirical evidence, in order to determine IUCs. Finally, they all justify the adoption of these non-empirical principles by reference to pragmatic considerations that are derived from scientific practice. Therefore, according to these accounts, it is scientific practice that provides the ultimate standard for having scientifically justified ICs of preference strengths.

However, the appeal to pragmatic arguments is not peculiar to third-person approaches, but it is a central feature of first-person approaches to IUCs as well. The main characteristic of first-person approaches is that they offer solutions to the problem of IUCs that are based on an introspective reduction of inter-personal comparisons to intra-personal comparisons of utility. Typically, such a reduction involves constructing an extended preference ranking[27]. Although this approach has been variously explored by many authors[28], I shall here focus only on its most influential version: the one of the later Harsanyi.

3.1 Harsanyi's "Morality and Social Welfare"

Harsanyi considers the following setting. Suppose there are n individuals 1, ..., n. Let A = x, y ..., z be the preference domain, which includes lotteries describing the possible

[26] Furthermore, List suggests another (normative) way to break the underdetermination of IUCs by the empirical evidence. The idea is that if it is intersubjectively agreed that certain options or states of affairs are normatively significant, we can make meaningful IUCs of the types considered above.

[27] However, there are also approaches where the intra-personal reduction takes place without the construction of an extended preference ranking. See GIBBARD, A. [1986].

[28] See SEN, A. [1970] and [1979a] and ARROW, K. [1977], amongst the others.

objective situations in which an individual can find himself. By assumption, the set is the same for all individuals. Suppose that individuals have preferences over the lotteries in the set A. Let us denote by P_i the set of these preferences, for each individual $i = 1, \ldots, n$. P_i describes the possible subjective states in which an individual can find himself. If the expected utility axioms hold, preferences can be represented through a utility function u, unique up to a positive affine transformation. We can then derive a profile of utility functions $\{u_i\}$, for each individual $i = 1, \ldots, n$.

Suppose that an observer k wants to interpersonally compare individual i's and individual j's utilities. Harsanyi considers the following questions[29]:

(1) What kind of judgments are IUCs?
(2) How can the observer make IUCs?
(3) Do IUCs have intersubjective validity?

According to Harsanyi, IUCs are logically equivalent to preferences between extended alternatives. An extended alternative is an option including both a possible objective situation, i.e. a state of the world, and a possible subjective situation, i.e. a state of mind, in which an individual can find himself. Formally, it has the following form: $[x, P_i]$, for all states of the world x and for all individuals i. Thus, according to Harsanyi, an observer makes IUCs if and only if he forms preferences between extended alternatives (i.e. he forms extended preferences). Notice that this conception of IUCs marks a significant change in Harsanyi's work. In his 1955 paper, Harsanyi takes IUCs to be "factual propositions". They represent the world as it is supposed to be and they can be either true or false. As such, they are the objects of doxastic attitudes, such as beliefs, which have a mind-to-world direction of fit. This means that they are supposed to fit the world. Instead, in the 1977 paper, Harsanyi takes IUCs to be logically equivalent to extended preferences. This shows that, in Harsanyi's view, IUCs cease to be representations of the world. In fact, preferences are attitudes that have a world-to-mind direction of fit. This means that, unlike beliefs, the world is supposed to fit them. The crucial notion here is not truth, but satisfaction: preferences can be either satisfied or not satisfied, perhaps with different degrees[30].

---

[29] See HARSANYI, J. [1977].

[30] See HARSANYI, J. [1977] and [1982]. There are probably several reasons why Harsanyi switches from the third-person approach defended in the 1955 paper to the first-person approach firstly defended in the 1977 paper. Here is one hypothesis. Recall that one of the problems of the second method for solving the "psychological problem" of IUCs was that it relied on knowledge of laws connecting psychological determinants to preferences. In 1955, Harsanyi is optimistic that further empirical research will lead to

How can the observer make IUCs then? The starting point consists in considering the preferences that each individual has in correspondence with each objective state of the world. The profile of utility functions $\{u_i\}$ conveys information about the structure and properties of each individual *i*'s set of preferences P_i. However, such information is insufficient to ground IUCs, because it does not imply anything about how different people's preferences compare. The reduction of inter-personal utility comparisons to intra-personal utility comparisons provides the way for comparing different people's preference strengths. The fundamental operation is imaginative empathy. According to Harsanyi, an observer *k* can compare the intensity of an individual *i*'s preferences for option *x* with the intensity of an individual *j*'s preferences for option *y* simply by imagining being in the place of, respectively, individual *i* in state of the world *x* and individual *j* in state of the world *y*. Through imagination, the observer *k* forms preferences between the extended alternatives $[x, P_i]$ and $[y, P_j]$. By considering all states of the world and all individuals, the observer *k* forms a set of preferences between extended alternatives combining different pairs of individuals/states of the world. Finally, if these preferences satisfy the expected utility axioms, they can be represented by an extended utility function of the following form: $v_k[x, P_i]$.

In order to reduce inter-personal comparisons of utilities to intra-personal comparisons of extended utilities, the following conditions need to be satisfied:

(1) Rationality of individual preferences over simple alternatives;

(2) Rationality of individual preferences over extended alternatives;

(3) The Principle of Acceptance.

Conditions (1) and (2) express the standard consistency requirements necessary for representing simple and extended preferences through, respectively, a utility function and an extended utility function. Condition (3) is crucial, because it connects simple to extended utilities. The Principle of Acceptance requires that, when considering extended

---

discovering these laws. Despite the possibilities offered by future scientific progress, however, the problem concerning IUCs made from a third-person perspective remains twofold. One difficulty regards the individuation of the relevant conditions determining an individual's preferences. The other difficulty regards the absence of well-established and precise psychological laws. Therefore, a third-person approach is potentially subject to two different kinds of mistake. By contrast, a first-person approach promises to reduce the impact of, at least, the second difficulty. On a realist interpretation, psychological laws represent the actual working of an individual's mind. In particular, they represent the way in which an individual exposed to certain environmental conditions forms his mental states. One may argue that, even if the individual does not have explicit knowledge of the relevant psychological laws, he may put them to work through imagination. If the individual's mind works in a similar way in the imagined as in the actual case, the problem of IUCs becomes only a matter of individuating the relevant conditions determining different individuals' preferences.

situations $[x, P_i]$, ..., $[z, P_i]$, the observer's extended preferences agree with individual *i*'s simple preferences. In other words, when the observer imagines being in different states of the world but with the same preferences as individual *i*, his extended ranking must agree with individual *i*'s simple ranking. According to Harsanyi, the Principle of Acceptance implies that, for any state of the world *x* and for any individual *i*, there exists a function u_i, such that $u_i(x) = v_k[x, P_i]$. Since, trivially, there exists a function u_k, such that $u_k(x) = v_k[x, P_k]$, the Principle of Acceptance connects each individual's simple utility to the observer's extended utility. As such, the Principle of Acceptance is the basic condition for the reduction of inter-personal utility comparisons to intra-personal extended utility comparisons.

Conditions (1) – (3) account for how an observer *k* makes IUCs. The question arises whether or not IUCs are intersubjectively valid. The Principle of Acceptance merely implies that, for *any* observer *k* and for all individuals *i*, the observer's extended preferences agree with each individual's simple preferences. However, the Principle of Acceptance does not imply that two different observers compare the extended alternatives including different individuals' subjective attitudes in the same way. For instance, suppose that an observer *k* and an observer *h* compare the extended alternatives $[x, P_i]$ and $[y, P_j]$, that is, form a preference relation over these extended alternatives. Nothing implies that their preferences will be the same. Even if the Principle of Acceptance holds, *k* and *h* may form different extended preferences. Does this mean that IUCs are inherently subjective and they do not possess intersubjective validity? Harsanyi thinks otherwise.

The starting point is the idea that IUCs are intersubjectively valid if and only if people form the same extended preferences. The further claim is that, if certain conditions are satisfied, people are bound to form the same extended preferences. Therefore, it is possible to have intersubjectively valid IUCs. Harsanyi's argument is based on three important assumptions that are worth examining. The first assumption is that (both simple and extended) preferences are determined by causal variables. More formally, for each individual *i*, there is a function f_i that maps a vector of causal variables $\langle c_{i1}, c_{i2}, \ldots, c_{im}\rangle$ into specific preference relations, where $c_{im} \in C_i$ and C_i is the set of causes. This assumption has one important implication. As we have seen, according to Harsanyi, comparing different individuals' preferences is logically equivalent to forming preferences over extended alternatives of the form $[x, P_i]$, where *x* is an objective state of the world and P_i is the set of individual *i*'s simple preferences. Now, if preferences are determined by causal variables, we can write the extended alternative $[x, P_i]$ as $[x, C_i]$, where C_i is the set of causal determinants of individual *i*'s preferences, because $[x, C_i]$ implies $[x, P_i]$.

Moreover, if we consider the numerical representation of extended preferences, we can write the extended utility $v_k[x, P_i]$ as $v_k[x, C_i]$, since the latter implies the former.

In accordance with the position already expressed in the 1955 paper, Harsanyi thinks that differences in either the determining causal variables or their properties explain the empirical fact that people have simple preferences with different intensity. At first sight, the same can happen in the case of extended preferences. That is, differences in the determining causal variables may lead people to form extended preferences with different intensity. Yet, Harsanyi thinks that people are bound to form the same extended preferences. The next two assumptions are crucial for his conclusion. We have seen in the previous paragraph that each observer considers extended alternatives that can be expressed in the following way $[x, C_i]$, where x is an objective state of the world and C_i is the set of causal variables determining individual i's preferences. According to Harsanyi, when considering such alternatives, each observer imagines being in the place of individual i, that is, each observer imagines facing the objective properties and the causal conditions that individual i faces. Crucially, Harsanyi thinks that imagining facing individual i's causal variables is the same as imagining being subject to these causal variables. This is Harsanyi's second assumption. This assumption implies that, by considering the same extended alternative $[x, C_i]$, different observers imagine being subject to the same causal variables C_i.

Suppose that the second assumption is sound. Something else is required for Harsanyi to conclude that different observers will form the same extended preferences. Indeed, even if they imagine being subject to the same causal variables, they may still form different extended preferences because the mechanisms governing preference formation are interpersonally different. Harsanyi's third assumption is that the laws of human psychology are the same for all individuals. In a more formal terminology, we can say that the function mapping causal variables into individual preferences is unique for all individuals. This means that, for any two individuals k and h, $f_k = f_h = f$[31]. This assumption is crucial. It excludes the existence of fundamental differences concerning the way in which people form their (both simple and extended) preferences. Harsanyi labels this assumption the "similarity postulate". According to it, "once proper allowances have been made for the empirically given differences in taste, education, etc., between me and another person, then it is reasonable for me to assume that our basic psychological reactions to any given

[31] Since the function mapping causal variables into individual preferences may not be absolutely unique, perhaps it is more precise to say that the *family* of functions mapping causal variables into individual preferences is the same across individuals.

alternative will be otherwise much the same"[32]. According to Harsanyi, the "similarity postulate" is "a nonempirical a priori postulate". Its justification is entirely pragmatic and relies on the idea that a theory embracing such a postulate is "less arbitrary" than any other empirically adequate theory[33].

If the third assumption holds, different observers considering extended alternatives of the form $[x, C_i]$ will not only be subject to the same causal variables C_i, but will also form the same extended preferences on the basis of such causal variables. Thus, different observers k and h will have the same extended utility functions $v_k = v_h = v$ and the same extended utilities $v_k[x, C_i] = v_h[x, C_i] = v[x, C_i]$. Given that, according to Harsanyi, $u_h(x) = v_h[x, C_h]$, the following identities hold when $h = i$: $u_h(x) = v_h[x, C_h] = v[x, C_h]$. Therefore, Harsanyi shows that there exists a universal extended utility scale, which is connected to each individual's simple utility function. The conclusion is that any two individuals k and h will have the same extended utility functions and, thereby, will make the same, intersubjectively valid, ICs of simple utilities.

To summarise: the Principle of Acceptance assures that the reduction from inter-personal utility comparisons to intra-personal extended utility comparisons is possible. The assumptions that preferences are determined by causal variables, that imagining facing certain causes is the same as imagining being subject to these causes and that the psychological laws on the basis of which preferences are determined are the same for all individuals assure that the introspective reduction determines intersubjectively valid IUCs. If IUCs are logically equivalent to extended preferences and extended preferences are the same for all individual, IUCs can be determined by the combination of empirical evidence and a non-empirical postulate, i.e. the "similarity postulate".

4. Troubles for Harsanyi's first-person approach

According to Harsanyi, forming identical extended preferences is a necessary and sufficient condition for having intersubjectively valid IUCs. In this section I will argue against the necessity claim. My line of thought is the following. Harsanyi argues that comparing different individuals' simple preferences requires forming extended preferences. Moreover, he argues that forming extended preferences requires imagining being subject to the same causal circumstances to which the individuals under comparison are subject. The latter stage is required in order to understand which preferences those individuals have in

[32] See HARSANYI, J. [1982], p. 50.
[33] See HARSANYI, J. [1982], p. 51.

various objective states of the world. Following the recent literature on mindreading, I shall claim that such understanding involves forming a specific type of hypothetical preferences, i.e. pretend preferences, which are different from extended preferences. Then, I shall argue that, if extended preferences are based on pretend preferences, forming extended preferences is entirely redundant, because pretend preferences are sufficient for making IUCs.

In the remaining of this section, I shall also consider some other reasons to reject Harsanyi's extended preference approach. In particular, I will sketch Broome's argument against the assumption that imagining facing the same causal variables determining an individual's preferences is the same as imagining being subject to the same causal variables determining his preferences. Finally, in the next section, I shall consider further reasons against Harsanyi's approach by examining and rejecting the justification that Harsanyi gives in support of his third assumption, i.e. the assumption that psychological laws are the same for all individuals.

4.1 Early objections

Harsanyi claims that imaginative empathy is the crucial capacity that an observer must possess for making IUCs. Imagination is required for understanding the intensity of other individuals' preferences in specific objective situations from a first-person perspective. In turn, such understanding is required for constructing an extended preference ranking with the properties that Harsanyi envisages. In the course of the years, Harsanyi's proposal has raised various objections, which are worth examining before presenting my own objection.

It seemed to several authors that, in Harsanyi's approach, understanding another individual's preferences through empathic identification presupposes that the observer can have the very same preferences as the observed individual. The first objection is that this is impossible. The idea is that an observer *k* cannot have the same preferences as an observed individual *i* while remaining himself[34]. However, and contrary to a widespread opinion, this objection is not too damaging. It can be dealt with by simply weakening the notion of sameness of preferences that the objection implicitly relies on. More precisely, in order for individual *k* to have the same preferences of individual *i*, it is not required that individual *k*'s preferences are individual *i*'s very own preferences. Rather, it is only required that individual *k*'s preferences are similar to individual *i*'s preferences in certain relevant respects, namely, their content, their strength and their other structural properties. One way

[34] See, for instance, MACKAY, A. F. [1986].

to capture this reply is by making the assumption that preferences are independent from personal identity[35]. If this assumption holds, two individuals can have identical preferences while remaining fundamentally themselves.

This leads to a second objection. The idea is that, even if we adopt a weaker notion of sameness of preferences, imagination does not guarantee that the observer *k* will have the same preferences as the observed individual *i*. According to MacKay, the assumption underlying Harsanyi's approach is that if an observer *k* imagines being subject to the same causal variables as an individual *i*, then he has the same preferences as individual *i*. However, this is an instance of a more general principle, according to which if an individual imagines satisfying certain conditions, then he obtains the same results that would follow were those conditions actually satisfied. It is easy to find counterexamples to this principle. Consider the following counterfactual: if I imagine having been raised in France, I can speak French fluently (*ceteris paribus*). Clearly, imagining having been raised in France does not make me able to speak a single word of French. Therefore, the principle is false. Imagining that certain conditions are satisfied does not imply that the effect determined by the actual satisfaction of those conditions ensues[36].

One way to meet MacKay's challenge consists in reformulating the assumption underlying Harsanyi's approach in the following way: if an observer *k* imagines being subject to the same causal variables as an individual *i*, then *k imagines* having the same preferences as *i*. Two issues arise. First, one may wonder what exactly it means to say that individual *k* imagines having the same preferences as *i*. Second, one may wonder what exactly the outcome of individual *k*'s imagination is. We can provide an answer to both questions by turning our attention to a current approach to mindreading, namely, Simulation Theory (ST)[37]. ST conceives imagination as a sort of replication or re-enactment of another individual's mental life. The main idea is that, in order to predict or explain another individual's behaviour, the simulator uses himself as an analogue model. More specifically, the simulator imagines being in the other person's shoes, that is, he pretends to have the initial mental states of the other individual. These mental states are inputs fed into his practical reasoning system, which, as it is often put, runs off-line. If the

[35] This assumption can be interpreted also in the reverse way, as saying that personal identity is independent from personal preferences. This means that one remains fundamentally the same individual even if one has the same preferences of another individual. Cfr. MONGIN, P. [2001], pp. 156-157.

[36] See MACKAY, A. F. [1986], pp. 316-322.

[37] Since I shall discuss ST in greater detail in the next chapter, here I shall confine myself to a basic illustration. The current debate concerning ST starts with GORDON, R. [1986], HEAL, J. [1986] and GOLDMAN, A. [1989]. Three useful collections of papers discussing the early debate between Simulation Theory and its alternative, Theory-Theory, are DAVIES, M. and T., STONE, [1995a,b] and CARRUTHERS, P. and P. K., SMITH, [1996].

inputs are individuated correctly and the relevant processes are sufficiently similar across individuals, the outputs of the simulation heuristic correspond to the other individual's targeted mental states in certain crucial respects. In particular, if the goal of the simulation heuristic is to predict another individual's behaviour, the output will be either a pretend decision or a pretend intention. On the other hand, if the goal is to explain another individual's behaviour in terms of his mental states, the output will be some other pretend mental state.

Harsanyi's argument can be reformulated in the jargon of current ST. One difficulty is that, in Harsanyi's framework, the inputs include also non-mental causal variables, whereas the inputs of simulation are pretend mental states only. If we ignore this complication, we can say that, if individual *k* pretends being subject to the same causal variables as individual *i*, then *k* pretends having the same preferences as *i*, that is, *k* forms pretend preferences. Notice that, in order to have intersubjectively valid IUCs, it must be the case that *k*'s pretend preferences correspond to *i*'s actual preferences in the relevant respects. However, Harsanyi's similarity postulate is not sufficient for this. Indeed, even if the psychological laws determining preferences on the basis of *actual* causal circumstances are the same across individuals, imagination may lead one individual to form pretend preferences that are radically different from another individual's actual preferences, on the basis of *imagined* causal circumstances. In order for having intersubjectively valid IUCs, the additional assumption is required that the off-line working of the individual's mind-system approximates its online working. If this is the case, then, if individual *k* pretends being subject to the same causal circumstances of individual *i*, he forms pretend preferences corresponding to individual *i*'s actual preferences.

4.2 Challenging the necessity claim

Reformulating Harsanyi's argument in accordance with current ST can accommodate MacKay's objection. However, the success of this manoeuvre comes at a significant cost. Harsanyi's strategy becomes vulnerable to another objection: if extended preferences are based on pretend preferences, the construction of an extended preference ranking is entirely redundant because pretend preferences are sufficient for making IUCs. More specifically, if two observers form the same pretend preferences, they can make intersubjectively valid IUCs, without having to form also extended preferences.

Let us start by refining the distinction between extended and pretend preferences. An *extended preference relation* is a preference relation taking the form $[x, C_i]\ R_k\ [y, C_j]$. We

can interpret it as saying that individual k prefers the extended alternative $[x, C_i]$, including i's objective and subjective conditions, to the extended alternative $[y, C_j]$, including j's objective and subjective conditions, for any options x and y and for any three individual $i, j, k = 1, \ldots, n$[38]. Instead, a *pretend preference relation* is a special type of simple preference relation. More specifically, pretend preferences are simple preferences formed by imagining being subject to some specified causal variables. As such, they are hypothetical simple preferences. In what follows, I shall use the notation $(x;_P C_k)$ to refer to a pretend alternative.

Harsanyi's main claim is that forming pretend preferences is necessary in order to form extended preferences. We can formally represent this claim as follows: $[x, C_i] R_k [y, C_j] \rightarrow (x;_P C_i) R_k (y;_P C_j)$. How can we interpret this conditional? The interpretation of the left hand side is straightforward. It says that individual k prefers the state of the world x and the causal circumstances C_i to the state of the world y and the causal circumstances C_j. On the other hand, the interpretation of the right hand side is trickier. At first sight, it says that individual k prefers the state of the world x when imagining being subject to causal circumstances C_i to the state of the world y when imagining being subject to causal circumstances C_j. The problem with this interpretation is that it seems to require individual k to imagine being subject *simultaneously* to both causal circumstances C_i and causal circumstances C_j. In fact, in order to entertain a preference relation between the options $(x;_P C_i)$ and $(y;_P C_j)$, individual k has to represent both x and y while, respectively, imagining being subject to C_i and imagining being subject to C_j.

The intuitive reply is that individual k forms his pretend preferences by reiterating the imagination process. First, the observer k imagines being subject to causal circumstances C_i and forms pretend preferences for x. Second, he imagines being subject to causal circumstances C_j and forms pretend preferences for y. Third, he combines those preferences into a single pretend preference relation. The problem is that the third stage is more contentious than it seems at first sight. Consider the following characterisation of the simulation heuristics[39]:

SIM1	(a)	IDENTIFY with the other individual
	(b)	IMAGINE being in the same causal situation
SIM2	(a)	PRETEND to have the other individual's mental states
SIM3	(a)	CLASSIFY reactions

[38] Notice that it may be the case that $x = y$. Moreover, it may be the case that $i = j$, or $k = j$ or $k = i$ or that $i = j = k$.

[39] Here I am borrowing from (and adapting) PERNER, J. [1996], pp. 92-93 and p. 97 in particular.

(b) DE-IDENTIFY and ATTRIBUTE the last reaction to the other individual

According to SIM3, the simulator classifies the output of the simulation process and attributes it to the simulated individual. By so doing he forms a *belief* that the other individual has a certain mental state. In the case under consideration, first, individual *k* classifies the pretend output as a preference for *x* and then ascribes it to individual *i*. Then, he reiterates the procedure, that is, he classifies the output as a pretend preference for *y* and ascribes it to individual *j*. Eventually, individual *k* forms a belief about *i*'s preferences for *x* and a belief about *j*'s preferences for *y*.

This characterisation shows that the third stage of the simulation heuristic leads to the formation of beliefs rather than pretend preferences of the kind envisaged in the conditional seen above. Thus, extended preferences are based on pretend preferences in the sense that the observer *k* forms extended preferences on the basis of beliefs about individual *i*'s and individual *j*'s actual preferences, which, in turn, are formed on the basis of separate pretend preferences. At this point, however, a worry arises: it appears that forming extended preferences is unnecessary for making IUCs. Indeed, extended preferences are formed by drawing both on the mechanisms and on the inputs used to form pretend preferences. Yet, these pretend preferences seem to be perfectly sufficient for making IUCs. Indeed, in order to compare *i*'s and *j*'s respective preferences, individual *k* has simply to combine the beliefs about their preferences, which he has formed on the basis of his pretend preferences. The outcome is a further belief, i.e. the belief about how *i*'s preference for *x* compares with *j*'s preference for *y* with respect to strength. If the observer can use the outcomes of simulation to make IUCs, extended preferences appear to be entirely redundant and, therefore, unnecessary. The formation of pretend preferences leads to the interpersonal comparison of the individuals' believed preference strengths. If the assumptions about the simulation mechanisms are satisfied, different observers form the same pretend preferences and, thereby, the same beliefs about other individuals' preference strengths. By combining these beliefs, they form intersubjectively valid IUCs. Therefore, pretend preference formation is sufficient for making intersubjectively valid IUCs[40].

4.3 Other objections against Harsanyi's approach

[40] This also suggests adopting a cognitivist view of IUCs. According to it, and *pace* Harsanyi, IUCs are cognitive judgments such as beliefs rather than non-cognitive judgments such as extended preferences. For a similar objection see MONGIN, P. [2001]. This conclusion has some positive consequences. By modifying Harsanyi's non-cognitivist view into a cognitivist one, Harsanyi's first-person approach fit the presupposition that IUCs are factual judgments explored in this thesis.

The previous sub-section rejects the claim that extended preferences are necessary for making IUCs. Even if they were, however, there would be other reasons to dismiss Harsanyi's approach as incapable of providing a solution to the epistemological problem of IUCs. In particular, further objections may be raised about the assumptions on which Harsanyi's argument is based. In the next section, I shall present an argument against the pragmatic justification that Harsanyi offers in support of his third assumption, i.e. the assumption that the laws of human psychology are the same for all individuals. Instead, in this section, I shall briefly illustrate Broome's argument against Harsanyi's second assumption, i.e. the assumption that imagining facing the same causal circumstances as another individual is the same as imagining being subject to those causal circumstances.

Broome's main idea is that Harsanyi conflates the object of preference with the cause of preference[41]. When the observer *k* imagines facing the same causal variables under which individual *i* is subject, he considers the set of causes C_i as *object* of his (extended) preferences. By contrast, when the observer *k* imagines being subject to the same causal variables as individual *i*, he considers the set of causes C_i as *determinant* of his (pretend) preferences. In the former case, the observer considers the extended alternative $[x, C_i]$. In the latter case, he considers the pretend alternative $(x;_P C_i)$. Broome highlights the differences between the two cases by representing the second situation in the following way: $[x; C_i]$, with emphasis on the fact that the set of causes is not part of the object of preference, but is simply its determinant.

Broome's distinction has important consequences for the soundness of Harsanyi's proposal. Harsanyi's analysis starts from the observation that, if different individuals were subject to identical causes and if the psychological laws governing preference formation were also the same, they would have identical preferences. In ordinary life, causes are typically different and so are people's preferences. At first sight, the same can happen for all sorts of preferences, including extended preferences. That is, different people can form different extended preferences because they may be subject to different causes. Instead, Harsanyi thinks that people are bound to form identical extended preferences. The reason is that, if they consider the same extended alternatives, they will be subject to the same causal variables. His second assumption is crucial for this conclusion. Yet, if this assumption stems from a conflation of the object of preference with the cause of preference, then the soundness of this assumption is, at best, not obvious and, at worst, highly questionable.

[41] See BROOME, J. [1993] and [1999].

If we embrace Broome's distinction, Harsanyi's assumption requires that when different observers consider the extended alternative $[x, C_i]$, they consider the (pretend) extended alternative $([x, C_i];_P C_i)$ or, in Broome's own formal representation, the extended alternative $[x, C_i; C_i]$, for any individual *i*. The problem for Harsanyi is that forming extended preferences does not imply imagining being *subject* to the same causal circumstances that figure as object of extended preferences. Indeed, people can form extended preferences while being (actually or imaginatively) subject to different causal variables. For instance, when considering the extended alternative $[x, C_i]$, different observers *k* and *h* may indeed be considering the extended alternatives $([x, C_i]; C_k)$ and $([x, C_i]; C_h)$, respectively, where $C_k \neq C_h$. If this is the case, there is no guarantee that they will form the same extended preferences. In turn, this shows that there is no guarantee that they will make intersubjectively valid IUCs.

I take Broome's distinction to be correct. So I shall quickly move on to a more general objection against approaches based on pragmatic considerations. The conclusion of this sub-section is that Harsanyi's approach fails to show that there is an introspective method by which we can make intersubjectively valid IUCs.

5. Against pragmatic solutions

Let us take stock. The central question about IUCs is the question of whether or not we can have scientific knowledge of, or scientifically justified, ICs of preference strength. Recall that, in the 'standard picture', the problem arises because IUCs are underdetermined by choice behavioural evidence. Since underdetermination is only relative to a body of evidence E, the most natural reaction consists in trying to ground IUCs on a larger body of evidence E^+. The surprising result is that, even when we ideally dispose of the richest empirical basis E^{max}, as in the cases considered by Christian List, IUCs remain underdetermined by the empirical evidence. This means that IUCs are empirically meaningless in a particularly robust way.

Empirical meaninglessness threatens the possibility of having scientific knowledge of, or, at least, scientifically justified, IUCs. From an evidentialist point of view, the same empirical evidence supports two incompatible beliefs about how different people's preferences compare in terms of strength. From a reliabilist point of view, the empirical evidence casts doubt on the reliability of processes of belief formation that can give raise to incompatible beliefs about how different people's preferences compare in terms of strength. However, empirical meaninglessness does not entirely compromise the possibility of a

positive solution. After all, empirical meaningfulness is, at best, only a sufficient condition for having scientifically justified IUCs. Although a purely empirical basis does not guarantee the satisfaction of the conditions required for scientific justification, other, non-empirical, considerations may be added to meet the requirement.

The authors considered in the previous sections propose to add similar, although not identical, non-empirical principles in order to determine IUCs. In his 1955 paper, Harsanyi suggests adding two non-empirical principles: the "principle of unwarranted differentiation", according to which "if two objects or human beings show similar behaviour in *all* their relevant aspects open to observation, the assumption of some unobservable hidden difference between them must be regarded as a completely gratuitous hypothesis and one contrary to sound scientific method"[42], in order to deal with the "metaphysical problem"; and "the principle of unwarranted correlation", according to which unchangeable variables have no influence on preferences, in order to deal with the "psychological problem".

In the wake of the early Harsanyi, in his 1972 paper, Waldner rejects the hypothesis that some individuals may be more "sensitive" than others by resorting to "the principle of not postulating any differences unless there is some reason to do so"[43]. In his 1977 paper, instead, Harsanyi somehow combines the "principle of unwarranted differentiation" and "the principle of unwarranted correlation" into a single "similarity postulate", according to which "once proper allowances have been made for the empirically given differences in taste, education, etc., between me and another person, then it is reasonable for me to assume that our basic psychological reactions to any given alternative will be otherwise much the same"[44]. Finally, in his 2003 paper, List suggests basing the choice between alternative ways of accounting for people's behaviour, in order to make IUCs, on a commonsensical "principle of parsimony".

These principles have two things in common. The first is that they serve the purpose of breaking the underdetermination of IUCs by the empirical evidence. The idea is that, if the union of an empirical and a non-empirical basis determines IUCs, then IUCs can be scientifically justified, provided that all the other requirements fixed by scientific standards are met. The second thing is that these non-empirical principles are justified on the basis of an inference to the best explanation type of argument. In other words, it is argued that their acceptance leads to better explanations of people's behaviour. In turn, the goodness of an explanation is assessed in terms of considerations of simplicity, parsimony and explanatory

[42] See HARSANYI, J. [1955], p. 317. [Emphasis in the original]
[43] See WALDNER, I. [1972], p. 102.
[44] See HARSANYI, J. [1982], p. 50.

power. Since these are virtues that allegedly characterise good scientific practice, we can say that the adoption of these principles is vindicated by showing that it conforms to good scientific practice. To summarise, the accounts examined so far suggest that IUCs can be scientifically justified if they are determined on the basis of the union of an empirical basis and a non-empirical basis, whose adoption conforms to good scientific practice. This approach is the most common in the literature on IUCs. However, I shall argue that it is not successful.

The first objection concerns the epistemic value of pragmatic virtues such as simplicity and parsimony. As we have seen, the underdetermination of IUCs by the empirical evidence implies that there are at least two empirically equivalent, but incompatible, theories: a theory T_1, which assumes that interpersonal utilities are co-scaled, and a theory T_2, which assumes that interpersonal utilities are not co-scaled. A pragmatic argument of the kind under discussion adjudicates between the two by reference to some pragmatic advantages (simplicity and parsimony) that one theory (T_1), the one embracing non-empirical principles, has over the other theory (T_2), the one which does not. Parsimony and simplicity are certainly theoretical virtues. As such, they have pragmatic value. For instance, they can play a relevant role in theory choice, in the case of empirical underdetermination. However, in the case of IUCs, they do not merely have a methodological role, but also an epistemic role. This means that if a theory T_1 is simpler than an alternative theory T_2, then believing T_1 is epistemically justified. In virtue of what is this the case? One may argue that a more pragmatically advantageous theory is not just instrumentally preferable, in the sense that it is better than its alternatives with respect to the uses for which it is intended, but is also epistemically preferable, in the sense that it is more likely to be true. In other words, pragmatic virtues would be reliable means to arrive at true statements about the world. However – the objection goes – this may be extremely difficult to prove. It is questionable whether or not parsimony and simplicity have any epistemic value at all. That is, it is questionable whether they can be used to infer something about how the world really is[45]. In the case of IUCs, this implies that the adoption of non-empirical principles on the basis of pragmatic virtues has, at best, methodological value, but no epistemic value at all. As a consequence, IUCs may be, at best, intersubjectively agreed upon, but not scientifically justified.

Although worth discussing, I shall not pursue this line of criticism here. I think there are independent grounds for rejecting this strategy. My argument shall try to prove that, even if

[45] However, it is worth pointing out that several attempts have indeed been made to show that pragmatic virtues do have epistemic value. See SOBER, E. [2001] and [2003].

pragmatic considerations have epistemic value, they do not vindicate the adoption of non-empirical principles to break the underdetermination of IUCs by the empirical evidence, because the theory T_1, which assumes that interpersonal utilities are co-scaled, is not the most pragmatically advantageous theory.

Let us start with explanatory power. For the purpose at stake, we can leave the notion of explanatory power at a fairly intuitive level and say that it concerns the extent to which a theory is capable of explaining the phenomena that falls within its range. Let us consider the following question: does T_1 explain more aspects of individual behaviour than T_2? The answer is negative. The explanatory power of both theories is the same, despite the fact that they differ with respect to the assumptions of whether or not interpersonal utilities are co-scaled. The fact that T_1 allows one to make IUCs does not add anything to the power of the explanation. Let us see why.

One of the goals in the explanation of human behaviour is to account for individual behaviour in terms of causal entities and processes. In the case under consideration, this means explaining behaviour in terms of individual preferences, their causal property of strength and their relation with other mental states. Typically, empirical evidence offers a ground for positing entities and properties insofar they can causally explain that very same evidence. Although non-causal properties can have pragmatic relevance, the explanatory power of a theory is typically based on the extent to which the entities and properties postulated by the theory are able to causally account for the empirical evidence.

This poses a problem for IUCs. The property of being interpersonally comparable in terms of strength is not a causal property of preferences. That is, comparability plays no causal role in accounting for individual behaviour. As a consequence, T_1 and T_2 share the same causal properties, since they differ only with respect to the assumptions made about interpersonal comparability. Therefore, they have the same explanatory power. IUCs do not add anything to the explanatory power of a theory about individual behaviour[46].

This conclusion can be challenged. One may object that, contrary to what I have claimed, IUCs do add something to the explanation. For instance, if IUCs can be meaningfully made, we can offer allegedly comparative explanations of the following kind: individual *i* shows moderate appreciation rather than repulsion for tomato soup because he prefers tomato soup more than individual *j*. The first clause highlights a feature of individual *i*'s behaviour (i.e. appreciation for tomato soup) by comparing it with a feature

[46] This explains why the collection of further empirical evidence is typically used for purposes different from the ascription of degrees of interpersonal comparability. On the one hand, further evidence helps improve the understanding of the content of an individual's preferences. On the other hand, it helps refine the ascription of each individual's relative preferential strength for different options. It is not used to ground interpersonal comparability, because interpersonal comparability does not contribute to increasing explanatory power.

of individual *j*'s behaviour (i.e. repulsion for tomato soup). The second clause allegedly explains individual *i*'s behaviour by means of an interpersonal comparison of utility levels (i.e. individual *i* prefers tomato soup more than individual *j*).

What kind of explanation do statements like this provide? Let us keep in mind that what we want to explain is individual behaviour, that is, in our case, individual *i*'s choice and expressive behaviour. Clearly, we can compare two individuals' behaviour. For instance, we can compare individual *i*'s and individual *j*'s choices and expressive reactions. However, on the one hand, the interpersonal comparison of their behaviour does not imply anything about how the preferences that are supposed to explain each individual's behaviour compare in terms of strength. On the other hand, and most importantly, it is unclear how IUCs are supposed to explain individual *i*'s and individual *j*'s behaviour. The claim that individual *i* prefers tomato soup more than individual *j* is not explanatory. Rather, it *suggests* an explanation, i.e. an explanation of individual *i*'s behaviour in terms of the relative intensity of his preferences. If there were a common scale of preference strength and if we were to know the relative intensity with which individual *j* prefers the various options in the preference domain, then the claim that individual *i* prefers tomato soup more than individual *j* would allow the following inference: individual *i*'s preference for tomato soup is relatively less distant, in terms of strength, from the option that he prefers the most than it is for individual *j*. This would explain why their observable behaviour differs. However, this defence of the explanatory role of IUCs ignores one thing: we can equally explain individual *i*'s behaviour and why it differs from individual *j*'s behaviour in terms of relative preference strength, without assuming anything about how *i*'s preferences compare with *j*'s. Therefore, the assumption that different individuals' utilities are co-scaled is not necessary.

Let us now consider parsimony. I shall take parsimony to be defined with respect to the number and/or kinds of properties postulated. Thus, the most parsimonious theory is the one that explains the evidence with the least number and/or kinds of property assumptions. However, the assumption that utilities are co-scaled does not lead to a more parsimonious theory. Let us consider again the example illustrated in the first chapter. Suppose $u_i(y) = u_j(x) = 0.6$ on the basis of a broader empirical basis. The empirical evidence is consistent with two incompatible theories, T_1 and T_2. On the one hand, T_1 holds that utilities are co-scaled and, thereby, that *i* and *j* have the same preference strengths. Call the attitude conveyed by T_1 one of 'optimism'. On the other hand, T_2 holds that utilities are not co-scaled and, thereby, that *i* and *j* have not the same preference strengths. Call the attitude conveyed by T_2 one of 'scepticism'. If T_1 and T_2 are the only theories available, it is hard

not to conclude that the first is more parsimonious than the second. It characterizes the individuals' behaviour by assuming that their preferences share the same, comparable, property of strength. In other words, it does not postulate any hidden difference in preference strengths when the empirical evidence is identical for both individuals. This is precisely what the principles seen above recommend. If so, it looks like their adoption does indeed lead to a more parsimonious theory, and is, therefore, justified.

However, there is another theory, T_3, that is compatible with the empirical evidence. T_3 registers the fact that the individuals' utilities are numerically identical, but does not infer anything as to whether or not they are comparable. In other words, T_3 does not take any position about IUCs. Call the attitude conveyed by T_3 one of 'neutrality'. Neutrality is a legitimate position, because the assumption of interpersonal comparability is simply not required for the explanation of individual behaviour. Since comparability plays no role, it is possible to remain agnostic about whether people's utilities are co-scaled or not. If what we care about is just parsimony, then the question to ask is the following: is it more parsimonious to have a theory that does not postulate any differences between individuals' utilities or to have a theory that does not postulate anything at all? Strictly speaking, the latter is more parsimonious than the former both with respect to the number and the kinds of properties postulated and, therefore, it should be favoured. I admit that we might have conflicting intuitions here. However, the fact that the issue cannot be easily solved is enough to reject parsimony as a conclusive reason in favour of the adoption of non-empirical principles in order to show that IUCs are scientifically justified.

Finally, let us consider simplicity. It is generally difficult to define what simplicity amounts to. One account reduces simplicity to parsimony. In this case, the previous remarks apply. An alternative account constructs simplicity as elegance, which, in turn, can be defined in terms of the ease with which a theory favours computation or decision-making. Is a theory (T_1) that assumes co-scaled utilities simpler than either a sceptical (T_2) or a neutral (T_3) theory? In order to answer this question, we need to consider the purposes for which such a theory can be used. Suppose we use T_1 to explain individual behaviour. Since the assumption that different people's utilities are on the same scale plays no role in accounting for individual behaviour, it follows that it does not make computation any easier for explanatory purposes. Therefore, insofar as the criterion of simplicity is the ease in the calculation, T1 cannot be deemed simpler than T_2 or T_3.

Suppose now that we use T_1 to take a decision affecting the interests of two or more individuals. In this case, the situation appears to be different. Undoubtedly, a theory that assumes that utilities are co-scaled considerably simplifies decision-making. However, in

the case under consideration, the justification based on simplicity ceases to be purely pragmatic and becomes rather close to a normative justification. We do not count as simpler a theory assuming interpersonal utility comparability merely because it leads to *a* decision; rather simplicity is valuable to the extent that it leads, or it favours reaching, decisions that are considered *fair* or *even-handed.* If this is true, however, it is fairness, or even-handedness, which provides the ultimate justification for the assumption that utilities are on the same scale. Simplicity plays a mere instrumental role. Assuming that people's utilities are co-scaled is justified only insofar as this helps us reach fair or even-handed results. Therefore, the justification is not pragmatic, but normative.

The conclusion is the following. The inference to the best explanation argument for the adoption of the non-empirical principles seen above fails. None of the pragmatic virtues considered offers conclusive grounds for the acceptance of a non-empirical basis for IUCs. Ultimately, this means that this strategy fails to demonstrate that IUCs can be scientifically justified and, *a fortiori*, that we can have scientific knowledge of how preferences compare in terms of strength.

6. Conclusion

The problem of IUCs is the problem of whether or not we can have knowledge of, or scientifically justified, IUCs. In this chapter, I considered solutions that appeal to an inference to the best explanation type of argument. The underlying idea is that IUCs can be scientifically justified if they are determined on the basis of the union of an empirical basis and a non-empirical basis, whose adoption conforms to good scientific practice. In turn, scientific practice is good if it is based on pragmatic virtues such as simplicity, parsimony and explanatory power.

I examined two approaches pursuing this strategy, namely, a third-person approach and a first-person approach. I argued that both approaches fail because the adoption of a non-empirical basis to determine IUCs is not pragmatically advantageous and, thereby, does not conform to good scientific practice. As an instance of the first-person approach, I considered the later Harsanyi's position. I argued that his proposal fails also on other grounds: his extended preference approach is both redundant and conflates the object of preferences with their causes.

CHAPTER 3

The argument from nativism

1. Introduction

One interesting feature of at least some of the more economic-oriented solutions to the problem of IUCs is that they make reference to the explanation of how ordinary people supposedly make ICs of preference strength in everyday life. Since this explanatory problem concerns mental states (i.e. preferences) and one of their properties (i.e. strength) in particular, one would expect the existence of a large literature in philosophy of mind addressing the issue. Instead, and quite surprisingly, philosophers of mind have almost completely ignored this explanatory problem. One significant exception is constituted by Alvin Goldman, who has attempted to bring the problem of IUCs in line with current debates in philosophy of mind and epistemology[1].

Typically, people's everyday practice of comparing mental states is viewed as a two-step process. First, they ascribe mental states to different targets (and, in some cases, to themselves). Second, they compare the targets' mental states with respect to strength. If this picture is correct, the explanation of how ordinary people make ICs of preferences should build on the explanation of how they ascribe preference strengths to other people and to themselves. For this purpose, we need to examine two different kinds of problems: the problem of the meaning of mental states, that is, the problem of what ordinary people mean when they employ mental terms[2]; and the problem of mindreading, that is, the problem of how ordinary people assigns mental states to other people.

There are two main theories of the meaning of mental states in current philosophy of mind: (commonsense) functionalism and experientialism. For introductory purposes, we can briefly describe these accounts in the following way. According to functionalism, the meaning of a mental state is given by the set of causal laws in which such a mental state figures and which relate it to inputs, other mental states and behavioural outputs.

---

[1] See GOLDMAN, A. [1995a].

[2] It is worth emphasising that the problem of the meaning of mental states differs from the problem of the nature of mental states. The former is a semantic problem, concerning what ordinary people mean when they employ mental terms. The latter is a metaphysical problem, concerning what mental states really are. This difference is often missed because there are as many theory of the meaning of mental terms as there are theories of their nature.

Instead, according to experientialism, the meaning of a mental state is given by the more or less conscious experiences that the subject has of it.

On the other hand, there are two alternative explanations of mental ascription: Theory Theory (TT) and Simulation Theory (ST). According to the former, ordinary people ascribe mental states to others by means of a 'theory of mind' that they, more or less tacitly, possess. According to the latter, ordinary people ascribe mental states to others by trying to replicate, or simulate, their mental life. As it has been recently shown in several papers[3], both mindreading accounts can be characterised at two different levels, namely, the sub-personal level of description and the personal level of description. The former level is concerned with the question of what the information-processing mechanisms are that underpin our folk psychological practice of mental ascription[4]. The personal level of description focuses on the way in which persons, as such, think about or interpret other people's mental and overt behaviour[5].

In this chapter, I want to pursue two goals. The first is to show how philosophy of mind can contribute to the debate about IUCs by extending Goldman's analysis. For the present purpose, Goldman's approach has two limitations. It focuses explicitly on ICs of happiness only and it is very specific. In particular, Goldman embraces experientialism as a theory of the meaning of mental states, ST as a theory of mindreading and reliabilism as a theory of justification. In this chapter, I shall extend Goldman's analysis by focusing on ICs of preference strength and by considering also functionalism as a theory of the meaning of mental states, TT as theory of mindreading and evidentialism as a theory of justification. Like Goldman, I shall be concerned only with the sub-personal level of description.

My second goal is to assess whether or not philosophy of mind can help us find a successful solution to the problem of IUCs. I shall devote a special interest to Goldman's own argument from nativism. According to it, the assumption that different people's utilities are co-scaled is justified if the assumption that ICs of preference strength are performed through innate mechanisms that are either hyper-similar across individuals or very closely representative of the workings of other individuals' mind-systems is sound. I shall argue that, when the notion of innate cognitive capacity or mechanism is properly spelt out, this argument reduces to an inference to the best

[3] See HEAL, J. [1994], [1998a], and [2000], DAVIES, M. [2000], DAVIES, M. and T., STONE [2000] and [2001].
[4] See GOLDMAN, A. [1989], [1992], [1995b], [2000], GALLESE, V. and A. GOLDMAN, [1998], STICH, S. and S., NICHOLS, [1992], [1995], [1996], [1997], NICHOLS *et al.* [1996] and NICHOLS, S. and S., STICH [1998] and [2003].
[5] See HEAL, J. [1994], [1998a,b] and [2000], GORDON, R. [1992].

explanation kind of argument. Therefore, this strategy fails to show that ICs of preference strength can be scientifically justified.

I shall proceed as follows. In section 2, I shall give a more detailed illustration of the main theories about the meaning of mental states. In section 3, I shall illustrate the general features of TT and ST. In section 4, I shall illustrate the way in which both accounts may explain how the folks form their beliefs about how different people's preferences compare in terms of strength. In section 5, I shall have a closer look at the conditions that ought to be satisfied for such beliefs to be scientifically justified. In section 6, I shall discuss the argument from nativism and ultimately reject it. I shall summarise my results in section 7.

2. The problem of meaning

Let us start by considering the two main theories of the meaning of mental states in contemporary philosophy of mind. The first is commonsense (or analytic) functionalism[6]. Historically, this account descends from logical behaviourism and dispositionalism. According to the former doctrine, mental states have a purely behaviourist meaning. For instance, preferences mean nothing but choice behaviour. Instead, according to the latter doctrine, mental states are conceived as dispositions towards some behavioural output. For instance, according to a narrow dispositional account, the meaning of preference is that of a disposition towards choice behaviour. On the other hand, according to a broad dispositional account, the meaning of preference is that of a disposition not only towards choice behaviour, but also towards other behavioural expressions.

As philosophers of mind have recognised long time ago, both accounts face some problems. Let us consider dispositionalism as a paradigmatic example. According to this doctrine, when we say that an individual prefers taking the umbrella rather than not taking it, we mean that the individual has a disposition to perform an action of the relevant type in suitable circumstances. However, this is the case only provided that we also assume that the individual's other relevant mental states remain the same. Indeed, when we say that an individual prefers taking the umbrella rather than not taking it, we would not mean that the individual has a disposition to perform the relevant action if we also thought that the individual believes that such an object is a baseball bat.

---

[6] See LEWIS, D. [1972].

The lesson is that the meaning of mental states cannot be defined independently from other mental states. Thus, some philosophers of mind have recommended the adoption of a different, functionalist, account of the meaning of mental states. Such an account takes into consideration the interdependencies between different mental states, while, at the same time, preserving a moderate behaviourist account of their meaning. More precisely, according to functionalism, the meaning of a mental state is given by the set of causal laws in which that mental state figures. Such causal laws specify how each mental state is related to environmental inputs, other mental states and behavioural outputs. The meaning of a mental state is then entirely exhausted by the causal relations in which it figures[7]. It may be the case that the agent employing mental concepts is incapable of specifying all these constitutive causal relations. Indeed, this may require a sophisticated analysis. If we think of the defining causal relations as forming a theory that the agent possesses, then we can say that such a theory operates tacitly, or, equivalently, that the theory is tacit.

There are at least two reasons to be interested in the prospects that functionalism offers for solving the problem of IUCs. First, functionalism has been the dominant view of the meaning of mental states, in philosophy of mind, for the past thirty years. Recently, attempts have been made to characterise preferences as well in functionalist terms[8]. It is natural to ask how this affects the traditional debate about IUCs. Second, since the origin of decision theory, beliefs and degrees of belief have been typically given a functionalist understanding[9]. Thus, conceiving preferences and degrees of preference along the same line is a way of maintaining a consistent understanding of the meaning of mental states.

Preferences can be defined in functionalist terms as mental states that are causally related to certain inputs, and that, in combination with other mental states, produce certain behavioural outputs[10]. What are these causal relations? According to some authors, decision theory is the research area that attempts to specify some of the relevant relations[11]. In particular, decision theory conceives preferences as mental states that lead

[7] Thus, the functionalist theory of the meaning of mental states goes beyond the dispositional theory in one crucial respect: it includes the relationship with environmental inputs and with other mental states as part of the definition of a mental term, in addition with its relationship with behavioural outputs.
[8] See PETTIT, P. [2006].
[9] See RAMSEY, F. P. [1990].
[10] The relation that preferences have with both inputs and outputs may turn out to be indirect, that is, mediated by other mental states that are connected to preferences.
[11] See LEWIS, D. [1986] and PETTIT, P. [1991], reprinted in PETTIT, P. [2002], and [2006].

to choices, in combination with beliefs and desires[12]. If we define preferences in functionalist terms, the property of preferential strength can be seen as a causal property of preferences. In other words, preference strength is the causally efficacious property that leads an individual to behave in a certain way, when subject to specific circumstances and in the presence of other mental states.

Functionalism is not unchallenged[13]. For instance, Goldman discusses three general difficulties that a functionalist account faces[14]. First, functionalism has trouble in specifying the laws in which mental states are supposedly embedded. The fact that even the "experts", e.g. philosophers and social scientists, have poor explicit knowledge of the causal relations that define mental terms seems to cast doubt on whether ordinary people's understanding of mental concepts is governed by knowledge, even if implicit, of functional laws.

Second, functionalism seems to be unable to capture the qualitative features of some of, or perhaps all, our mental states. Consider the 'inverted spectrum' problem. Two individuals may be functionally identical and yet they may have radically different subjective mental experiences. For instance, although functionally identical, they may have colour experiences that lie at the opposite poles of the colour spectrum. If this is a genuine possibility, it appears that functionalism sanctions the use of identical mental terms for mental states that are drastically different, because of its inability to register qualitative differences between mental states.

Third, functionalism does not seem to offer a plausible account of self-ascription of mental states. In order for an individual to classify one of his own mental states as, for instance, a headache, functionalism requires that he be able to identify the causes of such a headache, the relationship with other, both occurrent and non-occurrent, mental states and the behavioural headache expressions. This seems to burden self-ascription with excessive computational requirements. At least phenomenologically, it seems plausible that the individual can identify a mental state of his as a headache without undergoing this complex series of computations.

---

[12] Roughly speaking, there are three possible ways to conceive the relationship between desires and preferences. First, one can be eliminativist about preferences and claim that the notion of preferences is syncategoramatic. It is simply a way to conveniently describe an individual's desires and their relations. However, there are no real mental states corresponding to preferences. Second, one can be reductivist and claim that preferences are real mental states but mental states that reduce to desires in one sense or another, e.g. they constitute a specific, e.g. relational, class of desires. Finally, one can maintain that preferences are derivative on desires, in the sense that they are related to, and determined by, them; but they do not reduce to desires, except in the loose sense that they are both pro-attitudes of some sort. I think that the functionalist position fits more comfortably with the latter position, which I shall thereby adopt in what follows.

[13] The *locus classicus* for a critique of functionalism is BLOCK, N. [1980].

[14] See GOLDMAN, A. [1993] and [1995a].

In opposition to functionalism, Goldman recommends the adoption of an experientialist theory of the meaning of mental states. According to it, the meaning of a mental state is given by the more or less conscious experiences that the subject has of it[15]. An equivalent definition is that the meaning of a mental state is given by the agent's experience of 'what it is like' to have that mental state. Thus, according to this account, mental states are phenomenologically real and the agent has introspective – privileged, although not infallible – access to them. Likewise, the strength of a mental state is a real psychic magnitude, which the subject experiences and can introspectively discriminate.

Preferences can be defined in experientialist terms as mental states that give raise to certain experiences in a subject. It may be the case that there is no unique phenomenal experience that different individuals have in common when they are in a preference-state. However, it is enough that there is a family of experiences that are sufficiently similar to constitute a preference-type. According to an experientialist understanding, then, preference strength is a felt property, a qualitative experience of the individual that has preferences. The subject has introspective access and can discriminate the strengths of his preferences. As such, the meaning of preference strength arises "from points or intervals on the experiential scale"[16] that the term denotes.

There is at least one direct reason to be interested in experientialism for the problem of IUCs, together with the indirect reasons provided by the limits of functionalism. According to Goldman, experientialism offers a better account of what people means when they make ICs of preferences than functionalism and, ultimately, promises a solution to the problem of IUCs in combination with an ST account of mindreading.

3. The problem of mindreading

The explanation of mental ascription at the sub-personal level of description is concerned with the question of what information-processing mechanisms should be posited in order to explain the folks' mindreading capacity, that is, the capacity to ascribe mental states to other people.

TT characteristically accounts for this cognitive capacity by positing cognitive processes that exploit "an internally represented "knowledge structure" - typically a body of rules or principles or propositions - which serves to guide the execution of the

[15] See particularly the account offered by GOLDMAN, A. [1995a].
[16] GOLDMAN, A. [1995a], p. 713.

capacity to be explained"[17]. In short, TT explains mental ascription by arguing that the folks possess a 'theory of mind' (ToM), to which they have a more or less conscious access[18]. Nichols and Stich represent boxologically the basic architecture of each agent's mind-system, under the TT hypothesis, in the following way[19].

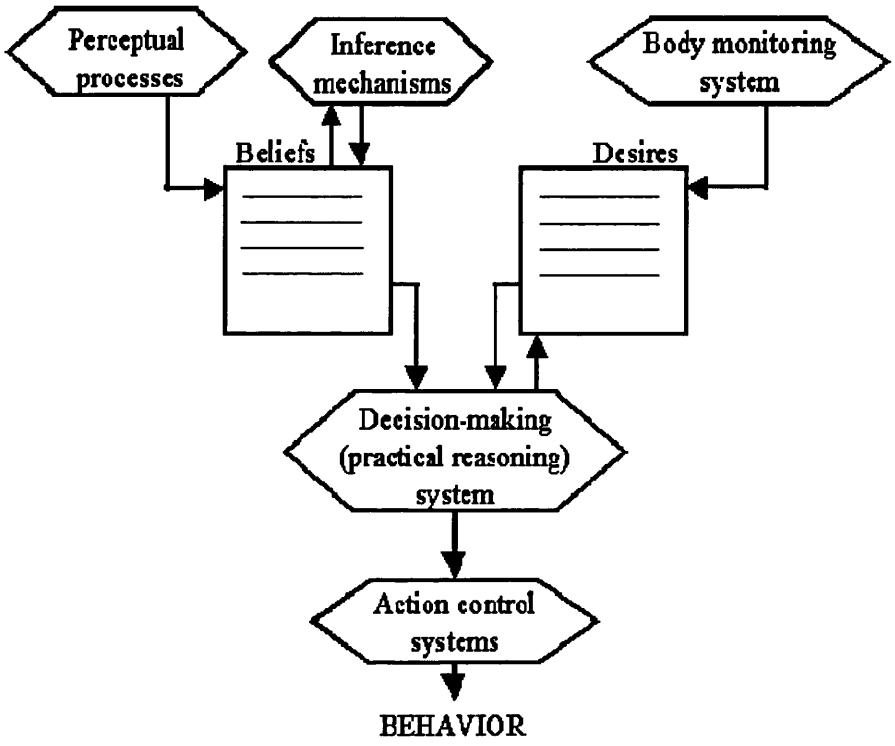

Figure 1

As far as the meaning of mental states is concerned, TT is generally associated with analytic functionalism, according to which the meaning of a mental state is given by the

[17] See STICH, S. and S., NICHOLS [1992], pp. 35-36. See also, STICH, S. and S., NICHOLS [1995] and [1997], NICHOLS *et al.* [1996] and NICHOLS, S. and S., STICH [2003].

[18] There are two variants of the TT approach to mindreading, namely, the scientific-theory theory (STT) and the modularity theory (MT). According to the former, the ToM that the folks use for mindreading is both learnt and stored in the mind in the same way as scientific theories are. In the course of their development, children proceeds as little scientists, formulating hypotheses on the basis of the information available and revising them in the light of new data. In other words, the ToM that the folks possess is included in the belief box. According to the latter, the ToM is neither learnt nor stored in the same way as scientific theories are, but it is rather included in one or more innate modules. As such, the ToM that the folks possess is connected, but distinct, from the belief box. For the purpose of this thesis, however, we can ignore the distinction between the two approaches. See WELLMAN, H. [1990], PERNER, J. [1991], GOPNIK, J. and H., WELLMAN [1992], [1994], GOPNIK, J. and A. N., MELTZOFF [1997] for a defence of the STT approach. See LESLIE, A. [1987], [1988], [1994], [2000], LESLIE, A. and T., GERMAN [1995] and BARON-COHEN, S. [1995] for a defence of the MT approach.

[19] This figure is borrowed from STICH, S. and S., NICHOLS [1992].

set of causal laws in which such a mental state figures[20]. As far as mental ascription is concerned, TT assumes that the folks ascribe mental states to other people by observing external events (i.e. inputs and outputs) and inferring the relevant mental states by reference to the causal relations postulated by the ToM that they possess.

ST offers an alternative account of mental ascription. First of all, let us distinguish two different kinds of simulation, namely, simulation in reality and simulation in imagination. In the former case, simulation involves replicating the behaviour of an object in specific circumstances by using an object of the same kind in similar circumstances. By contrast, in the latter case, simulation takes place in imagination and involves replicating the behaviour of an object in specific circumstances by imagining how the object would behave in similar circumstances. Mental simulation is an instance of simulation in imagination, since it involves replicating another individual's mental life in specific circumstances by imagining being subject to the same or relevantly similar circumstances.

The basic idea is that the folks ascribe mental states to other people by taking their own information-processing mechanisms 'off-line' and feeding them with pretend inputs, which correspond to the other people's initial mental states. The relevant mechanisms run 'off-line' and produce pretend outputs, which correspond to the other people's targeted mental states. For instance, when the goal is to predict another individual's behaviour, the relevant information-processing mechanism is the practical reasoning system and the pretend inputs are pretend beliefs and desires. Under the ST hypothesis, each agent's mental system can be represented boxologically in the following way[21].

[20] However, functionalism is compatible with two different accounts of third-person mental ascriptions. On the one hand, it is compatible with the theory-theory approach under discussion. On the other hand, it is compatible with a theory-driven simulation approach. In both approaches, the interpreter employs mental concepts that are defined with respect to the role that they have in the underlying theory.

[21] This figure is borrowed from STICH, S. and S., NICHOLS [1997].

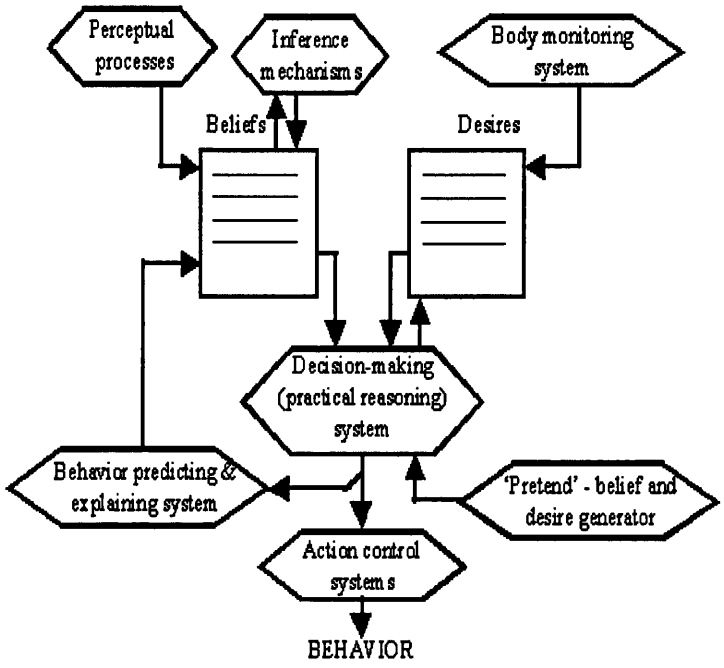

Figure 2

The main proponent of a simulationist approach at the sub-personal level of description is Alvin Goldman[22]. According to Goldman, the folks simulate other people's mental life by using themselves as "analogue models". The simulator, first, asks himself what mental states he would have if he were subject to the initial mental states of the simulated agent; then, he introspects his own mental states and ascribes them – by analogy – to the simulated agent[23].

We can characterize Goldman's account with respect to three dimensions:

(1) the level of information;

(2) the direction of gaze;

(3) the epistemological status.

[22] See, in particular, GOLDMAN, A. [1989], [1992], [2002], [2006].
[23] See GOLDMAN, A. [1989].

With respect to the first dimension, Goldman's account conceives mental simulation as characteristically "process-driven". This means that mental simulation does not rely upon any body of knowledge (or, at most, it relies upon a very minimal body of knowledge), but simply uses the same processes of the targeted object[24]. This view contrasts with the view of simulation as "theory-driven", that is, as relying upon a body of knowledge, which may be either explicit or tacit and more or less rich.

With respect to the second dimension, Goldman's view is, as Davies and Stone put it, that "the gaze of the simulator [is] first inward and then outward to the person being simulated and it [is] the inward gaze that distinguish[es] simulation from the use of a third-personal empirical theory about psychological processes"[25]. This view contrasts with the one advanced by authors such as Gordon and Heal, according to whom, as Davies and Stone put it, "the simulator's gaze is neither inward nor upon the person being simulated but, primarily, upon the (imagined) circumstances about which the person being simulated is thinking"[26].

Finally, with respect to the third dimension, Goldman presents ST as an empirical hypothesis within cognitive science. As such, it draws upon empirical research and can be either confirmed or disconfirmed by it. This view contrasts with aprioristic accounts, such as the one offered by Heal, according to which ST is an *a priori* hypothesis, which cannot be disconfirmed by empirical research.

Goldman's account presupposes that the simulator is capable of introspecting his own mental states in order to ascribe mental states to another individual by analogy. One possibility is that the simulator recognises his own mental states with respect to the function that they occupy in his mind-system. However, this possibility is precluded, since it would precipitate mental ascription into a "theory-driven" simulation and collapse Goldman's ST account into a TT account. As far as the meaning of mental states is concerned, Goldman adopts an experientialist view[27]. According to Goldman, "experientialism is "the traditional view that mental language gets its meaning, primarily and in the first instance, from episodes of conscious experience of which the

---

[24] See GOLDMAN, A. [1989].

[25] See DAVIES, M. and T., STONE [2000], p. 2 at http://philrsss.anu.edu.au/~mdavies/papers/simrep.pdf.

[26] See DAVIES, M. and T., STONE [2000], p. 2 at http://philrsss.anu.edu.au/~mdavies/papers/simrep.pdf.

[27] See GOLDMAN, A. [1993]. It is worth noticing that Goldman has recently changed his mind about the meaning of mental states. At present, he defends the view that mental concepts pick out categorical properties of mental states that are nonetheless not phenomenal properties. See GOLDMAN, A. [2002] and [2006]. Here I shall still focus on the version of ST associated with experientialism for two reasons. The first is that this is the best developed version of ST. The second is that Goldman's most recent account of the meaning of mental states is not yet elaborated in sufficient details to provide a basis for an accurate discussion of the problem of ICs of preference strength.

agent is more or less directly aware"[28]. This means that the simulator recognises his own mental states on the basis of their phenomenology, that is, on the basis of 'what it is like' to have them in specific circumstances.

4. Mindreading and the problem of IUCs

Let us now examine more closely the relationship between the problem of mindreading and the problem of IUCs. The first question to ask is the following. In what way can TT and ST explain, at the sub-personal level of description, how the folks form their beliefs about how different people's preferences compare in terms of strength?

Let us consider TT first. Suppose for simplicity that the judge is one of the two individuals whose preferences are to be compared in terms of strength. As seen above, TT is typically associated with a functionalist understanding of mental states. If we adopt such a view of preferences, TT may account for the judge's beliefs in the following way. The first step concerns third-person mental ascription. The judge observes the relevant external events (i.e. instances of the input-types and output-types that are included in the definition of preferences) and infers both the other relevant mental states (i.e. tokens of the mental state-types that are included in the definition of preferences) and the relevant preferences, by reference to the causal relations postulated by the ToM that he – more or less tacitly – possesses.

The second step concerns first-person mental ascription. Orthodox TT suggests that first-person mental ascription entirely parallels third-person mental ascription. This means that self-ascription is based on inferences mediated by the ToM that the subject possesses. Less orthodox TT approaches relax this position by conjecturing that first-person mental ascription may involve the use of recognitional devices or mechanisms – which either make the use of the ToM invisible, but not completely irrelevant, or confine it to certain specific purposes – and ends up with the self-ascription of both the relevant mental states and the relevant preferences. However the core idea of TT remains that the judge ascribes preferences with a specific content and strength both to himself and to the other individual on the basis of the ToM in his possession. Finally, in the last stage, he compares the intensity of his preferences with the intensity of the other individuals' preferences.

[28] See GOLDMAN, A. [1995a], p. 712.

Let us now consider ST. As seen above, Goldman's approach is associated with an experientialist understanding of mental states. If we adopt such an experientialist view, it is immediately clear that the comparison of two individuals' preference strengths presents a serious problem. Each individual has introspective access to, and ability to discriminate, the intensity of his own preferences only. However, neither individual can introspect the other individual's preference strengths. This seems to threaten the very possibility of making ICs of preference strength. The matter is even more complicated if the comparison is performed by a judge, since a third party cannot access either individual's preference strengths.

The problem generalises. If the observer must be able to introspect another individual's mind in order to ascribe mental states to him, then, since the observer has introspective access to his own mental states only, it follows that third-person ascription of mental states is impossible. The upshot is not only that ICs are impossible, and, *a fortiori*, cannot be scientifically justified; it is also a full-blown scepticism about the existence of other people's minds. Nevertheless, this formulation of the problem suggests a possible way out. Since we do make both third-person mental ascriptions and ICs of the intensity of other individuals' mental states, perhaps if we explain how we attribute mental states to other people, we may also explain how we make ICs of preference strength.

Goldman follows this strategy. According to him, the problem of ICs of preferences is a particular case of the more general problem of third-person mindreading. In general, ST holds that the folks ascribe mental states to other people by running their practical reasoning system 'off-line', after feeding it with pretend mental states corresponding to the other agent's initial mental states. Once again, suppose for simplicity that the comparer is one of the two individuals whose preferences are to be compared in terms of strength. Goldman's account explains how the simulator makes ICs of preference strength in the following way. The first step concerns third-person mental ascription. The simulator introspects what preferences he would have if he were to have the simulated agent's initial mental states. By so doing, he recreates in imagination the same qualitative experiences of the individual whose preference strengths he wants to compare and discriminates them through introspection. Then, the simulator classifies these experiences as experiences of preferences with a specific intensity and ascribes such preference strengths to the other agent by analogy. The second step concerns first-person mental ascription. The simulator introspects his own preferences, discriminates their intensities and ascribes the detected preference strengths to himself. Finally, in the

last step, the simulator compares the intensity of his own preferences with the intensity of the simulated agent's preferences.

It is worth noticing that both TT and ST set only minimal conditions for the *possibility* of forming beliefs about how different people's preferences compare in terms of strength. In both cases, the explanation of how ICs of preference strength are made is consistent with the possibility that different observers massively disagree about them. For instance, if different individuals possess different 'theories of mind', it is likely that they will form different ICs. By the same token, if different individuals' practical reasoning systems work in a different way, it is likely that they will form different ICs. In the worst case scenario, different individuals might form different beliefs about how different people's preferences compare in terms of strength even on the basis of the same external events or the same 'pretend' inputs. This is a problem for both justification of a garden-variety sort and scientific justification. If one adopts an evidentialist perspective, the same evidence may equally support incompatible ICs of preference strength. If one adopts a reliabilist perspective, the same type of cognitive process may lead to different ICs of preference strength. However, it is hard to see how two cognitive processes can both be reliable and deliver incompatible conclusions from the same initial data.

Two questions arise. First, what conditions should TT and ST satisfy in order to lead to scientifically justified ICs of preference strength? Second, what reason do we have to assume that those conditions can be satisfied? I shall try to answer these questions in the next sections.

5. The conditions for scientific justification

5.1 Simulation Theory

Goldman has done a lot of work to demonstrate that ICs of happiness can be scientifically justified. His analysis provides an excellent starting point for our discussion about ICs of preference strength. Therefore, in this section I shall reverse the order of exposition and start from ST.

According to the evidentialist version of scientific justification, a belief about how different people's preferences compare in terms of strength is scientifically justified when it is inferred from evidence that is (a) public, (b) replicable; (c) such as to lead to accurate and precise measurements of the relevant variables. According to the reliabilist

version, a belief about how different people's preferences compare in terms of strength is scientifically justified when it is acquired through methods and techniques that are (a) reliable and *known* to be reliable on the basis of scientific evidence; (b) replicable; and (c) such as to lead to accurate and precise measurements of the relevant variables. Moreover, if belief formation involves the use of non-empirical principles, these principles must be justified by means of considerations that are acceptable for the scientific community.

Since both versions share the replicability and the measurement conditions, I shall start from these conditions. First, let us consider replicability. Goldman suggests that the fact that "well-informed, skilled deployers of the simulation heuristic"[29] often reach intersubjective agreement about ICs is a proof that simulation is replicable. However, intersubjective agreement cannot, by itself, be the mark of replicability. Rather, the extent to which simulation is replicable depends on whether or not it is based on information-processing mechanisms that different simulators similarly possess. Perhaps, the fact that different simulators often reach intersubjective agreement in mental ascriptions offers a reason to think that they share relevantly similar information-processing mechanisms. Yet, this is a substantive (and crucial) issue. A specific position about it cannot be assumed without arguing for it. I shall consider the issue in more detail below.

Second, let us consider measurement. So far we have examined how ST may explain ICs of preference strength. However, ST seems to lead to purely ordinal comparisons and possibly imprecise or vague ones. By contrast, at least some kinds of IUCs require a cardinal representation of preferences and a great deal of accuracy and precision. Can ST satisfy such a condition? I think so. The starting point is the collection of information about the mental states that the individuals to be compared would have in hypothetical situations. Then, the simulator plugs such 'pretend' mental states into his practical reasoning system and, by means of repeated simulations, he derives an accurate set of preferences that he finally ascribes to the other individual. If these preferences satisfy the axioms of expected utility theory, they can be represented by a utility function, unique up to a positive affine transformation. In a nutshell, the reduction of inter-personal comparisons to intra-personal comparisons grounds as much precision and accuracy in the measurement of preference strength as in the individual case.

[29] GOLDMAN, A. [1995a], p. 722.

Third, let us consider the other conditions for scientific justification. In the evidentialist framework, ICs of preference strength must be supported by public evidence. In Goldman's account, the evidence is provided by the simulator's beliefs about his own preference strengths and the simulated agent's ones, which he can both access through introspection. Thus, introspection constitutes the main source of evidence. Does it count as publicly acquired evidence? Some authors are convinced that it does not. For instance, Robbins' attack against the scientific legitimacy of IUCs stems precisely from a rejection of introspection as admissible source of evidence[30]. Although it is difficult to establish exactly what the requirements of publicity are, introspection seems to lack a public character at least in an intuitive sense. This is a problem for the scientific justification of simulation-based ICs of preference strength within an evidentialist framework.

The issue is different within the alternative reliabilist framework. In the reliabilist framework, beliefs about how different people's preferences compare in terms of strength must be acquired through reliable processes. Summarising Goldman's own position and a large literature on ST, we can distinguish three requirements that are thought to be jointly sufficient for the reliability of simulation, in general. First, simulation must be based on the 'correct' inputs. Second, the simulator's relevant information-processing mechanisms must operate in the same way in imagination as in reality. Third, the simulator and the simulated agent must be similar at the level of the relevant information-processing mechanisms. I shall discuss these requirements in the next sub-section. For the present purpose, one crucial difference between evidentialism and reliabilism is that, in the former case, what needs to be public is the scientific evidence supporting ICs of preference strength, whereas, in the latter case, what needs to be public is the scientific evidence about the reliability of the processes determining ICs of preference strength.

5.2 ST and the reliability requirements

Consider the first reliability requirement. If the simulator feeds his information-processing mechanisms with incorrect inputs, then he is likely to reach wrong conclusions about the simulated agent's mental states. The same is true if we move from the general case to the case of ICs of preference strength. In Goldman's account, ICs of preference strength are formed on the basis of the inputs that are fed into the simulator's

[30] See ROBBINS, L. [1932], specially pp. 139-142.

practical reasoning system. These inputs are the simulator's pretend mental states, which, on the one hand, the simulator has supposedly unproblematic access to and, on the other hand, supposedly correspond to the simulated agent's actual mental states.

However, the orthodox ST account poses two problems. The first is that it is not entirely clear which pretend mental states the simulator should feed into his practical reasoning system. Under a functionalist understanding, preferences are defined in connection with other mental states. At the input level, in particular, preferences are causally connected to beliefs and desires. However, under an experientialist understanding, it is not immediately clear why the simulator should consider pretend desires and pretend beliefs. More generally, it is not clear how to individuate the types and the contents of the pretend attitudes that the simulator must feed into his 'off-line' practical reasoning system in order to derive pretend preferences. Many authors sympathetic to a simulationist approach recognise the need for a minimal body of knowledge to fill this gap. Although this is a move towards a more hybrid account, it may not be problematic for ST, in its general form. Indeed, the core idea that mental ascription is performed by replicating another individual's mental life would remain intact[31].

The second problem is that, even if the simulator can correctly individuate attitude-types and attitude-contents, he must still be able to correctly individuate also the intensity of the pretend mental states that should be used as inputs for the simulation. It is plausible to maintain that the interpersonal comparison of the simulator's pretend mental states and the simulated agent's actual mental states raises the same difficulties associated with the interpersonal comparison of two different individuals' actual mental states. Thus, the assumption that the simulator can feed the 'correct' inputs into his 'off-line' system simply begs the question.

In the light of both problems, it seems better to take the simulated agent's environmental circumstances, rather than pretend mental states, as inputs of simulation. This suggestion brings ST closer to Harsanyi's causal approach. In fact, the simulator does not begin by asking himself what preferences he would have if he were to have another individual's initial mental states. Instead, he begins by asking himself what preferences he would have if he were in the other individual's initial circumstances. This move shows that introspection is not the only source of evidence. The observation and collection of data concerning the simulated agent's environmental circumstances and personal history is also necessary. ST must be complemented with causal

[31] This is the approach pursued by Goldman himself in GOLDMAN, A. [2006].

knowledge about the relations between environment and mental states such as beliefs and desires. Moreover, it must be complemented with knowledge about the history of the simulated agent, which should be used to identify which environmental circumstances constitute relevant inputs in specific situations, amongst the infinite ones that the mere observation of the simulated agent's situation allows one to consider. Once again, this moves ST towards a more hybrid formulation.

Let us consider now the second requirement. We can alternatively characterise it by saying that the 'off-line' simulation must approximate the 'on-line' working of the relevant information-processing mechanisms[32]. Consider the distinction between simulation in reality and simulation in imagination. In the former case, simulation is a reliable guide to the behaviour of the targeted object, because, if the objects are of the same kind, "the same processes occur in the simulation as would be operative in generating the behaviour of the object being simulated"[33]. Thus, simulation in reality is 'process-driven'. At first sight, instead, simply imagining the behaviour of an object in specific circumstances does not warrant any conclusions concerning its actual behaviour[34]. In order for simulation in imagination to be reliable, simulation must be based on a body of knowledge about the simulated object, which guarantees that imagination correctly mimics the actual processes generating the object's behaviour. Thus, typically simulation in imagination must be 'theory-driven'. The problem is that, if mental simulation is 'theory-driven', then it collapses into the alternative explanation of third-person mental ascription that ST is supposed to challenge, namely, TT. Indeed, according to TT, the ability to ascribe mental states to other people stems precisely from our possession of a body of information that guides our folk psychological practice. One way to avoid the collapse is to claim that mental simulation can be 'process-driven', provided that "at least some mental processes operate in just the same way when we imagine being in a particular situation as they would if we were really in that situation"[35].

What are the relevant information-processing mechanisms that must operate in the same way in imagination as in reality in the case of ICs? Goldman's account focuses on the practical reasoning system only. However, if simulation starts from environmental

---

[32] Goldman's claim is that "psychological systems must operate on feigned pretend input states in the same way they operate on genuine states, at least to a close enough approximation". See GOLDMAN, A. [1995a], p. 722.

[33] See DAVIES, M. and T., STONE [2000], p. 1 at http://philrsss.anu.edu.au/~mdavies/papers/simrep.pdf.

[34] This is the objection raised by MacKay against Harsanyi's extended preference approach, which we discussed in chapter 2. See MACKAY, A. F. [1986], pp. 316-322.

[35] See DAVIES, M. and T., STONE [2000], p. 2 at http://philrsss.anu.edu.au/~mdavies/papers/simrep.pdf.

circumstances, rather than pretend mental states, it is necessary to impose the requirement on other information-processing mechanisms as well. In particular, we must expect that the simulator's response to pretend environmental inputs approximates his response to actual environmental inputs. If we consider Nichols' and Stich's boxological representation and if the relevant pretend mental states are pretend beliefs and pretend desires, the requirement is that the 'off-line' working of the perceptual system, the inference system and the body monitoring system approximates their 'on-line' working.

Let us consider now the third requirement, according to which the simulator and the simulated agent must be similar at the level of the relevant information-processing mechanisms. Let us refer to it as the assumption of interpersonal psychological similarity. In order for the ascription of preference strength to be reliable on the basis of evidence about behavioural outputs, the assumption of interpersonal similarity must hold not only for the perceptual system, the inference system and the body monitoring system, but also for the action control system. In fact, a behavioural output is caused by a corresponding preference relation only if the latter produces the former in the appropriate way. The crucial question is how similar the simulator and the simulated agent must be in order for simulation to produce reliable ICs of preference strength. It is worth noticing at the outset that this is the dimension that distinguishes the problem of ICs from the more general problem of mental ascription. In fact, the similarity requirement in the case of a belief about ICs of preference strength is more stringent than the corresponding requirement in the case of a belief about another individual's mental states. Let me explain why.

Consider a simulator *i* and a simulated agent *j*. Suppose *i* uses the 'correct' inputs for simulation and the 'off-line' working of the relevant information-processing mechanisms approximates their 'on-line' working. That is, suppose that the first two requirements for the reliability of simulation are satisfied. Typically, the simulator *i* ascribes preferences to the simulated agent *j* with content and strength that best predict or explain *j*'s behaviour. For this purpose, however, individual *i*'s experiential scale must be similar to individual *j*'s experiential scale only up to the point of capturing the relevant facts about *j*'s behaviour. Yet, there is a variety of mental ascriptions that are consistent with such facts. In particular, there is a variety of mental ascriptions that, despite offering acceptable predictions or explanations of *j*'s behaviour, sanction alternative and incompatible ICs of preference strength. The upshot is the following. In order for simulation to be reliable for ICs, the psychological similarity between

simulator and simulated agent must be particularly high. In the most favourable case, the simulator's and simulated agent's relevant information-processing mechanisms are perfectly identical. In a less than favourable, but yet acceptable, case, the simulator and the simulated agent are psychologically similar up to the point where the simulator's ascription of preferences not only leads to adequate predictions or explanations of the simulated agent's behaviour, but also to correct ICs of preference strength.

5.3 Theory Theory

Let us consider TT now. Once again, since both the evidentialist and the reliabilist version of scientific justification share the replicability and the measurement conditions, I shall start from these conditions. First, let us consider replicability. According to TT, mental ascription is based on the deployment of a ToM. Even if different individuals use different 'theories of mind', mental ascription may be replicable provided that such bodies of knowledge can be transferred from one individual to another. If this is the case, preference ascription is clearly replicable and, thereby, forming ICs on the basis of the previous ascription is replicable too.

Second, let us consider measurement. The starting point is the collection of information about the individuals to be compared, in hypothetical situations. In a TT framework, the judge uses the information in combination with the ToM in his possession to infer the preferences that the compared individuals might have in hypothetical situations. If these preferences satisfy the axioms of expected utility theory, they can be represented through a utility function, up to a positive affine transformation.

Third, let us consider the other condition for scientific justification. In the evidentialist framework, ICs of preference strength must be supported by public evidence. If we adopt a functionalist understanding of preferences, the relevant evidence is represented by the observation of external events. More precisely, the evidence is represented by both the input-types and the output-types that enter the functionalist definition of preferences. This evidence is public and clearly counts as scientific. Indeed, one of the virtues of a functionalist approach to mental states is that it offers a naturalistic and, thereby, 'scientific' account of the meaning of mental states.

In the reliabilist framework, beliefs about how different people's preferences compare in terms of strength must be acquired through processes that are both reliable and known to be reliable. The requirements for reliability in the case of TT parallel, although do not entirely coincide with, the requirements seen in the case of ST. First,

the ToM used by the subject making ICs must be applied on the basis of 'correct' evidence. Second, the ToM must closely represent the working of the relevant information-processing mechanisms of the targeted agent.

5.4 TT and the reliability requirements

Consider the first requirement. If the evidence that the observer considers while applying the theory in his possession is not correct, then he is likely to reach wrong conclusions about the observed agent's mental states. We have seen above that the relevant evidence is constituted by the elements that enter the functionalist definition of preferences. Choice behaviour is a natural candidate. Choice behaviour provides evidence for the ascription of individual preferences, because the ascription of preferences, with suitable content and strength, explains an individual's choices. However, choice observation may not be the only relevant evidence. After all, choices may be only one of the behavioural outputs of preferences. For the sake of the argument, we can be very liberal in deciding what counts as relevant behavioural outputs for preference ascription and consider latency of choice (e.g. Waldner), verbal expressions (e.g. Harsanyi), expressive reactions (e.g. Weintraub), facial expressions, body temperature and other proxies (e.g. List)[36]. However, if we adopt a functionalist understanding of preferences, behavioural outputs do not exhaust the elements included in the set of relevant evidence. Indeed, preferences are determined with respect to both outputs and inputs. Therefore, information about the agent's history and surrounding environment too should be included in the set of relevant evidence for the ascription of preferences.

Let us now consider the second requirement, according to which the ToM used by the observer for making ICs must closely represent the working of the relevant information-processing mechanisms of the targeted agent. Let us refer to it as the assumption of ToM-to-mind similarity. Two questions arise in the TT case as in the ST case. First, what are the relevant information-processing mechanisms? Second, how closely must the theory represent their workings? The answers are parallel. With respect to the first question, the ToM must closely represent both the agent's response to the environmental inputs and the interaction between the agent's different mental states. If we consider Nichols' and Stich's boxological characterisation, this means considering the following requirements. First, if the mental states that are functionally connected to

[36] See HARSANYI, J. [1955] and [1977], WALDNER, I. [1972], WEINTRAUB, R. [1998], LIST, C. [2003].

preferences at the input-level are beliefs and desires, the theory must closely represent the working of the perceptual system, the inference system and the body monitoring system. Second, the theory must closely represent the interaction between beliefs and desires, on the one hand, and preferences, on the other; that is, it must closely represent the agent's practical reasoning system. Finally, in order for the ascription of preference strength to be reliable on the basis of evidence about behavioural outputs, the theory must closely represent the working of the action control system.

With respect to the second question, once again, we must posit a more stringent similarity requirement than in the case of mental ascription. The argument is analogous to the one seen in the case of ST. The main idea is that there is a variety of mental ascriptions that offer both empirically adequate predictions and explanations of different individuals' behaviour and incompatible ICs of their preference strengths. The upshot is the following. In order for the observer's ToM to be a reliable instrument for forming a belief about ICs of preference strength, the similarity between the theory representation and the observed agent's relevant information-processing mechanisms must be particularly high. In the most favourable case, the observer's ToM perfectly represents the observed agent's information-processing mechanisms. In a less than favourable, but yet acceptable, case, the observer's ToM represents the observed agent's information-processing mechanisms up to the point where the ascription of preferences not only leads to adequate predictions or explanations of the agent's behaviour, but also to correct ICs of preference strength.

5.5 ST and TT compared

As we have seen, there is an asymmetry in the reliability requirements in the case of ST and in the case of TT. Simulation reliability depends on the satisfaction of three requirements, whereas the ToM reliability depends on the satisfaction of two requirements only. More precisely, unlike ST, TT does not presuppose interpersonal psychological similarity between all individuals. Let us see why. Consider two mindreaders *k* and *h* and two agents *i* and *j*. In accordance with the requirements seen above, *k* and *h* form reliable beliefs about how *i*'s and *j*'s preferences compare in terms of strength if they consider the correct inputs and if the 'theories of mind' that they apply for mental ascription closely represent *i*'s and *j*'s relevant information-processing systems. The satisfaction of the latter requirement is possible provided that both *k* and *h* possess the same (or very similar) ToM about *each* of the agents under consideration.

However, this does not imply that *k* and *h* must possess the same (or very similar) ToM about *all* the other agents. The assumption that mental ascription requires the subject to apply the same (or very similar) ToM to all individuals is not needed in order for a belief about ICs of preference strength to be reliably acquired. Indeed, mental ascription may be reliably performed even if the targeted agents are different, provided that the subject uses the correct ToM for each of them. Therefore, TT does not need the assumption of psychological similarity across all individuals.

At worst, the previous point presents a scenario where there are as many 'theories of mind' as there are agents to mindread. This may suggest that each agent is treated as different, in kind, from all the other agents[37]. Although this should be conceptually granted, there are reasons to think that a less than radical TT account may be more plausible and more symmetrically in line with a ST account. Quite independently of how each subject acquires the body of knowledge on which his mindreading capacity is based, it may be plausible to assume that, *if* the second reliability requirements is met, that is, *if* the subject's ToM very closely represents the working of the relevant information-processing mechanisms of another individual, then the subject uses the same theory for ascribing mental states to all other individuals. The assumption of interpersonal psychological similarity would follow thereby. The reason why this assumption is reasonable is that the possibility that a subject acquires such highly specific knowledge representing the working of another individual's information-processing mechanisms is more plausible if it applies to all other individuals. In fact, the cost of acquiring different, but deeply specific, nomological knowledge about each targeted agent would be intolerably high to constitute a real possibility. Therefore, the assumption of interpersonal psychological similarity may still follow *if* the assumption of ToM-to-mind similarity holds.

6. Discussion

6.1 Preliminaries

[37] Incidentally, this possibility motivates some of the criticisms raised against the TT approach. Accordingly, as Heal puts it, TT asks us "to view other people as we view stars, clouds or geological formations. People are just complex objects in our environment whose behaviour we wish to anticipate but whose causal innards we cannot perceive. We therefore proceed by observing the intricacies of their external behaviour and formulating some hypotheses about how the insides are structured". See HEAL, J. [1986], p. 135. The worst case scenario is even more radical, since each person is represented as a different kind of object. According to Heal, this is an unacceptable consequence of the TT approach.

The previous section raises some preliminary issues. The first concerns the meaning of preferences. Shall we opt for an experientialist or a functionalist account of the meaning of mental states? The second concerns the choice between alternative accounts of mental ascription. Shall we opt for a ST or a TT account of the folks' mindreading capacity, at the sub-personal level of description? Within the former field, does Goldman's version represent the best simulationist account or is there a better alternative? I shall set these issues aside. Instead, I shall focus on the question of whether or nor these accounts show, in their own terms, that the folks' beliefs about how different people's preferences compare in terms of strength can be scientifically justified.

In the reliabilist framework, if the conditions of reliability, replicability and measurement are satisfied, then ICs of preference strength can be scientifically justified. In the context of ST, even if ICs are underdetermined by empirical evidence concerning environmental inputs and behavioural outputs, they can indeed be determined if they are formed on the basis of both empirical evidence and cognitive processes that are highly similar across individuals. In the context of TT, even if ICs are underdetermined by empirical evidence concerning environmental inputs and behavioural outputs, they can indeed be determined if they are formed on the basis of both empirical evidence and a ToM that represents the interpreted agent's information-processing mechanisms with a high degree of similarity.

At first sight, the situation is complicated within an evidentialist framework, even in the case where the conditions of publicity, replicability and measurement are satisfied. After all, the evidence concerning environmental inputs and behavioural outputs is not sufficient to determine ICs and, clearly, the evidential situation is unaltered both in the ST context and in the TT context. However, the asymmetry is just apparent. If it is possible to show that ICs are formed on the basis of, respectively, interpersonally similar cognitive processes or representationally adequate 'theories of mind', then ICs can indeed be determined and, thereby, scientifically justified.

The result is the following. Both within the reliabilist framework and within the evidentialist one, the crucial question is whether or not there is scientific evidence vindicating either the assumption of interpersonal psychological similarity, in the case of ST, or the assumption of ToM-to-mind similarity, in the case of TT.

Before proceeding, it is worth noticing one thing. One of the attacks raised by Goldman against functionalism was that functionalism does not have the resources to capture the meaning, and to offer an explanation, of ICs of preference strength. If the

previous analysis is correct, the opposite is true. Functionalism, in its TT version, has the same resources as, or at least similar resources to, experientialism.

6.2 A nativist solution?

Goldman considers four arguments in support of the assumption of interpersonal psychological similarity. I shall extend these arguments to the assumption of ToM-to-mind similarity. I shall call them, respectively, the argument from mindreading predictive success, the argument from evolution, the argument from scientific practice and the argument from the analogy with linguistics. The former two are based on empirical considerations only, whereas the latter two include also non-empirical considerations.

The argument from mindreading predictive success claims that the fact that mindreading is reliable for predictive purposes provides *prima facie* evidence that mindreading is reliable also for the purpose of making IC judgments. The reason is that IC judgments are based on the same mental ascriptions that lead to reliable predictions. As illustrated in section 5, the objection against this argument is that success at predicting an agent's behaviour requires a looser degree of similarity than the one required in order to have scientifically justified ICs. For instance, in the case of ST, although it may be true that "empirically observed success at empathy-based predictions of behaviour does go some distance toward supporting psychological isomorphism"[38], it is not true that predictive success goes far enough in showing that such a psychological isomorphism leads to correct ICs of preference strength. The reason is that reliable predictions are consistent with different and incompatible IC judgments. At best, predictive success shows that simulation is reliable for predictive purposes. However, it does not offer a reason to think that simulation is reliable also for making ICs of preference strength. The same is true, *mutatis mutandis*, in the case of TT.

The argument from evolution claims that evolutionary pressure might have favoured the development of a close match between the simulator's and the simulated agent's information-processing mechanisms, in the case of ST, or between the observer's theory and the target's information-processing mechanisms, in the case of TT. The reason is that this would have maximised the expected fitness of the members of a relevant group by endowing them with competitively advantageous features for the typical environment encountered by the group. The crucial variable in the argument is expected

[38] GOLDMAN, A. [1995a], p. 724.

fitness. What contributes to fitness? When would fitness be maximised? If each individual's fitness is assessed with respect to the capacity to predict or to explain another group member's behaviour, then fitness would be maximised – or, perhaps, optimised – if all members were endowed with a looser degree of similarity than the one necessary in the case of ICs. In fact, the expected fitness would be equally maximised if the match were sufficiently high to guarantee the agreement between the members of the group. As we shall see in more detail below, intersubjective agreement presupposes a less demanding degree of similarity than the one required for having reliable IUC judgments. As a consequence, a higher degree of similarity would be unnecessarily costly. In the light of evolution's traditional 'economy', the upshot is that this argument does not support either the assumption of interpersonal psychological similarity or the assumption of ToM-to-mind similarity.

The failure of arguments based only on empirical considerations is not surprising. In the wake of the analysis of chapter 2, it should be clear that the assumptions of interpersonal psychological similarity and ToM-to-mind similarity are non-empirical. This means that ICs of preference strength can be scientifically justified provided that the use of non-empirical considerations can be vindicated in a way that is acceptable for the scientific community. Goldman explores two arguments based on non-empirical considerations. The first is the argument from scientific practice. According to it, the hypothesis of a high match between the simulator's and the simulated agent's information-processing mechanisms, in the case of ST, or between the observer's theory and the target's information-processing mechanisms, in the case of TT, is the simplest and most parsimonious hypothesis and, thereby, the most likely to be true. This is a variant of the argument that we have considered and rejected in the previous chapter[39]. Goldman himself is sceptical and prefers to pursue a more interesting nativist approach, which explores the analogy with Chomsky's nativist approach in linguistics

The starting point is Chomsky's influential "poverty of the stimulus argument" in support of nativist theories of language acquisition. Chomsky's analysis starts from the observation that children belonging to the same community end up acquiring the same grammar. This fact is particularly striking because grammar acquisition is radically

[39] There is indeed an important difference between the two arguments. The argument in chapter 2 is that we are justified in assuming that different people's utilities are co-scaled because *this* assumption is part of the best explanation of their behaviour. By contrast, the argument in this chapter holds that we are justified in assuming that different people's utilities are co-scaled because the assumption of *interpersonal similarity* of different people's mind-systems (either in the ST or in the TT form) is part of the best explanation of their behaviour. Despite this difference, the same objections made against the first argument apply to the second as well.

underdetermined by the empirical evidence. According to Chomsky, it is not plausible to assume that children use purely pragmatic criteria, such as simplicity and parsimony, in order to learn a specific and common grammar amongst the infinitely many possible ones that are consistent with the available empirical evidence. Instead, Chomsky suggests that children possess an innate and universal body of knowledge, which guides them in the process of language learning[40]. Such an innate and universal body of knowledge is not only important during the acquisition process. Indeed, it is the very body of knowledge on which the grammaticality judgments of adult competent speakers are based.

Goldman invites us to conceive the problem of IUCs in analogy with linguistics. The starting point is the observation that different observers reach frequent intersubjective agreement about ICs of preference strength. Their intersubjective agreement seems to suggest that they form the same beliefs about how different people's preferences compare in terms of strength. This fact is particularly striking because ICs are radically underdetermined by the empirical evidence. As the analogy with linguistics suggests, it is not plausible to assume that different observers form the same beliefs, amongst the infinite ones licensed by the empirical evidence, on the basis of purely pragmatic considerations[41]. Rather, it is more plausible to hold that they form the same beliefs on the basis of the possession of either innate and highly similar information-processing mechanisms, in the case of ST; or an innate and highly representative ToM, in the case of TT[42]. According to Goldman, if the nativist hypothesis gives linguistics "epistemic respectability", so does it with ICs of preference strength[43]. Since the nativist hypothesis is non-empirical, this means that both ST and TT may solve the epistemological problem of IUCs provided that we accept a non-empirical postulate, whose acceptance is supported by the same considerations that warrant postulating the existence of an innate and universal grammar in linguistics.

6.3 Three questions about innateness

The nativist account postulates the existence of innate cognitive mechanisms (an innate body of knowledge), with certain specific properties, in order to explain people's

[40] See CHOMSKY, N. [1980].
[41] This is equivalent to rejecting Harsanyi's assumption that considerations of arbitrariness regulate our practice of third-person mental state ascription.
[42] In the light of the remarks made in section 5.5, we might add 'universal' to the attributes of the individuals' ToM.
[43] See GOLDMAN, A. [1995a], pp. 725-726.

capacity to make comparative judgments concerning preference strengths. The crucial concept is that of 'innateness'. Three questions arise. First, what is innateness? Second, what epistemological implications does nativism have? Third, what reasons – if any – support the nativist hypothesis in the case of IUCs?

Let us consider the first question. The answer is particularly controversial and has generated, in the past few years, a particularly intense philosophical debate[44]. Although the literature presents an evident lack of agreement, the most recent positions suggest taking 'nativism' as equivalent to 'psychological primitivism'[45]. Accordingly, innate cognitive capacities are psychological primitives. In turn, psychological primitives are entities or processes that, on the one hand, are mentioned in the correct – or, perhaps, in the best - psychological explanations of human behaviour; and whose acquisition cannot be explained by any psychological theories, but only by a theory at a lower level, on the other hand.

With this definition at hand, we can move to the second question. It is worth noticing that nativism is not a theory of justification and, therefore, it has no epistemological implications by itself. However, nativism has epistemological implications when combined with either an evidentialist or a reliabilist theory of justification. In both cases, ICs of preference strength can be scientifically justified if there is scientific evidence showing that ordinary people make them on the basis of innate information processing-mechanisms that are highly similar across individuals, in the case of ST; or on the basis of an innate ToM that is highly representative of the other individuals' mind system, in the case of TT[46].

The most important issue is the one posed by the third question, i.e. the question of whether or not there is scientific evidence supporting the nativist hypothesis in the case of ICs of preference strength. It is worth noticing here that I am interpreting the nativist hypothesis in a broad way. In fact, in the case of ST, the issue of whether or not the information processing-mechanisms that the subjects use to make ICs of preference strength are innate is distinct from the issue of whether or not the information

[44] See COWIE, F. [1999], GRIFFITHS, P. [2002], SAMUELS, R. [2002], KHALIDI, M. A. [2007].
[45] See specially COWIE, F. [1999] and SAMUELS, R. [2002].
[46] Suppose that the non-empirical postulate of innate similarity can be vindicated. Suppose also that simulation (the individual's ToM) is generally reliable for the purpose of making ICs. The question still remains of whether or not each particular interpersonal comparison of preference strengths is scientifically justified. After all, a typically reliable mechanism such as vision may produce false beliefs under certain unfavourable circumstances, e.g. when the individual is drunk or is hallucinating. An alternative consists in qualifying the original reliability condition by saying that actual scientific justification requires that a belief about how different people's preferences compare in terms of strength is acquired through properly working cognitive mechanisms. This move brings scientific justification closer to a specific version of reliabilism, namely, proper functionalism. See PLANTINGA, A. [1993].

processing-mechanisms that the subjects use to make ICs of preference strength are highly similar across individuals. Goldman's argument may work only provided that we take interpersonal psychological similarity itself to be an innate feature of the mind. Thus, the nativist hypothesis can be reformulated as the assumption of innate interpersonal psychological similarity. Likewise, in the case of TT, the issue of whether or not the ToM that the subjects use to make ICs of preference strength is innate is distinct from the issue of whether or not the ToM that the subjects use to make ICs of preference strength is highly representative of the other individuals' mind-system. Goldman's argument may work only provided that we take ToM-to-mind similarity itself to be an innate feature of the mind. Thus, the nativist hypothesis can be reformulated as the assumption of innate ToM-to-mind similarity.

In order to assess the nativist hypothesis so conceived, we need to consider the elements defining the notion of innateness. With respect to the first element, the assumption that a cognitive capacity is innate is justified provided that it is part of the correct – or part of the best – psychological explanation of human behaviour. In the case under consideration, the assumption of innate interpersonal psychological similarity (innate ToM-to-mind-similarity) is justified provided that it is part of the best explanation of why the folks reach frequent intersubjective agreement about ICs of preference strength. With respect to the second element, the assumption that a cognitive capacity is innate is justified provided that its acquisition cannot be explained by any psychological theories, but only by a theory at a lower level. In the case under consideration, the assumption of innate interpersonal psychological similarity (innate ToM-to-mind-similarity) is justified provided that its acquisition cannot be explained in terms of the interpersonal similarity of other psychological processes (bodies of information).

6.4 Objections

Let us consider the first claim. The first question to ask is whether it is really the case that the folks reach frequent intersubjective agreement about ICs of preference strength. This is an empirical question. Therefore, the answer to this question requires at least some empirical data. However, for the present purpose, I am prepared to grant that the assumption of frequent intersubjective agreement is likely to be corroborated by the empirical evidence.

The second and most important question to ask is whether the assumption of innate interpersonal psychological similarity (innate ToM-to-mind-similarity) is really part of the best explanation of people's intersubjective agreement. Notice that this argument is nothing but an inference to the best explanation argument, of the same kind of those that we have examined in chapter 2. Thus, if the previous analysis is correct, Goldman's argument from nativism is bound to fail. Some doubts about it come directly from the analysis conducted in the previous chapter. There, we saw that a decision-theoretic explanation of human behaviour does not require any assumption concerning the comparability of preferences. The scope of the explanation remains the same without any such assumption. Shall we, thereby, reject the nativist hypothesis from the start?

One might think that the matter is more complex. Decision theory offers an explanation of human behaviour in terms of the content and strength of each individual's desires and beliefs. For instance, decision theory explains how a judge makes ICs of preference strength in terms of the content and strength of his desires and beliefs. However, one can pose the question of why the judge has those specific beliefs about how different individuals' preference strengths compare. Typically, decision theory remains neutral about the cognitive mechanisms that lie behind the judge's process of belief formation. Thus, a different, but in no way incompatible, explanation of how the judge makes ICs of preference strength may start precisely from an assumption concerning his cognitive architecture. Such an explanation would strengthen the decision-theoretic approach by showing how the evidence possessed by the judge generates specific comparative beliefs through the workings of certain cognitive mechanisms.

Although this strategy offers a more favourable prospect, there are at least three other objections that can be raised against the nativist hypothesis. The first objection challenges nativism. The charge is that the interpersonal psychological similarity (ToM-to-mind-similarity) across individuals is not an innate feature of the mind. One may explain the folks' intersubjective agreement by reference to a capacity that they learn either by theorising or by enculturation, rather than possess innately. For instance, one may conjecture that the folks are taught how to compare the intensity of other people's preferences in certain token circumstances and, then, generalise such ICs to circumstances of the same type[47]. However, this objection is not particularly damaging.

[47] Notice that this objection is open only to a TT approach to mental ascription. In fact, ST always – at least to my knowledge – relies on nativist accounts concerning the acquisition of the mindreading capacity.

What matters is whether the assumption of interpersonal psychological similarity (ToM-to-mind-similarity) holds, independently of whether similarity is innate or not.

The second objection questions the location of similarity between individuals. The charge is that we can explain the folks' intersubjective agreement about ICs of preference strength without postulating highly similar information-processing mechanisms (a highly representative ToM). For instance, one idea is that intersubjective agreement is due to the recognition of certain facts as particularly salient. This account presupposes a certain degree of interpersonal isomorphy in belief formation. However, it does not imply that isomorphy concerns the information-processing mechanisms (the body of causal knowledge) used to form beliefs about how different people's preferences compare in terms of strength.

The third and most powerful objection is that, even if we locate similarity between individuals where the nativist hypothesis suggests, the assumption of interpersonal psychological similarity (ToM-to-mind-similarity) is not part of the best explanation of people's intersubjective agreement. On the contrary, at best, the assumption of interpersonal similarity (ToM-to-mind-similarity) is explanatorily on a par with the assumption that the folks *take* each other to be similar at the level of the relevant information-processing mechanisms (the ToM) used to make ICs of preference strength. At worst, it is explanatorily inferior. Let me illustrate why with an example.

Consider Goldman's ST account. Suppose there are two individuals, *i* and *j*, simulating each other's mental life. Suppose there is evidence that the first and the second reliability requirements are satisfied. Suppose also that both individuals are completely identical at the level of their practical reasoning systems. Finally, suppose that they are completely identical at the level of their response mechanisms, except for the fact that the individual *i* forms desires with intensity ten times greater than the individual *j*'s, when responding to the same environmental stimuli. Will the two individuals reach intersubjective agreement about ICs of preference strength? The answer is affirmative. Under Goldman's ST account, both individuals ascribe preferences to each other and to themselves on the basis of their own cognitive machinery. This means that each subject *takes* the target individual to be just like him. In other words, each subject *takes* the assumption of interpersonal psychological similarity to be satisfied. On the basis of this assumption, individual *i* ascribes the same preference strengths both to himself and to individual *j*, when they are subject to the same environmental circumstances; and so does individual *j*. By so doing, both individuals conclude that they have the same preference strengths, despite the fact that,

by stipulation, the intensity of *j*'s desires is ten times greater than the intensity of *i*'s desires.

The point of the exercise is the following. The folks' intersubjective agreement about ICs of preference strength can be equally explained by two different accounts. The first assumes that the folks are psychologically similar at the level of the relevant information-processing mechanisms. The second assumes that they merely *take* each other to be psychologically similar at the level of the relevant information-processing mechanisms. However, simulation is reliable for the purpose of making ICs of preference strength only if the former account is true, whereas it is unreliable if the latter account is true. In the former case, ICs of preference strength can be scientifically justified; in the latter case they cannot.

So far we have shown that the assumption that the folks *take* each other to be highly similar at the level of information-processing mechanisms is explanatorily on a par with the assumption of interpersonal psychological similarity. This is enough to show that the nativist project fails to offer a conclusive solution to the problem of IUCs, because, as we have seen in the previous chapter, no additional non-empirical considerations can help us adjudicate between the two assumptions. However, it may be tempting to argue that the former assumption is also explanatorily better than the latter. This can be done by resorting to the evolutionary argument discussed in section 5.2. As we have seen above, evolution might favour a degree of interpersonal psychological similarity that *optimises*, rather than maximises, the individuals' expected fitness. In turn, if expected fitness is assessed with respect to the benefits coming from intersubjective agreement, on the one hand, and the costs coming from the development of highly specific information-processing mechanisms, it follows that evolution might have favoured the development of both a less stringent degree of similarity than the one required in order to have reliable ICs of preference strength *and* the folks' attitude of taking each other to be alike, or highly similar, in certain relevant respects.

The same objection applies to the TT case. The folks' intersubjective agreement about ICs of preference strength can be equally explained by two different accounts. The first assumes that the ToM that the folks use is highly representative of the other individuals' mind. The second assumes that the folks merely take the ToM that they use to be highly representative of the other individuals' mind. However, TT is reliable for the purpose of making ICs of preference strength only if the former account is true, whereas it is unreliable if the latter account is true. If there is no reason to favour the former over the latter, the nativist project fails to offer a conclusive solution to the

epistemological problem of IUCs. Furthermore, it is possible to argue in evolutionary terms that the assumption that the folks take the ToM to be highly representative of the other individuals' mind is explanatorily better than the assumption of ToM-to-mind similarity.

If the previous point is not sufficient to discard nativism as a solution to the problem of IUCs, the sceptic might get further support by considering the second condition required to justify the assumption of innateness. According to it, the assumption that a cognitive capacity is innate is sound provided that its acquisition cannot be explained by any psychological theories, but only by a theory at a lower level. If we read this condition in a weak sense, this means only that it must be possible to offer an explanation of how certain psychological mechanisms are realised at the physical level. If we read the condition in a stronger sense, this means that the account at the physical level must be able to vindicate the nativist assumption at the psychological level.

As far as the problem of ICs is concerned, it is certainly possible to explain the acquisition of highly similar information-processing mechanisms (a highly representative ToM) at the neurophysiological level. However, even if we grant the claim that we are somehow made of the same neurophysiological 'stuff', this is not sufficient to vindicate the nativist hypothesis at the psychological level. To begin with, although there is evidence that some mental states, e.g. disgust, are located in specific brain regions and, thereby, that different individuals undergoing those states share common neural properties, the same is not true for other mental states, like preferences. In other words, there is yet no evidence that undergoing a preference state activates the same neural region in different individuals. However, this may simply be a problem of limited empirical evidence. It is possible that one day scientific research will discover the neural correlates of preferences.

Even if we grant this possibility, the prospects for ICs are dim. The existence of a common neural dedicated to preference formation does not show, *per se*, that preferences are formed by means of highly similar information-processing mechanisms. The reason is that, even if different individuals' neurons fire with the same intensity, it does not follow that their preference strengths are identical. Consider ST first. Under the experientialist conception of preferences, it is possible that identical neuronal activation across individuals corresponds to preference experiences that are very different at the level of strength, at least *if* we grant the possibility that the qualitative character of experiences is not fully accounted in terms of their neurophysiological character. This

is the same as admitting that interpersonal isomorphy at the physical level does not necessarily imply interpersonal isomorphy at the subjective level.

Consider TT now. Since preferences are now given a functionalist understanding, the argument must be different, as reference to alleged interpersonal differences at the experiential level are excluded by definition. Nevertheless, we can equally argue that, even if different individuals' neurons fire with the same intensity, it does not follow that their preference strengths are identical. The reason is the following. Preference strengths are individuated not only with respect to external inputs and outputs, but also with respect to other preferences and mental states. Crucially, these mental states can be both occurrent *and* non-occurrent. The problem is that neural activation registers only occurrent preferences and mental states. Thus, in order to conclude that different individuals' preference strengths are the same when their neurons fire with the same intensity, we need to assume that they are identical with respect to all the other non-occurrent preferences and mental states, which might impact on their preference strengths. However, it seems to be epistemically impossible to verify whether or not this assumption holds. The result is that, once again, interpersonal isomorphy at the physical level does not imply interpersonal isomorphy at the functional level.

To conclude, even if it is possible to account for the acquisition of each individual's mindreading capacity at a lower level, no empirical support can be offered for the assumption of innate interpersonal psychological similarity (innate ToM-to-mind similarity) at such level.

7. Conclusion

In this chapter, I considered the question of whether or not a nativist argument shows that ICs of preference strength can be scientifically justified. The argument is made in the context of current debates in philosophy of mind concerning the explanation of mental ascription and the meaning of mental states. I considered both ST and TT accounts of mindreading at the sub-personal level of description, together with the associated experientialist and functionalist accounts of the meaning of preferences.

Within the ST framework, the nativist argument holds that we are justified in assuming that different people's utilities are co-scaled if it is an innate feature of the mind that the information-processing mechanisms that people use to make ICs of preference strength are highly similar to the information-processing mechanisms that

other individuals use to form their preferences. I referred to it as the assumption of innate interpersonal psychological similarity.

Within the TT framework, the nativist argument holds that we are justified in assuming that different people's utilities are co-scaled if it is an innate feature of the mind that the ToM that the subject uses to ascribe preferences is highly representative of the information-processing mechanisms through which different people form their preferences. I referred to it as the assumption of innate ToM-to-mind similarity.

In this chapter I rejected the nativist argument in both forms. I argued that the reasons offered in support of the nativist hypothesis do not establish the soundness of either the assumption of innate interpersonal psychological similarity or the assumption of innate ToM-to-mind similarity. The conclusion is that we still lack a reason to think that our beliefs about how different people's preferences compare in terms of strength can be scientifically justified.

CHAPTER 4

Three 'possibility' arguments

1. Introduction

The arguments examined in the previous chapters fail to show that we can have scientific knowledge of, or, at least, scientifically justified, ICs of preference strength. Neither the appeal to an inference to the best explanation type of argument nor the appeal to a nativist argument offers a positive solution to the problem. These results increase the pressure brought by the sceptical challenge. One issue concerns the assessment of when a belief about how different people's preferences compare in terms of strength can be said to be true. Consider the contrast between predictions about an agent's behaviour and IUCs. In the former case, it is relatively easy to assess whether or not a prediction is correct. We simply have to look at the agent's behaviour and see if it corresponds to the predicted one. By contrast, unlike behaviour, mental states are unobservable. We cannot simply observe whether or not different people display the attributed mental states. As a consequence, we cannot simply observe whether or not the comparison of different people's preference strength is correct.

This issue, combined with the difficulties in finding a solution to the epistemological problem of IUCs, raises the doubt that there may be no fact of the matter about IUCs. In other words, the radical thought is that the alleged impossibility of solving the epistemological problem of IUCs does not stem only from epistemological limitations but, more radically, from the ontological incomparability of preferences with respect to the dimension of strength. In order to address this challenge, some authors have elaborated 'in principle' solutions to the problem of IUCs. These solutions are based on 'possibility' arguments. Their primary goal is to show that different people's preference strengths are indeed comparable. Their secondary goal is to show that it is possible, in principle but not by means of empirical or pragmatic considerations only, to have scientific knowledge of, or scientifically justified beliefs about, how different people's preferences compare in terms of strength.

In this chapter I want to examine three arguments of this sort. Broome offers the first argument that I shall consider. Although Broome takes the betterness relation as his object of interest, his argument can be extended, with few modifications, to the preference

relation. In section 2, I shall illustrate the main features of Broome's strategy and discuss some objections against it. The second argument that I will consider is based on a functionalist understanding of the nature of preferences. I shall illustrate this argument and present some objections against it in section 3. Bradley offers the third argument. It moves from an analogous understanding of the nature of preferences but argues for the interpersonal comparability of preferences in a different way. I shall illustrate this argument and discuss some objections against it in section 4. Finally, I shall summarise my results and conclude in section 5.

2. Broome's strategy

2.1 Broome's argument

In this section I shall illustrate the approach that Broome has put forward in his recent book *Weighing Lives*[1]. Broome is not explicitly concerned with preferences, but with well-being. According to him, "wellbeing is not an empirical concept"[2]. Although "economists generally hope to measure wellbeing by means of people's preferences", so to make preferences "the basis for measuring wellbeing empirically", Broome thinks that individual well-being should be founded on the non empirical notion of a person's betterness relation. Broome assumes that the betterness relation satisfies the expected utility axioms and, thereby, can be represented by an interval utility function, unique up to a positive affine transformation. Thus, in his approach, utility defines the value of a function that measures degrees of personal goodness.

Broome's argument is based on four assumptions. The first is that a person's goodness is supervenient upon "how things are for that person". These are features of the world that appear from that person's perspective and that affect that person's goodness. Broome calls the set of such features a person's 'life'. Thus, his first assumption is that personal goodness supervenes upon a person's life. What features of the world can figure as component of one's person's life is an open question, which depends on the substantive theory of personal goodness that one embraces. Nevertheless, according to Broome, there is at least one point on which most accounts of well-being can plausibly be expected to agree: a person's bare identity does not figure amongst the features of the world on which personal goodness supervenes. This means that personal goodness is independent from

[1] BROOME, J. [2004].
[2] See BROOME, J. [2004], pp. 78-79.

personal identity, or, which is the same, that "the goodness of a life is independent of who lives it"[3]. This is indeed Broome's second assumption.

This assumption implies the conceptual, or metaphysical, possibility that the same life can be lived by two different persons. In other words, it implies the existence of a possible world where the life that an individual *j* lives in the actual world is lived by another individual *i*. The underlying idea is that, once we exclude bare identity from an individual *i*'s life, it may be possible for another individual *j* to occupy *i*'s position with respect to all the features that figure from *i*'s perspective. This is particularly important, because it provides the basis for a metaphysical reduction of the intra-personal case to the intra-personal case. In turn, this metaphysical reduction provides the grounds for the conceptual possibility of having meaningful ICs of different individuals' lives. The central idea is that *if* the same person can live other persons' lives, those lives become comparable in terms of personal goodness[4]. Being lives that *i* can live, they can be compared in terms of how good they are for *i*. In other words, a betterness relation exists amongst all the lives that *i* can possibly live. Moreover, this relation holds independently of whether or not these lives are actually lived by other persons. Indeed, all the lives that *i* can possibly live remain comparable even when they are actually lived by individuals different from *i*.

A caveat. Broome's second assumption does not imply that each individual can live any other individual's life. Nor does it imply that, for *any* individual *j*, there is at least one possible world where another individual *i* lives individual *j*'s life. Thus, the argument in the previous paragraph shows only that it is possible to compare *some* individuals' lives, i.e. those lives that can be equally lived by different individuals. In order for the comparisons of *all* individuals' lives to be possible, Broome needs to make further assumptions. In order to highlight what these are, let us consider Broome's argument in more detail.

Recall that, as we have seen in chapter 1, in order to fix an interval scale of measurement, we need to fix two points, corresponding to the (arbitrary) zero and the (arbitrary) unit. Suppose that an individual *i* can live another individual *j*'s life. If personal goodness is independent from personal identity, it follows that the value of that life is identical for both *i* and *j*. This fixes a common point in their utility scales. In order to claim that their utility scales are the same, we need to find another point in common. This is possible if there is another life that both individuals can possibly live. Given that personal goodness is independent from personal identity, the value of this life is identical for both *i* and *j*. Once these two points are fixed, we can conclude that individual *i*'s and *j*'s utility

[3] BROOME, J. [2004], p. 94.
[4] See BROOME, J. [2007].

functions are co-scaled. If two persons share at least two possible lives, they form an "overlapping pair"[5]. Being an overlapping pair is a necessary condition for two persons' personal goodness to be comparable. Broome's third assumption is thus that every individual belongs to at least one overlapping pair.

This is yet not enough to guarantee that all individuals' lives are comparable. It is easy to understand why. Suppose, as before, that *i* and *j* form an overlapping pair. It follows that their utilities are co-scaled. Suppose now that there are two other individuals *k* and *h*, who form another overlapping pair. Their utilities are also co-scaled. However, suppose that there are no possible lives that the first overlapping pair, i.e. *i* and *j*, have in common with the second overlapping pair, i.e. *k* and *h*. If this is the case, although it is possible to make ICs of personal goodness *within* each overlapping pair, it is not possible to make ICs of personal goodness *across* different overlapping pairs. Given that there is no common point between, say, *i*'s and *k*'s utility scales, it is not possible to claim that they are co-scaled. On the other hand, if *i* and *k* also form an overlapping pair, then their utilities are co-scaled and their personal goodness is comparable. Moreover, it also follows by transitivity that *i*'s personal goodness is comparable with *h*'s, on the one hand, and that *k*'s personal goodness is comparable with *j*'s, on the other hand. The idea is that it is possible to compare *all* individuals' personal goodness if everyone is suitable related to everyone else by means of a "chain of overlapping pairs". This is indeed Broome's fourth and last assumption. Together, these assumptions imply that personal goodness can be measured on a universal scale and compared across all individuals and states of the world.

2.2 *Objections*

Broome's argument is based on four assumptions, which we can summarise as follows:

(B1) Personal goodness supervenes on a person's life;
(B2) Personal goodness is independent from personal identity;
(B3) Every person belongs to at least one overlapping pair;
(B4) Every person is related to everyone else by a chain of overlapping pairs.

In what follows, I will focus in particular on the implications that (B1) and (B2) have for the problem of IUCs. Before proceeding further, it is worth noticing one point. The goal of this section is to discuss whether or not Broome's approach shows that preferences are

[5] BROOME, J. [2004], p. 96.

interpersonally comparable in terms of strength. However, Broome does not consider the preference relation as his object of interest. Rather, he focuses on the betterness relation, which he explicitly characterises as a non-empirical relation. What is then the relevance of Broome's argument for the issue at stake? We can answer in two ways. The first is more indirect. It consists in claiming that, as a matter of fact, Broome's argument *is* compatible with a preference satisfaction view of personal goodness *and* that, in turn, the degree to which preferences are satisfied is given by the intensity with which the individual prefers the option that the world realises. Thus, we can apply Broome's argument to preference satisfaction and see whether or not it helps us addressing the problem of the interpersonal comparability of different people's preferences.

The second answer is more direct. The idea is that Broome's argument not only is, but also *must* be, compatible with a preference satisfaction view of personal goodness. One of the underlying preoccupations in Broome's work concerning the structure of personal goodness is to remain neutral between alternative substantive conceptions of it. Hence, his solution to the problem of IUCs must be applicable to various, and possibly very different, specification of what constitutes personal goodness. Although Broome has argued elsewhere against a preference-based theory of well-being[6], his solution to the problem of IUCs must be independent from the soundness of that criticism. That is, it must be able to accommodate the case where the preference satisfaction theory of well-being is indeed the correct theory. At the very least, it must be able to accommodate the case where preferences are a component of well-being[7].

For simplicity, in what follows, I shall assume that personal goodness is entirely supervenient on facts about individual preferences. More specifically, I shall assume that there is a direct relation between personal goodness and preference satisfaction. Moreover, I shall assume that the degree to which preferences are satisfied is given by the intensity of people's preferences. From these assumptions, it follows that a person's life is constituted *entirely* by that person's (realised) preferences. We can thus rewrite Broome's first assumption in the following way

[6] See BROOME, J. [forthcoming].

[7] In a recent article, Broome has explicitly withdrawn his subscription to this neutrality requirement. More precisely, Broome now argues that his account about the structure of personal goodness is not compatible with preference satisfaction accounts of personal goodness. See BROOME, J. [2007]. Broome's latest position marks a significant change from the earlier position expressed in *Weighing Lives*. Most importantly, it has two important implications. On the one hand, it significantly weakens the strength of his project. Indeed, if personal goodness turns out to be of the substantive kind that Broome rejects, his work is not really about the structure of personal goodness. On the other hand, his new position seems irremediably *ad hoc*. Indeed, as we shall see below, Broome's approach is problematic for *all* accounts of personal goodness which include some mental states, e.g. preferences, desires, emotions, as features of a person's life

(B1*) Personal goodness supervenes on a person's (realised) preferences.

Let us now consider Broome's assumptions in more detail. According to the standard definition of supervenience, a set of properties supervenes on another set of properties if two things cannot differ with respect to the former set without differing also with respect to the latter. In the case under consideration, this definition admits of two readings. According to a weak reading, personal goodness supervenes on a person's life if, *for the same person*, two possible lives of his cannot differ in terms of their goodness without differing also at the level of his realised preferences. However, Broome favours a stronger reading, according to which personal goodness supervenes on a person's life if, *for any two individuals*, two possible lives of theirs cannot differ in terms of their goodness without differing also at the level of their realised preferences. Clearly, the difference is that the latter reading includes an interpersonal element, whereas the former reading does not. Thus, Broome's reading implicitly assumes that the supervenience relation is the same for different individuals, whereas the weaker reading allows for the possibility that the supervenience relation is different for different individuals.

It is worth noticing that Broome's approach does not need the stronger reading. His argument requires only that there be at least some pair of individuals such that one can live the other's life. In other words, his argument requires that, for at least some pairs of individuals *i* and *j*, *i* can live *j*'s life. Indeed, if *i* can occupy *j*'s position, then *i* has the same personal goodness as *j*. The reason is that, by occupying *j*'s position, *i* acquires also the same supervenience relation occurring between *j*'s life and *j*'s personal goodness. Whether this relation is the same as the one between *i*'s life and *i*'s personal goodness is not relevant. The two may as well be different. Indeed, once it is established that *i*'s personal goodness is the same as *j*'s, it is also established that it is possible to compare *i*'s and *j*'s well-being with respect to that life. For this purpose, it only necessary to assume that the possibility of living someone else's life entails the acquisition of the very same supervenience relation occurring between that individual's life and his personal goodness. If this is granted, the weaker reading of the supervenience relation is sufficient for Broome's argument.

The main question is thus whether or not it is indeed possible for one individual to live another individual's life, as (B2) seems to imply. The first objection is that, if we embrace (B1*), Broome's argument is question-begging. In order for individual *i* and individual *j* to have the same personal goodness, it must be the case that, once individual *i* is endowed with individual *j*'s own life, we can say that individual *i* has the same preferences as *j*.

Amongst other things, this means that it must be possible for individual *i* to have the same preference *strengths* as *j*. However, this presupposes what needs to be proven, namely, that different people's preference strengths are interpersonally comparable. If we cannot meaningfully assume that *i* has the same preference strengths as *j*, then we cannot conclude that *i* has the same personal goodness as *j*. It thus appears that we should reject (B2), at least when it is combined with an account of a person's life that makes references to preferences, like (B1*).

There is one possibility to avoid the objection. One may claim that it is possible to retain (B2) by arguing that, contrary to what one may think, (B2) does not imply the question-begging assumption that preferences are interpersonally comparable in terms of strength. The idea is that (B2) simply implies the possibility that individual *i* can occupy individual *j*'s mental location. At such location, individual *i* has individual *j*'s very own mental states and, in particular, the content and structural properties of *j*'s very own preferences. As a matter of metaphysical possibility, this can happen independently of whether preferences are interpersonally comparable in terms of strength. Thus, it may be true that, if personal goodness is independent from personal identity, then individual *i* can live individual *j*'s life and assume *j*'s individual preferences. However, endowing individual *i* with *j*'s preferences does not presuppose that *j*'s preferences are interpersonally comparable in terms of strength with individual *i*'s own preferences.

This move comes at a cost. The result is that Broome's argument does not prove that different people's preferences are comparable in terms of strength, but rather presupposes that in fact they may not be. The reason is that this reply allows for the possibility that individual *i* has individual *j*'s own preference strengths and, at the same time, that their preferences are not comparable in terms of strength. More generally, this move allows for the possibility that individual *i* can live individual *j*'s life, without sanctioning the conclusion that individual *i*'s life is comparable with individual *j*'s life in terms of personal goodness.

There is another possibility to deal with the initial objection. It consists in modifying (B1*) in the following way.

(B1**) Personal goodness supervenes on the relevant facts about a person's (realised) preferences.

This move suggests a more indirect, non question-begging, way to argue for the interpersonal comparability of preferences. More specifically, the assumption is that

individual *i* can be endowed with those facts about individual *j*'s preference strengths that figure in *j*'s life. Possibly, these facts refer both to the behavioural outputs of *j*'s preferences and to the causal variables determining *j*'s preferences, along the lines suggested by Harsanyi's approach. The difference with Harsanyi's approach is that these are the very facts belonging to *j*'s life and not facts of the same type. The idea is that, once endowed with these facts, necessarily, individual *i* maps them into preference strengths that are identical to *j*'s. If the facts determining those preferences are interpersonally comparable, the conclusion is that individual *i* forms exactly the same – interpersonally comparable – preferences as individual *j*.

Two further objections can be raised. First of all, the success of Broome's strategy depends on whether *all* the facts about *j*'s preference strengths are interpersonally comparable. In fact, if the inputs of preferences include other mental states with a specific intensity, we encounter the same problem as in the case of preferences. Endowing individual *i* with individual *j*'s mental states presupposes that these mental states are interpersonally comparable in terms of strength. Once again, this is question-begging.

Second, the move under consideration presupposes either that the relations connecting causal variables to preference strengths, on the one hand, and preference strengths to behavioural outputs, on the other hand, are the same for all individuals, or, at least, that it is possible to account for potential differences across individuals. Indeed, even if we assume that individual *i* can be endowed with facts about individual *j*'s preference strengths and that all these facts are objective, the conclusion that individual *i* has preference strengths identical to individual *j*'s follows only if the previous assumption holds. The problem is that we cannot take the assumption concerning the interpersonal sameness of the relevant causal relations for granted and, as we have seen in the previous chapters, showing that such an assumption is sound is not a trivial matter at all.

To conclude, Broome's thought experiment shows, at best, that, once endowed with all the facts about individual *j*'s preference strengths, individual *i* has *j*'s very own utility function. However, it does not show that it is possible to compare individual *i*'s and individual *j*'s preferences in terms of strength. Thereby, it does not show that individual *i*'s and individual *j*'s utilities are commensurate. Therefore, Broome's approach is unsuccessful.

3. Can functionalism rescue IUCs?

An alternative strategy consists in exploring the possibilities offered by the adoption of a functionalist understanding of preferences. Unlike the previous chapter, this chapter focuses on functionalism as a theory about the *nature* of mental states. Thus, the relevant version is psycho-functionalism, not analytic functionalism. According to psycho-functionalism, preferences are mental states individuated in terms of their causal relations with certain inputs, other mental states, and certain behavioural responses[8]. I shall consider two solutions. In common they have the idea that functionalism sanctions the assignment of identical preference strengths to different individuals if it is possible to individuate cases where their preferences play the same causal role. On the other hand, they differ with respect to the cases that are supposed to be relevant for the comparability of different people's preferences. In this section, I shall start by reviewing some of the reasons to think that preference strengths are interpersonally incomparable. I shall then proceed by illustrating and discussing the first functionalist solution.

3.1 Reasons for incomparability

In general, one of the conditions for the interpersonal comparability of mental states is that such mental states have properties in common across individuals. More specifically, if one adopts a functionalist understanding of the nature of mental states, sameness of properties across individuals is identified by cases where the same type of mental state plays the same causal role in different individuals' minds. In the case of preferences, the problem is that it appears to be impossible to identify cases where different people's preferences have the same causal role.

We need to distinguish two types of impossibility and, correspondingly, two types of argument that can be made in support of the thesis that preferences are interpersonally incomparable in terms of strength. The first is an epistemological argument for incomparability. It starts by observing that it is *epistemically impossible* to identify cases where different people's preferences play the same causal role and argues that the best explanation for this impossibility is the ontological incomparability of different people's preference strengths. The second is a conceptual argument for incomparability. It starts by observing that it is *conceptually impossible* to identify cases where different people's preferences play the same causal role and argues that the reason for this impossibility is the ontological incomparability of different people's preference strengths. Both arguments can

[8] The relation that preferences have with both inputs and outputs may turn out to be indirect, that is, mediated by other mental states that are connected to preferences.

be made with respect to three cases. The first is the case where different individuals are isomorphic at the level of both inputs and outputs. The second is the case where different individuals are isomorphic at the level of the relevant functional relations. The third is the case where different individuals are isomorphic at he level of both inputs and outputs and at the level of the relevant functional relations.

Let us consider the first case. Suppose there are two individuals, *i* and *j*. Suppose also that the empirical evidence available at time *t* is the same for both individuals. Under a functionalist understanding of the nature of preferences, the admissible empirical evidence is constituted by both behavioural outputs and environmental inputs. Suppose that $u_i(y) = u_j(y) = 0.6$. Can we conclude that their preferences are interpersonally comparable and that individual *i* prefers option *y* with the same strength as individual *j*? Since functionalism conceives mental states in terms of causal relations between inputs, other mental states and behavioural outputs, identical empirical evidence determines IUCs only if the relevant causal relations are the same for all individuals. For instance, suppose that the requirement that different individuals respond to the same environmental inputs in the same way is violated. It is conceptually possible to imagine a situation where the same environmental inputs lead two different individuals to form preferences leading to the same behavioural outputs on the basis of different initial mental states. However, if the individuals' initial mental states are different, the intensity of their preferences may also be different, even though both the environmental inputs and the behavioural outputs are identical. Suppose now that the requirement that different individuals' mental states interact in the same way is violated. It is conceptually possible to imagine a situation where identical initial mental states determine preferences with different intensity and yet, ultimately, lead to the same behavioural outputs. This result may be due to a double difference somewhere in their mind-systems: for instance, the first fault may occur in the conversion of beliefs and desires with identical strengths into preferences with identical strengths; the second fault may occur in the conversion of preferences with different strengths into intentions leading to choices with identical strengths. Coming back to our problem, the epistemological argument claims that it is epistemically impossible to identify cases where the relevant causal relations are the same for different individuals. By contrast, the conceptual argument claims that this impossibility is conceptual. Both arguments conclude that preferences are interpersonally incomparable in terms of strength.

Let us consider the third case, from which the second can be derived as an application. Suppose there are two individuals, *i* and *j*. Suppose that the empirical evidence available at time *t* is the same for both individuals. Suppose also that *i* and *j* respond to the same

environmental inputs and form their mental states in the same way. If $u_i(y) = u_j(y) = 0.6$, can we conclude that their preferences are interpersonally comparable and that individual *i* prefers option *y* with the same strength as individual *j*? One objection is the following. Each individual's preference domain may include infinite options. Preferences for specific options become manifest in certain choice situations and, *ceteris paribus*, are revealed by the behavioural outputs included in the set of admissible empirical evidence. However, other preferences do not become manifest because no corresponding choice or behavioural opportunity is presented to the individual. Since the intensity of manifest preferences is relative to the intensity of all the options in the preference domain, including those that are not manifest in observable behaviour, it is conceptually possible to imagine a situation where two individuals have different preference strengths even though they are identical with respect to the admissible empirical evidence and the relevant causal relations. More generally, the epistemological argument claims that it is epistemically impossible to identify cases where different individual's preference strengths are the same, even when both the empirical evidence and the relevant functional relations are supposed to be identical. By contrast, the conceptual argument claims that this impossibility is conceptual. Both arguments conclude that preferences are interpersonally incomparable in terms of strength.

3.2 Are preferences unbounded?

The first functionalist argument is concerned with the conceptual argument for incomparability and with the second case of conceptual impossibility. More precisely, its goal is to show that different people's preferences are conceptually comparable when the individuals are assumed to be isomorphic at the level of the relevant causal relations.

The starting point is the idea is that a functionalist view of preferences leaves room for the existence of both a most preferred and a least preferred option, relative to a specific preference domain. Once an individual's preference domain is fixed, so are the top and the bottom options in his preference ranking. Since the preferences for the top and bottom options play the same causal role for different individuals, in relation to their respective preference domain, it follows that preferences are interpersonally comparable in terms of strength.

The main objection against this argument is that the existence of a most preferred and a least preferred option is conceptually impossible, even when they are relative to a specific preference domain, because the number of options included in the preference domain may

be infinite. The idea is that infinity precludes the existence of both an upper and a lower bound in an individual's preference ranking.

There are at least two ways to counter this objection. The first consists in denying that infinity is sufficient to preclude the existence of a most preferred and a least preferred option in an individual's preference ranking. In a nutshell, the thesis is that, even if the number of items in the preference domain tends to infinity, this may not, *per se*, prevent the existence of both a best and a worst option in an individual's preference ranking. In other words, infinity alone cannot be the source of conceptual impossibility.

The second consists in arguing that an individuals' preference domain never includes an infinite number of options. This reply invites to consider an individual's preference domain across his entire life. The idea is that, although an individual can potentially consider all sorts of options as objects of preferences, at the end of his life he will have considered only a certain number of options. This means that, although his lifetime preference domain may contain an uncountable number of options, it does not contain an infinite number of options. This is enough to make the existence of a most preferred and a least preferred option conceptually possible[9].

If either one of these replies works, the result is that it is possible to provide a solution to the metaphysical problem of IUCs. How about the epistemological problem of IUCs? Recall the two goals of a 'possibility' argument. The primary goal is to show that it is conceptually possible to compare different people's preferences in terms of strength. The secondary goal is to show that this result can be used to defend the epistemic possibility of meaningful ICs of preference strength. One objection is that, *even if* the first functionalist argument can provide a solution to the conceptual problem concerning the interpersonal comparability of preference strengths, it does not provide a solution to the epistemological problem of IUCs. The main charge is that, although it might be conceptually possible to defend the existence of a best and a worst option in each individual's preference ranking, it is not epistemically possible to identify these options. Both when the preference domain contains an infinite number of items and when it contains a finite number, the empirical evidence leaves the identification of the top and the bottom options in each individual's preference ranking underdetermined. The reason is that, in both cases, the individual may form preferences that never become manifest in behavioural outputs, so that no amount of empirical evidence is sufficient to detect what the upper and lower bound in the individual's preference ranking really is. The conclusion is that the first functionalist

[9] There is at least another way to counter this objection. It consists in arguing that it is conceptually possible to reduce an infinite preference domain to a finite one. This is an approach that invites further investigation. In what follows, however, I shall ignore this possibility.

solution does not show that we can have epistemic access to cases where different individual's preferences play the same causal role.

There is a stronger version of this objection. Recall that the first functionalist argument assumes that different individuals are isomorphic at the level of the relevant causal relations. The argument under consideration does not offer any reason to think that it is conceptually possible to compare different people's preferences in terms of strength when this assumption is relaxed. *A fortiori*, the argument under consideration does not offer any reason to think that it is possible to have epistemic access to the facts about the relevant causal relations. Indeed, as we have seen in the previous chapters, both empirical and non-empirical strategies fail to vindicate the assumption that different individuals both respond to the same environmental inputs and form their mental states in the same way. The conclusion is that the first functionalist solution fails to solve the epistemological problem of IUCs.

4. Bradley's strategy

4.1 Bradley's argument

Bradley offers the second functionalist solution that I consider in this chapter[10]. Bradley's analysis too is concerned with the conceptual argument for incomparability and with the second case of conceptual impossibility. More precisely, Bradley's goal is to show that different people's preferences are conceptually comparable when the individuals are assumed to be isomorphic at the level of the relevant causal relations. Bradley confronts the problem of ICs of degrees of preference with the problem of ICs of degrees of belief. Amongst other things, beliefs differ from preferences because they are supposed to be interpersonally comparable. Indeed, the mainstream view holds that there are cases where different people's beliefs play the same causal role, whereas there are no cases where different people's preferences play the same causal role.

Bradley thinks that this conclusion is unwarranted. As shown by Ramsey's method – the *locus classicus* of the literature concerning the measurement of degrees of belief – degrees of belief are derived from a preference ordering and a background theory of action. Bradley's suggestion is that, if degrees of belief are interpersonally comparable, *relative to the background theory of action used to measure them*, there is no conceptual or technical reason to hold that degrees of preference are not interpersonally comparable, *relative to the*

[10] See BRADLEY, R. [2007b].

same background theory of action. The idea is that, relative to the same theory of action, it is possible to conceptually identify both cases where different people's beliefs play the same causal role *and* cases where different people's preferences play the same causal role. To see how, let us examine Ramsey's method in more detail[11].

Ramsey's problem is that of determining a measure of both degrees of belief and degrees of preference from evidence about an agent's observable (either verbal or non-verbal) behaviour. The problem lies in the difficulty of measuring two unknown variables (i.e. degrees of belief and degrees of preference) while disposing of only one known variable (i.e. observable behaviour). Ramsey makes three assumptions. First, he assumes that preferences range over a particularly rich set of prospects, which includes conditional prospects, that is, prospects that can be expressed with (indicative) conditionals of the form $x \rightarrow y$. Second, he assumes that, for an arbitrary individual i, the empirical evidence, in the form of observable behaviour, is sufficient to determine a preference ordering amongst prospects. Third, he adopts the expected utility theory as his background theory of action.

In order to measure the intensity of the individual's attitudes over prospects, Ramsey introduces the notion of an ethically neutral proposition. According to Ramsey, "an atomic proposition p is called ethically neutral if two possible worlds differing only in regard to the truth of p are always of equal value"[12], or, as Bradley puts it, "a proposition whose truth or falsity is a matter of indifference to the agent irrespective of what else is the case"[13]. Ramsey uses the notion of ethically neutral proposition as his Archimedean point. In fact, if one assumes the expected utility theory as one's background theory of action, an ethically neutral proposition p turns out to be believed with degree 0.5 if, for any simple prospects x and y such that $x \, P \, y$, the agent is indifferent between the pairs of more complex prospects $(p \rightarrow x)\ (\neg p \rightarrow y)$ and $(\neg p \rightarrow x)\ (p \rightarrow y)$. Indeed, this is the only way in which an agent preferring x to y can be indifferent to prospects $(p \rightarrow x)\ (\neg p \rightarrow y)$ and $(\neg p \rightarrow x)\ (p \rightarrow y)$ if he is an expected utility maximizer.

Once the degrees of belief for the ethically neutral proposition p have been fixed, it is possible to assign utility values, that is, numerical values representing degrees of preference, to all the prospects in the preference domain. Suppose x and y are prospects such that $x \, P \, y$. Suppose also that we arbitrarily fix $u(x) = 1$ and $u(y) = 0$. Finally, suppose that p is an ethically neutral proposition of degree of belief 0.5. Then, we can find a prospect z, such that z is mid-way between x and y, if the agent is indifferent between the pairs of complex prospects $(p \rightarrow x)\ (\neg p \rightarrow y)$ and $(\neg p \rightarrow z)\ (p \rightarrow z)$. Given that $u(x) = 1$

[11] See RAMSEY, F. P. [1990].
[12] See RAMSEY, F. P. [1990], p. 73.
[13] See BRADLEY, R. [2007b], p. 6.

and u(y) = 0, we can assign the utility value 0.5 to z, so that u(z) = 0.5. By reiterating this procedure, we can assign a measure of preference strength to all the prospects in the preference domain. The utility scale so defined is unique up to a positive affine transformation.

Once the utility values are fixed, it is possible to assign a probability value to all propositions, including those that are not ethically neutral. The task is easier. Given Ramsey's adoption of the expected utility theory as background theory of action, degrees of belief are determined accordingly. For instance, the degree of belief on the proposition z seen above is given by the following formula:

$$\Pr(z) = \frac{u(z) - u(y)}{u(x) - u(y)}$$

From this, we can get that the degree of belief on the necessary propositions is equal to 1 and the degree of belief on the impossible propositions in equal to 0. Thus, Ramsey's method shows that the probability scale so defined is absolutely unique.

As we have seen above, degrees of belief are supposed to be comparable across people, whereas degrees of preference are not. According to Bradley, in the former case comparability is assured by the fact that it is possible to conceptually identify two points with respect to which different people's beliefs play the same causal role. These are the necessary proposition and the impossible proposition. Ultimately, then, different people's degrees of belief are comparable for two reasons: the first is that "it belongs to the concept of partial belief" that there is a maximum and a minimum; the second is that such a maximum and minimum are common for different individuals[14]. By contrast, according to Bradley, it does not belong to the concept of degrees of preference that a maximum and a minimum exist. Moreover, it does not belong to the concept of degrees of preference that the items that occupy the top (the bottom) position in different individuals' preference rankings are preferred with the same strength, for the reasons that we have seen in chapter 1.

However, Bradley holds that Ramsey's method has the resources for the conceptual identification of at least one point with respect to which different people's preferences play the same causal role. The use of Ramsey's method is partly complicated by his vague characterisation of the relationship occurring between the objects of beliefs, i.e.

[14] See BRADLEY, R. [2007b], p. 8.

propositions, and the objects of preferences, i.e. prospects. However, as Bradley shows, it is possible to avoid inconsistencies by simply postulating that both beliefs and preferences range over prospects. According to Bradley, then, in the case of preferences, it is a mistake to look for interpersonally common points by trying to find an absolute maximum and an absolute minimum. Rather, one of the relevant points is given by the ethically neutral prospect, i.e. the prospect with formal properties analogous to the ethically neutral proposition. Bradley's main idea is that different people's preferences for ethically neutral prospects are comparable in terms of strength because they have the same causal force, or better, because they manifest "absence of force"[15]. Although what counts as an ethically neutral prospect may differ for different individuals, it nonetheless is the object of zero strength preferences, for all individuals. If this is true, the ethically neutral prospect identifies the 'natural' zero in the utility scale. Moreover, this 'natural' zero is common for different individuals. This means that, for any two individuals *i* and *j*, with ethically neutral prospects *p* and *q*, respectively, it is the case that $u_i(p) = u_j(q) = 0$.

Bradley's proposal is completed by a suggestion as to how to fix a common unity of the scale used to measure preferences. The idea is that the total desirability of all the prospects in the preference domain provides a common point for different individuals. Once we assume a functionalist understanding of the nature of preferences and identical causal relations across individuals, the total available preferential strength is supposed to be the same for different individuals. If we understand preferences in terms of their role in the individual's mind, then we may think that their role is fixed by the set of alternatives that the individual considers. Thus, everything else being equal, once the set of alternatives is fixed, the total available preferential strength is also fixed and is the same for different individuals. If this is the case, Bradley's account shows that "given our background theory of action, comparisons of relative strengths of preference are meaningful"[16]. More formally, Bradley's suggestion is to co-scale different people's utilities through the application of the following formula:

$$\frac{u_i(x)}{\sum_{\text{for all } x} u_i(x)}$$, for all individuals *i*

[15] See BRADLEY, R. [2007b], p. 9.
[16] See BRADLEY, R. [2007b], p. 9.

This formula effectively assigns a constant utility value equal to 1 to the total desirability of the prospects. Moreover, it preserves the assignment of a utility value equal to 0 to the ethically neutral prospect. The result is that it is possible to obtain a zero-one utility representation that is comparable across individuals. In other words, Bradley's proposal shows that it is conceptually possible to compare different people's preferences with respect to strength.

4.2 Objections

Bradley's proposal allows us to measure degrees of preference on a ratio scale on the grounds that the ethically neutral prospect identifies a 'natural' zero in the preference scale. One may ask whether or not this is really the case. Recall that, according to Ramsey, "an atomic proposition p is called ethically neutral if two possible worlds differing only in regard to the truth of p are always of equal value"[17]. At first sight, it does not follow from this quote that p and $\neg p$ bring *no* value to both possible worlds, but only that they bring *equal* value to them. In this sense, the ethically neutral prospect is not a prospect that the agent does not value, but a prospect that the agent values as much as its negation. In terms of preferences, this means that, although the intensity of the agent's preference for the ethically neutral prospect p is equal to the intensity of the agent's preference for its negation $\neg p$, it does not follow that they have null intensity. If this is correct, it is a mistake to say that an ethically neutral prospect is a prospect such that the agent is "disposed neither to make it or its contrary true"[18]. Likewise, it is a mistake to say that the utility measure of an ethically neutral prospect picks out "absence of force"[19]. Rather the ethically neutral prospect is a prospect such that the agent is *equally* disposed to make it or its contrary true. Likewise, the utility measure of an ethically neutral prospect p picks out the *same* force as the utility measure of its contrary $\neg p$, but not absence of force. In this sense, neutrality is not absence of force, but equal force between one prospect and its negation.

This is only a preliminary point. It has no bearing on the issue of comparability. In fact, even if the ethically neutral prospect does not identify a 'natural' zero in the preference scale, comparability can still be proven provided that preferences for ethically neutral prospects play the same causal role in different people's minds. In other words, even if the utility measure of an ethically neutral prospect does not pick out "absence of force", IUCs

[17] See RAMSEY, F. P. [1990], p. 73
[18] See BRADLEY, R. [2007b], p. 9.
[19] See BRADLEY, R. [2007b], p. 9.

can be meaningful provided that the utility measure of an ethically neutral prospect picks out the same causal force across individuals.

As we have seen, the ethically neutral prospect may be different for different individuals. Moreover, each individual can have several ethically neutral prospects. If they have to have the same causal force inter-personally, such prospects must also have the same causal force intra-personally. Minimally, this means that the agent must be indifferent between different ethically neutral prospects. If, for each individual, different ethically neutral prospects are preferred with different strength, that is, if each individual sets neutrality at different levels, then it follows that ethically neutral prospects do not pick out the same causal force across individuals. At the very least, it appears that we need to impose a requirement of rationality in order to constraint the agent's preferences over ethically neutral prospect. According to it, rationality requires an agent to be indifferent between ethically neutral prospects.

However, it is possible to show that this is not an additional constraint on the agent's preferences, but rather a mathematical implication of the definition of ethically neutral prospect[20]. Indeed, for p to be an ethically neutral prospect, it is not sufficient that the agent be indifferent between the prospect p and its opposite $\neg p$. Instead, p is ethically neutral if and only if the agent is indifferent between the prospect $(p \,\&\, q)$ and the prospect $(\neg p \,\&\, q)$, for any q. If p is ethically neutral in this sense, then it is the case that, for all q, $u(p \,\&\, q) = u(\neg p \,\&\, q)$. Moreover, the following identities hold:

$$\begin{aligned} u(q) &= Pr(p \mid q)\, u(p \,\&\, q) + Pr(\neg p \mid q)\, u(\neg p \,\&\, q) \\ &= Pr(p \mid q)\, u(p \,\&\, q) + Pr(\neg p \mid q)\, u(p \,\&\, q), \text{ since } u(p \,\&\, q) = u(\neg p \,\&\, q) \\ &= u(p \,\&\, q). \end{aligned}$$

Now, if we suppose that q is *also* an ethically neutral prospect, it follows by the same token that $u(p) = u(p \,\&\, q)$. Therefore, by transitivity of identities, it is the case that $u(p) = u(p \,\&\, q) = u(q)$. This shows that, if two prospects are ethically neutral, necessarily, the individual prefers them with equal strength and is, thereby, indifferent between them. In turn, this implies that we do not need to impose any requirement of rationality as the possibility that the agent is not indifferent between ethically neutral prospects is simply excluded by the definition of ethical neutrality.

The previous argument takes care of the worry that preferences for ethically neutral prospects may not play the same causal role in the *intra-personal* case. At the same time,

[20] I owe this point and its mathematical proof to my examiners Christian List and Wlodek Rabinowicz.

the argument provides further indirect support for the claim that preferences for ethically neutral prospects play the same causal role in the *inter-personal* case as well. After all, these preferences are identified in terms of functional relations that are common across individuals. Bradley's proposal may work. What about his other suggestion that different individuals possess identical total preferential strength? Bradley invites us to conceive the problem of forming preference strengths as the problem of allocating the total preferential strength available amongst the options in the preference domain. Thus, if the functional relations characterising preferences are the same across individuals and if the set of prospects under consideration is fixed, Bradley argues that the total strength available to each individual is the same.

We can raise the following objection. It is perhaps true, as Bradley suggests, that once the set of alternatives is fixed, the total preferential strength available to each individual is also fixed. However, it is tempting to resist Bradley's further claim that, if everything else is equal, i.e. if the causal relations are the same across individuals, necessarily, the total available preferential strength is the same for all individuals. After all, the preference domain may not be the same for different individuals and it may include a different number of options. Even if the underlying processing mechanisms are the same across individuals, it may be the case that the preference strength assigned to each option is simply added, in one way or another, rather than being divided between these options. According to Bradley, from a functionalist point of view, the possibility of a difference in total strength is simply an illusion. However, as the previous remarks suggest, it is not clear why we should think that this is really the case.

Once again, however, here I am more interested in a different kind of objection, concerning the implications that Bradley's argument has for the epistemological problem of IUCs. At first sight, Bradley's solution is more promising than the first functionalist solution. If the ethically neutral prospect is the object with respect to which different people's preferences play the same causal role, then there are no conceptual obstacles preventing us to have epistemic access to this kind of object. In the same way in which we can epistemically identify both the necessary proposition and the impossible proposition in the case of degrees of belief, so we can epistemically identify the ethically neutral prospect in the case of degrees of preference.

The matter is more complicated when we consider the other point that different individuals are supposed to have in common. Identifying the total desirability available to an individual seems to demand the identification of all the prospects in his preference domain. However, this task runs into problems similar to those that we have seen in the

context of the first functionalist solution. In fact, both when the preference domain contains an infinite number of items and when it contains a finite number of items, the empirical evidence leaves the identification of all the options underdetermined because some preferences may never become manifest. On the other hand, if identifying the total available preference strength does not require us to identify all the other prospects in an individual's preference domain, then we need a principled explanation of how we can have epistemic access to total desirability. The problem is that not only an explanation of this kind is not currently available but that no such explanation seems to exist.

Nevertheless, it is worth noticing one point. If it is possible to epistemically individuate different people's preferences for ethically neutral prospects, then Bradley's proposal achieves at least one interesting result. Indeed, ethically neutral prospects define a zero-line that is common across individuals. Moreover, this is independent from whether ethically neutral prospects define a 'natural' zero or whether they simply define an 'arbitrary' zero. Hence, Bradley shows that we can have scientific knowledge of at least one kind of ICs of preference strength, namely, ICs with respect to an interpersonally significant zero-line, of the kind proposed by List[21]. As we have seen in chapter 1, this means that we can make meaningful ICs of utility levels between individuals with utility, respectively, greater than / equal to / less than the utility value associated to their ethically neutral prospects.

Before concluding, I want to point out that there is yet a stronger objection that one can present against Bradley. As a matter of fact, Bradley's proposal is relative to a background theory of action, namely, the expected utility theory. Under a realist interpretation, such a theory is supposed to represent the relevant causal relations underlying the formation of different individuals' preferences. Thus, the second functionalist argument is based on the assumption that different individuals are isomorphic at the level of the relevant functional relations. Once again, the argument under consideration does not offer any reason to think that it is conceptually possible to compare different people's preferences in terms of strength when this assumption is relaxed. Most importantly, the argument under consideration does not offer any reason to think that it is possible to have epistemic access to the facts about the relevant causal relations. Since both empirical and non-empirical strategies fail to vindicate the assumption that different individuals both respond to the same environmental inputs and form their mental states in the same way, the conclusion is that the second functionalist solution fails to solve the epistemological problem of IUCs.

5. Conclusion

[21] See LIST, C. [2003].

In this chapter, I considered three 'possibility' arguments attempting to solve the problem of IUCs. Their primary goal is to show that different people's preferences are comparable in terms of strength. Their secondary goal is to show that it is possible to have scientific knowledge of, or, at least, scientifically justified beliefs about how different people's preferences compare in terms of strength.

The first argument is based on Broome's work on personal goodness. It claims that, if individual preferences are independent from personal identity, then different people's preferences are interpersonally comparable in terms of strength, provided that each person belongs to at least one overlapping pair and that everyone is connected to everyone else by a chain of overlapping pairs. If this is the case, each person's utility scale is universal and it is thereby possible to have commensurable interpersonal utilities. In this chapter I rejected this argument on the grounds that it is either question-begging or not sufficient to show that preference strengths are interpersonally comparable.

The second and the third argument are based on a functionalist understanding of the nature of preferences. Both arguments claim that it is conceptually possible to identify two points with respect to which different people's preferences play the same causal role. If this is the case, then functionalism allows us to conclude that preferences are interpersonally comparable in terms of strength. The former argument claims that these points are given by the most preferred and by the least preferred option, respectively. The latter argument – offered by Bradley – claims that one point is given by the ethically neutral prospect, while the other is given by the total desirability of all prospects.

If either one of these solutions works, it provides a solution to the metaphysical problem of IUCs. This notwithstanding, I argued that both solutions fail to solve the epistemological problem of IUCs. On the one hand, it is epistemically impossible to identify those cases where different people's preferences play the same causal role. On the other hand, neither solution shows that the relevant causal relations are really the same across individuals.

CHAPTER 5
Transcendental arguments

1. Introduction

The solutions to the problem of IUCs based on an inference to the best explanation type of argument and on a nativist argument fail to show that we can have scientific knowledge of, or, at least, scientifically justified, ICs of preference strength. This raises the suspicion that the difficulty in solving the problem of IUCs stems from the incomparability of preferences with respect to the dimension of strength and not just from epistemological limitations. As a consequence, some authors elaborate 'in principle' solutions to the problem of IUCs. These solutions are based on 'possibility' arguments. Their primary goal is to show that different people's preference strengths are comparable. Their secondary goal is to show that it is possible, in principle but not by means of empirical or pragmatic considerations only, to have scientific knowledge of, or scientifically justified beliefs about, how different people's preferences compare in terms of strength.

In chapter 4, we have examined three 'possibility' arguments. Interestingly, the previous analysis emphasises that, although these arguments differ with respect to their capacity to solve the conceptual problem concerning the interpersonal comparability of preference strengths, they all fail to solve the epistemological problem of IUCs on the same ground, i.e. they do not prove that different individuals respond to the environment and form mental states in the same way. In this chapter, I want to examine another 'possibility' argument, whose goal is precisely to argue in favour of the interpersonal comparability of preference strengths by defending the assumption that the causal relations determining preferences are the same across individuals.

Typically, this argument is formulated in the context of the explanation of the ordinary people's mindreading capacity at the personal level of description, that is, at the level at which persons, as such, think about or interpret other people's mental and overt behaviour. More specifically, it is formulated in the context of mindreading accounts that are based on an *a priori* assumption of rationality. According to these accounts, a necessary and *a priori* condition for ascribing mental states to another agent or system is that such an agent or system be rational, at least to a large extent. This

assumption is the basis for a transcendental solution to the problem of IUCs. In general, the first step consists in anchoring mental state comparability to rationality. The assumption that different individuals are commonly rational – it is argued – implies the assumption that their mental states are interpersonally comparable. The second step consists in defending the necessary role of the background assumption of rationality for the possibility of correctly interpreting other people's behaviour, by means of a transcendental argument. If this defence is successful, it follows that different people's mental states are indeed comparable.

In this chapter I shall pursue two goals. The first is to show that this strategy is unsuccessful. The main objection against it is that it is based on a 'strong' transcendental argument, which invalidly infers a conclusion about the world – i.e. the interpersonal comparability of preferences in terms of strength – from a fact about our psychological reality – the fact that we take correct interpretation to be possible. The second goal is to show that a strategy of the same kind can nonetheless achieve results of anti-sceptical significance on the basis of a more 'modest' transcendental argument. Crucially, this argument avoids inferring a conclusion about the world and validly infers a conclusion about our psychological reality – i.e. that we *take* preference strengths to be comparable in terms of strength. I shall argue that, if we combine such a 'modest' transcendental argument with coherentism about justification, it is possible to show, at least, that ICs of preference strength can be (scientifically) justified.

I shall proceed as follows. In section 2, I shall offer some very general remarks about the role of the background assumption of rationality for mindreading and ICs of preference strength. In section 3, I shall illustrate Davidson's argument in favour of comparability, which allegedly provides the paradigm of a transcendental strategy to solve the problem of IUCs. In section 4, I shall reconstruct the general form of his 'strong' transcendental argument. In Davidson's framework, the background assumption of rationality is part of a broader set of conditions, which are claimed to be necessary for the possibility of correct interpretation. I shall try to characterise these conditions in more detail in section 5. In section 6, I shall present some objections against transcendental arguments of a strong form. These objections cast more than one doubt on the transcendental strategy for solving the problem of IUCs. In section 7, I shall present and try to defend a more 'modest' transcendental argument. I shall discuss its implications for the problem of IUCs in section 8. Finally, I shall summarise my findings in the conclusion.

2. The background assumption of rationality

In chapter 3, we have considered two explanations of the ordinary people's mindreading capacity, i.e. Simulation Theory (ST) and Theory Theory (TT), at the sub-personal level of description. However, according to some authors, ST and TT should be conceived as hypotheses formulated at the personal level of description, that is, at the level at which persons, as such, think about or interpret other people's mental and overt behaviour[1]. The move from the sub-personal to the personal level of analysis is not particularly helpful for the purpose of finding a positive solution to the problem of IUCs if one assumes that the epistemological status of both ST and TT is empirical. Yet, this is not the only possibility. The move to the personal level is often associated with an understanding of the epistemological status of both ST and TT as *a priori*. This stance is typically adopted by versions of both ST and TT that are based on a background assumption of rationality. Let us consider these accounts in more detail.

Jean Heal is the main exponent of a ST account of mindreading based on a background assumption of rationality[2]. As we have seen in chapter 3, mental simulation involves replicating another individual's mental life in specific circumstances by imagining being subject to the same, or relevantly similar, circumstances. At the sub-personal level of description, simulation is conceived in terms of information-processing mechanisms. By contrast, at the personal level of description, simulation is conceived as the activity of "thinking about the same subject matter". According to Heal, in order to replicate another agent's mental life, the simulator must be capable of thinking about the same content of the simulated agent's mental states, on the one hand; and of having the same attitudes as the simulated agent, on the other hand. As such, simulation is a form of "co-cognition"[3]. According to Heal, simulation so conceived involves construing the interaction between the agent's mental states as rational, at least when simulation concerns certain subject matters. This means that the replication of another individual's mental life is based on the assumption that his mind's working satisfies certain criteria of rationality. In the same vein, TT accounts of mental ascription that are based on a background assumption of rationality suggest that the folks' mindreading capacity is based on the possession of a 'theory of mind' that represents other people's mental life as rational.

---

[1] See HEAL, J. [1994], [1998a,b], [2000] and GORDON, R. [1992].
[2] See, in particular, HEAL, J. [1998a,b].
[3] See HEAL, J. [1998a], pp. 483-484.

It is controversial whether the background assumption of rationality turns these mindreading accounts into rationality theory (RT) accounts or whether it is compatible with the key features of TT and ST. The basic idea underlying RT is that mental ascription consists in rationalising an agent's behaviour, that is, in ascribing to him the mental states that it would be rational for him to have. The main exponents of a RT approach are Davidson and Dennett[4]. RT accounts seem to contrast with, rather than instantiate, both ST and TT accounts for at least two reasons. The first is that rationality is usually construed as a normative notion and its normative character does not fit well with the causal approaches advocated by ST and TT. The second is that RT seems to imply that the meaning of mental terms is given by the set of normative requirements that the criteria of rationality set on mental states. This contrasts with both analytic functionalism, according to which the meaning of mental states is given by the set of causal laws in which those mental states are embedded[5]; and with experientialism, according to which the meaning of mental states is given by the more or less conscious experience that the subject has of them.

Although these reasons are often thought to be sufficient to discard both RT accounts and the background assumption of rationality, recent works on rationality suggest the possibility of a reconciliation between RT, on the one hand, and TT and ST, on the other hand. Many authors have suggested that we should not understand rationality as a normative notion, but rather as a quasi-normative notion. Rationality establishes the standards for the proper working of the mind-system[6]. This means that an individual is rational to the extent that his mind-system works properly. This is not equivalent to saying that he is rational if he does what he ought to do, from a normative point of view. Any system – even a causal one – works properly with respect to certain specific criteria. The fact that proper working is assessed with respect to these standards is compatible with the fact that the elements constituting the system interact in a causal way. If we understand rationality in this way, then rationality-based accounts suggest that an individual's mind works properly if and only if the interaction between his mental states conforms to certain standards, or requirements, of rationality. In this sense, RT accounts may as well be instances of both ST and TT accounts.

How can ST and TT explain the folks' ICs of preference strength, at the personal level of description, under a background assumption of rationality? Let us consider TT first. The first step concerns third-person mental ascription. The judge observes the

[4] See DAVIDSON, D. [1984] and DENNETT, D. [1987].
[5] See GOLDMAN, A. [2002] and [2006], chapter 3.
[6] See SCANLON, T. [1998] and [forthcoming], KOLODNY, N. [2005], YASGUR, S. [2008]

relevant external events (i.e. instances of the input-types and output-types that are included in the definition of preferences) and infers both the other relevant mental states (i.e. tokens of the mental state-types that are included in the definition of preferences) and the relevant preferences, by reference to the causal relations postulated by the ToM that he – more or less tacitly – possesses. The background assumption of rationality constraints the subject's inferences by assuming that the interaction of people's mental states obeys, *ceteris paribus*, certain standards of rationality.

The second step concerns first-person mental ascription. It is unclear whether or not, and how, the background assumption of rationality operates in the first-person case. According to the orthodox view, it constraints the subject's inferences by postulating that the interaction of his own mental states obeys, *ceteris paribus*, certain standards of rationality, in exactly the same way as it does in third-personal mental ascription. On the basis of the ToM in his possession, the subject ascribes preferences with a specific content and strength both to himself and to the other individual. Finally, in the last stage, the subject compares the intensity of his preferences with the intensity of the other individuals' preferences.

Let us now consider ST. Third-person mental ascription involves, first, observing the external events that are relevant for the individuation of the simulated agent's initial mental states and, then, replicating the interaction of the simulated agent's mental states. Once again, the background assumption of rationality constrains simulation by holding that the interaction of people's mental states obeys, *ceteris paribus*, certain standards of rationality. On the basis of simulation, the interpreter ascribes preferences with a specific content and strength both to himself and to the other individual. In the last step, he compares the intensity of his own preferences with the intensity of the simulated agent's preferences.

3. Davidson's strategy

In this section I shall illustrate Davidson's position. The analysis of his argument is particularly instructive in order to highlight some of the features of a transcendental solution to the problem of IUCs. Indeed, Davidson is one of the leading figures defending the thesis that the very possibility of correctly interpreting other people's behaviour implies that preferences are interpersonally comparable in terms of strength. Although his overall project is a paradigm of a RT approach to mindreading, his

argument can be applied, with few modifications, both to TT and ST rationality-based accounts.

The starting point of Davidson's investigation is the characterisation of the theoretical framework in which the problem of IUCs arises as a sequence of three steps, dealing with[7]:

(1) the determination of individual preferences and their representation through a (family of) utility function(s);
(2) the interpersonal comparison of utilities;
(3) the formulation of the judgment of interest.

In this framework, the problem of IUCs arises at the second stage. However, according to Davidson, the difficulties in finding a positive solution suggest that this theoretical framework *is* the source of the problem and should be rejected. There are at least two possibilities. The first consists in denying that IUCs are factual statements and maintaining that they are (part of) either normative or evaluative judgments. This means that the second and the third step are much more interdependent and difficult to distinguish than suggested by the 'standard picture', up to the point where they mesh together[8]. By contrast, Davidson emphasises the interdependence existing between the first and the second stage in the 'standard picture'. Without rejecting the connection between IUCs and normative judgments, Davidson claims that the very attribution of preferences to another individual involves an interpersonal comparison between the interpreter's and the agent's preferences. As a consequence, according to Davidson, the basis for IUCs is provided by the principle that guides the ascription of preferences and other mental states to other individuals, namely, the Principle of Charity (PoC, for short). This is how Davidson expresses the point: "I think interpersonal comparisons have a basis in the following sense: in the process of attributing propositional attitudes like beliefs, desires, and preferences to others, interpersonal comparisons are necessarily made". In the case of evaluative attitudes like preferences, this does not mean that "the attributer consciously or unconsciously makes a comparison, but that in the process of attribution the attributer necessarily uses his own values in a way that provides a basis for comparison; a comparison is implied in the attribution"[9].

[7] See DAVIDSON, D. [1986], reprinted in DAVIDSON, D. [2004].
[8] Cfr. ROBBINS, L. [1932], JEFFREY, R. [1971], SCHICK, F. [1971], and HAMMOND, P. [1991].
[9] DAVIDSON, D. [2004], p. 67.

Let us try to be more specific about the nature of this comparison. In a Davidsonian framework, interpretation has two features. Like scientific inquiries in other domains, it starts by characterising the object to be explained as a system, where theoretical entities interact within a certain structure. Unlike scientific inquiries in other domains, it proceeds by assuming that both the explaining agent and the system to be explained, i.e. the interpreter and the interpreted agent, are systems of the same kind. This second feature has two crucial implications for interpretation. On the one hand, it implies that the interpreter's and the interpreted agent's mind-systems share the same theoretical entities, i.e. the same mental state-types, and the same structure, i.e. the way mental states interact. On the other hand, it suggests that the default interpretive procedure consists in projecting the interpreter's standards for individuating the content, the type, the properties and the structure of mental states into the interpreted agent. In particular, as far as the issue of content individuation is concerned, the interpreter projects his own standards of truth, in the case of doxastic attitudes, and his own standards of value, in the case of evaluative attitudes. On the other hand, as far as the issue of structure is concerned, the interpreter projects his standards of rationality in order to establish a relation between different mental states.

In the light of these features, interpretation involves a comparison between the interpreter's and the interpreted agent's mental states, in certain relevant dimensions, from the start. First, the interpreter assumes that the observed agent's mental states play a role similar to the role played by his own mental states. This implies that the interpreter assumes that the agent's preferences, desires and beliefs possess a dimension of strength. Second, the interpreter assumes that, *ceteris paribus*, the observed agent believes, values, desires *p* if and only if he believes, values, desires *p*. This means that the interpreter assumes that the agent has, *ceteris paribus*, the same beliefs, values, desires, etc. that he has. Third, the interpreter assumes that the agent's mental states obey standards of rationality similar to his. According to Davidson, this means assuming that the agent's preferences are consistent in a specified way.

The projection establishes a comparison in the sense that it sanctions the ascription of similarities and differences between the interpreter's and the agent's mental states across the relevant dimensions. For instance, in the case of preferences, the interpreter can attribute irrational preferences to the agent, when they do not satisfy the standards of rationality recognised by the interpreter. Moreover, the interpreter can relate the agent's preferences to objects different from those that provide the content of his preferences. Finally he can ascribe preferences to the agent towards the same objects but

with different intensity. According to Davidson, differences can be tolerated up to a certain extent, that is, up to the point where the agent's behaviour remains intelligible. However, this is enough to establish an interpersonal comparison between the interpreter's and the agent's mental states.

It is worth emphasising that, although the interpreter's projection implies a comparison, "[it] does not amount to a comparative judgment", but rather "[i]t establishes a basis for comparative judgments"[10]. In other words, the ascription of mental states to another agent does not involve making ICs. That is, it does not end up with a conscious or unconscious comparative judgment. Instead, by establishing a comparison between the interpreter's and the agent's mental states, it provides the ground for the explicit comparative judgment, which is made on the basis of such an ascription at a later stage.

According to Davidson, the PoC is the principle recommending the interpreter's projection as the starting point for the interpretation of other people's behaviour. Davidson argues that the PoC is required in order to optimize agreement between the interpreter and the agent and, thereby, to make understanding possible. According to him, this is not "an empirical claim or an assumption for the sake of science". The PoC is neither discovered, nor normatively chosen, but it is an *a priori* principle[11] and "a necessary condition of correct interpretation"[12].

We may read Davidson as offering an *a priori* argument for thinking that preferences are interpersonally comparable in terms of strength[13]. The interpreter's projection establishes a comparison between the interpreter's and the agent's preferences in certain relevant dimensions. Since such a projection endows the agent's preferences with a dimension of strength, it crucially establishes a comparison between the interpreter's and the agent's preference strengths. The projection is required by the PoC. Crucially, it is not just the case that the interpreter *takes* the agent's preferences to be interpersonally comparable in terms of strength, as a result of the projection. Rather, as the PoC is an *a priori* principle, which is necessary for the *correct* interpretation of other people's behaviour, Davidson offers an *a priori* reason to conclude that it is also the case that the interpreter's and the agent's preferences really *are* interpersonally comparable in terms of strength.

---

[10] DAVIDSON, D. [2004], p. 71.
[11] DAVIDSON, D. [2004], p. 73.
[12] DAVIDSON, D. [2004], p. 72.
[13] For instance, WEINTRAUB, R. [1998] reads Davidson as offering an argument of this kind and rejects it.

4. A 'strong' transcendental argument

The success of Davidson's strategy crucially depends on the soundness of two assumptions. The first is that correct interpretation of other people's behaviour is possible. The second is that the PoC is necessary for correct interpretation to be possible. Thus, the argument takes the form of a 'strong' transcendental argument. The general idea behind a 'strong' transcendental argument is that we can reason from the fact that we possess a cognitive ability, first, to the individuation of the conditions that support our ability and, second, to the truth of those conditions. The argument goes as follows:

(1) We posses the ability to Φ;
(2) We could not Φ without X;
(3) Hence, X is true.

In the case under consideration, X is the 'PoC' while Φ is not just 'interpreting each other, but, more strongly, '*correctly* interpreting each other'. The argument becomes:

(1) We posses the ability to correctly interpret each other;
(2) We could not interpret each other correctly without the PoC;
(3) Hence, the PoC is true.

This argument can be used to defend the thesis that preference strengths are interpersonally comparable in virtue of the following additional assumption:

(3.1) The PoC implies that the interpreter's and the interpreted agent's preferences are comparable in terms of strength;

If (3) is sound, that is, if the PoC is true, the conclusion is that the interpreter's and the interpreted agent's preferences are indeed comparable in terms of strength.

This conclusion completes the first part of a 'possibility' argument. What about the second part, that is, the part aimed at showing that it is possible to have scientific knowledge of, or scientifically justified beliefs about, how different people's preferences compare in terms of strength? Indeed, the 'strong' transcendental argument can offer an indirect solution also to the epistemological problem of IUCs. Recall that

the arguments examined in the previous chapter are subject to two objections. The main objection is that they do not prove that the relevant causal relations are the same across individuals. The second objection is that they require an epistemically impossible individuation of all the prospects in each individual's preference domain. The 'strong' transcendental argument is invulnerable to either objection.

Let us consider the first. The 'strong' transcendental argument grounds the interpersonal comparability of different people's preferences on the fact that they possess both mental states with identical properties and mind-systems with identical structure. In other words, on the one hand, the 'strong' transcendental argument grounds interpersonal comparability on the assumption that different people respond to the environmental inputs and form their preferences in the same way and, on the other hand, it shows that this assumption is sound. This refuses the first objection.

Let us consider the second. As seen above, although interpretation does not involve making any explicit or implicit IC judgments, it establishes a basis for the IC judgments made at a later stage. Such judgments are based on the ascription of preferences with a specific content and strength. In turn, the content of preferences is individuated on the basis of the evidence available in specific circumstances. However, this does not presuppose individuating all the prospects included in the individual's preference domain. Therefore, the 'strong' transcendental argument is not subject to the second objection either.

If preference strengths are interpersonally comparable from the start, we can argue that, for any two individuals, if all the relevant evidence is the same, we are justified in assuming that their preference strengths are also the same. Within an evidentialist framework, justification comes from the fact that ICs of preference strength are determined on the basis of both non-empirical evidence concerning the relevant causal relations and empirical evidence concerning the object and the properties of the individual's preferences. Within a reliabilist framework, justification comes from the fact that ICs of preference strength are reliably determined on the basis of a non-empirical principle, i.e. the PoC, in combination with the relevant empirical evidence[14]. Finally, if their preference strengths are the same, the individuals have the same utility functions. For any two individuals, such an isomorphic situation provides the benchmark for the ascription of justifiable differences in their utility values, when the

[14] On the one hand, being a necessary condition for correct interpretation, the PoC is reliable for the purpose of interpreting other people's behaviour. On the other hand, the PoC is also reliable for the purpose of making ICs of preference strength, because correct interpretation is necessarily achieved by establishing a comparison between the interpreter's and the interpreted agent's mental states,.

relevant empirical evidence is not the same. The conclusion is that, if sound, the 'strong' transcendental argument shows that it is possible to have scientifically justified IUCs.

5. The Principle of Charity and The Principle of Similarity

Given the centrality of the PoC, the first step for assessing the transcendental strategy consists in providing a more precise definition of the principle itself. This task is surprisingly difficult. To begin with, none of the authors that rely on the PoC has offered any explicit formulation of it. Moreover, given that the PoC is a set of different conditions, there may be competing definitions depending both on what conditions are included and on how they are defined. Here I shall consider some alternative characterisations.

A caveat is in order. In order to keep my analysis applicable to both TT and ST accounts, I shall conceive rationality as the proper working of an individual's mind. Thus, the conditions more directly connected to rationality, in each of the following characterisations, should be read as specifying what it means for an individual's mind to work properly. Broadly speaking, there are two competing approaches to rationality. The classic approach develops the idea of rationality as consistency. Accordingly, an individual's mind works properly if and only if the interaction between his mental states obeys the requirements of logic. A more recent approach connects rationality to reasons. Accordingly, an individual's mind works properly if and only if the interaction between his mental states constitutes an appropriate response to reasons[15].

Let us start by characterising the PoC in accordance with the first approach to rationality. From his overall work on radical interpretation, we can formulate Davidson's version of the PoC as the combination of the following principles.

$\underline{\text{PoC}_1}$

1. *The principle of correspondence*: for any interpreter *i* and agent *j*, *ceteris paribus*, if *i* and *j* were to be subject to the same environmental causes, they would have the same beliefs, values, desires, etc.
2. *The principle of coherence*: for any agent *j*, *ceteris paribus*, *j* has, for the most part, logically consistent mental states.

[15] See YASGUR, S. [2008]. On rationality as appropriate response to reasons, see also RAZ, J. [1999].

3. *The principle of truth*: for any agent *j*, *ceteris paribus*, *j*'s beliefs are, for the most part, true.
4. *The principle of good*: for any agent *j*, *ceteris paribus*, *j*, for the most part, desires, values, prefers what is desirable, valuable, preferable[16].

The principle of correspondence is very close to Harsanyi's 'similarity postulate' Like Harsanyi, Davidson believes that "similar causes beget similar evaluations in interpreter and interpreted"[17]. The first difference is that Davidson extends this idea to doxastic attitudes as well. The second difference concerns the justificatory strategy. Harsanyi attempts to justify his 'similarity postulate' by appealing to pragmatic considerations. Instead, Davidson argues *a priori* that the principle of correspondence, as part of the PoC, is necessary for correct interpretation[18].

The principle of coherence claims that the agent's attitudes are, for the most part, logically consistent. According to Davidson, this is a requirement of minimal rationality. Since the principle of correspondence specifically applies to particular beliefs and evaluations, that is, to beliefs and evaluations prompted by the surrounding environment, the principle of coherence provides a bridge between particular mental states and more abstract and theoretical ones. More precisely, it allows the interpreter to tentatively infer the latter from the former on the basis of the rules of inference of logic.

Finally, the principle of truth claims that most of the agent's beliefs are true. Similarly, the principle of good claims that most of the agent's evaluations are correct. Davidson insists particularly on the principle of truth, whereas he never explicitly mentions the principle of good as part of the PoC. However, such a principle is the counterpart of the principle of truth for evaluative attitudes. Therefore, as a matter of internal consistency, it seems to be a crucial part of the PoC[19].

---

[16] The *ceteris paribus* clause takes into account contingent errors made by either the interpreter or the interpreted agent. The 'for the most part' clause takes into account errors due to the state of knowledge of the communities of which the interpreter and the interpreted agent are part.

[17] See DAVIDSON, D. [2004], p. 72.

[18] Perhaps, there is another difference. For the success of his overall project, Davidson may need a stronger formulation of the principle of correspondence, according to which, for any interpreter *i* and agent *j*, *ceteris paribus*, *i* and *j* believe, value, desire *p* if and only if they are subject to the same environmental causes. This is because the role of the Principle of Correspondence consists in improving a weaker Principle of Agreement, according to which, for any interpreter *i* and agent *j*, *ceteris paribus*, *j* believes, values, desires *p* if and only if *i* believes, values, desires *p*. For the success of Davidson's overall theory of radical interpretation, Agreement is not enough because the interpreter may come to agree with the agent in a spurious way, that is, on the basis of either completely different reasons or mistakes or deviant causal chains, etc. By contrast, the stronger version of the principle of correspondence would guarantee that exposure to the same environmental conditions is, *ceteris paribus*, a necessary and sufficient condition for agreement. Cfr. LEPORE, E. and K., LUDWIG [2005], chapter 12.

[19] Dennett's version of the PoC differs only slightly from Davidson's and includes the following principles:

Let us consider now how the PoC can be characterised when rationality is defined in terms of the individual's appropriate response to reasons. Clearly, the definition of rationality depends both on how the notion of reasons is conceived and on how the idea of 'appropriate response' is spelt out. As far as the former issue is concerned, I shall adopt, without arguing for it, an externalist understanding of reasons, according to which reasons are facts[20]. If we adopt such a view of reasons, we may understand external inputs as providing an individual with reasons for specific mental states and/or behaviours.

As far as the latter issue is concerned, I shall maintain, without arguing for it, that an individual responds appropriately to reasons if and only if he recognises the relevant reasons and forms his mental states in accordance with the reasons that he recognises. Following Yasgur, we can characterise the notion of appropriate response to reasons in terms of the notions of reasonableness and rationality[21]. The problem consists in specifying these notions more precisely. We can say that an individual is reasonable if and only if, *ceteris paribus*, he recognises (at least) a reason for believing, valuing, desiring *p* (and only provided that such a reason exists)[22]. Moreover, we can say that an individual is rational if and only if he forms his mental states in accordance with the reasons that he recognises.

With these notions in mind, we can try to formulate an alternative version of the PoC.

PoC$_2$

1. *The principle of reasonableness*: for any interpreter *i* and agent *j*, *ceteris paribus*, if *i* and *j* were to be subject to the same environmental circumstances,

---

1. *The principle of coherence*: for any agent *j*, *ceteris paribus*, *j* has, for the most part, logically consistent mental states.
2. *The principle of closure*: for any agent *j*, *ceteris paribus*, *j*'s belief-set is closed under entailment.
3. *The principle of truth*: for any agent *j*, *ceteris paribus*, *j*'s beliefs are mostly true.
4. *The principle of good*: for any agent *j*, *ceteris paribus*, *j*, for the most part, desires, values, prefers what is desirable, valuable, preferable.

This characterisation follows closely, although it is not identical to, Fodor's and Lepore's analysis. See FODOR, J. and E., LEPORE [1992].

[20] See RAZ, J. [1975] and [1999]. The externalist view contrasts with an internalist understanding of reasons, according to which reasons are mental states. See WILLIAMS, B. [1981].

[21] This way of defining 'rationality', 'reasonableness' and the corresponding principles is adapted from Yasgur analysis of reasons and rationality, in YASGUR, S. [2008].

[22] Reasonableness comes in degrees. Thus, in order to be reasonable, it is not necessary that the individual recognises *all* the reasons that there are for believing, valuing, desiring *p*.

then they would recognise the same reason(s) for believing, valuing, desiring p[23];

2. *The principle of rationality*: for any interpreter i and agent j, *ceteris paribus*, if i and j were to recognise the same facts as reasons for beliefs, values, desires, etc, then they would form the same beliefs, values, desires, etc[24].

The formulation of the principle of reasonableness is too strong. In particular, it leaves unexplained the case in which the interpreter ascribes mental states to a mistaken agent. Indeed, in some cases, the interpreted agent recognises as reasons facts that are not truly reasons. His mistakes may refer to either non-evaluative or to evaluative properties of the world. The interpreted agent's mental states operate as if they were based on truly reason-giving facts, although they are not. The relevant distinction is the one between operative reasons and normative reasons. Operative reasons are constituted by facts that the agent recognises as reason-giving, although they are not. By contrast, normative reasons are constituted by facts that are truly reason-giving[25]. For the purpose of interpretation, the interpreter must be able to identify the reasons that the agent recognises, whether or not they are truly reasons (i.e. normative reasons) or simply facts that the agent mistakenly recognises as reasons (i.e. operative reasons).

It is questionable whether or not the PoC, with its strong emphasis on the principle of truth and the principle of good, is suitable to capture this feature of interpretation. As a consequence, several authors have suggested grounding interpretation on a different principle, namely, the Principle of Similarity (PoS)[26]. The PoS does not imply either the principle of truth or the principle of good. Rather, it greatly emphasises the similarity between interpreter and interpreted agent, both with respect to their response to external inputs and with respect to the interaction between their mental states. One way to formulate the PoS, with respect to reasons and rationality, is the following[27].

---

[23] It is worth noticing that the principle of reasonableness implies both the principle of truth and the principle of good, if combined with the assumption that the agent recognises *most of* the reasons that there are for believing, valuing, desiring p. Indeed, if, *ceteris paribus*, an agent recognises a fact as a reason only if that fact is reason-giving, then, if he recognises most reasons, it follows that he has, *ceteris paribus*, mostly true beliefs about, and mostly correct evaluations of, the surrounding environment. In turn, this means that a particularly high reasonableness degree entails both the principle of truth and the principle of good.

[24] It is worth noticing that the principle of rationality does *not* imply the principle of coherence. The exclusion of coherence is due to the fact that, according to reason-based accounts of rationality, coherence is thought to be simply a by-product of the recognition of reasons, rather than a requirement of rationality itself (see, for instance, KOLODNY, N. [2005]).

[25] See RAZ, J. [1975] and [1999].

[26] The proposal to substitute the Principle of Charity with the Principle of Similarity originally comes from GRANDY, R. [1973].

[27] Notice that the Principle of Similarity arises as an attempt to block the objections raised against the

$\underline{\text{PoS}_1}$

1. *The principle of recognition*: for any interpreter *i* and agent *j*, *ceteris paribus*, if *i* and *j* were to be subject to the same circumstances, they would recognise the same facts as providing (either believed or normative) reasons for beliefs, values, desires, etc.
2. *The principle of rationality*: for any interpreter *i* and agent *j*, *ceteris paribus*, if *i* and *j* were to recognise the same facts as (either believed or normative) reasons for beliefs, values, desires, etc, then they would form the same beliefs, values, desires, etc.

6. Objections

The next step in our assessment consists in examining whether or not the premises of the transcendental argument are sound. Let us consider the first premise, which states that we posses the ability to correctly interpret each other. An interpretation is correct if and only if the explanation of the interpreted agent's (mental or overt) behaviour in terms of his mental states is correct. Presumably, this means that an interpretation is correct if and only if it is true both that the agent had those mental states and that they led him to the targeted behaviour. There is no doubt that we possess the ability of engaging in an activity that falls within the concept of interpretation. However, the soundness of the first premise presupposes the possibility to assess whether or not the interpretation is correct independently from the appreciation of the conditions that are supposedly necessary for its correctness, that is, independently from the second premise. This is problematic. One of the salient features of transcendental arguments is that the first premise states certain facts that even the sceptic ought to accept. Typically, these are facts about thought or experience, whose truth the sceptic too may grant. However, in the current case, the soundness of the first premise is not immediately acceptable by, or obvious to, the sceptic. Perhaps, the sceptic might concede that the first premise is sound if we conceive correctness in the weaker sense of intersubjective agreement. In this sense, an interpretation is correct if and only if it is intersubjectively agreed upon by, or if it optimises agreement between, interpreter and interpreted agent. However, even if we grant the empirical point that people very often reach

---

Principle of Charity. It is possible to do this by marginally modifying the previous formulation of the latter principle as shown below.

intersubjective agreement in interpretation, this is no proof of the fact that the intersubjectively agreed interpretation is correct in the stronger sense seen above. In other words, the fact that people agree that an interpretation is true is no proof that such an interpretation is really true. As we shall see below, this poses an insurmountable threat to the transcendental strategy under consideration.

Let us consider the second premise. According to it, we could not interpret each other correctly without the PoC. Since Davison's argument offers a paradigmatic case of transcendental analysis applied to interpretation, the literature devotes large attention both to his overall framework and to his version of the PoC, i.e. PoC_1, in order to assess the second premise. One objection concerns the relationship between radical interpretation and the actual practice of interpretation. Davidson explicitly claims that he is not concerned with how people actually interpret each other, but only with a highly idealised procedure, which is supposed to uncover certain key facts about meaning. Arguably, the PoC is true in the context of radical interpretation. However, if radical interpretation bears only little resemblance to actual interpretation, the transcendental strategy may not work when applied to the actual case. This objection is not too damaging. The only consequence is that we cannot show that the second premise is sound by appealing to Davidson's idealised framework. The question of whether or not the PoC is required for the correctness of actual interpretation remains open.

Yet, most of the literature remains sceptical. The main target of criticism is Davidson's principle of truth[28]. For instance, McGinn defends the conceptual possibility that correct interpretation may involve ascribing mostly false beliefs to an agent. McGinn's motivation takes the form of a *reductio*, based on the idea that "if Davidson were right about the inherently charitable nature of interpretation, then we could dismiss certain kinds of traditional scepticism; but it is absurd to suppose that scepticism could be dismissed in this oblique and roundabout way"[29]. On a similar vein, Stroud argues that, if the principle of truth were true, we could dismiss scepticism as a logical impossibility. However, scepticism is a logical possibility. Therefore, the principle of truth must be false[30]. Instead, Lepore and Ludwig contend that Davidson's defence of the principle of truth can, at best, show that, necessarily, most of our general beliefs are true. However, it does not show that any of our particular beliefs, that is, those beliefs

[28] Clearly, similar attacks could be raised against the principle of good. The only reason why they are not made is that Davidson never mentions such a principle explicitly.
[29] See McGinn, C. [2002], p 183.
[30] See Stroud, B. [2000], pp 177-202.

prompted by the surrounding environment, are true. As a consequence, *pace* Davidson, correct interpretation may as well end up with the ascription of mostly false particular beliefs to the interpreted agent[31].

These objections apply to all the versions of the PoC that include, or imply, the principle of truth, namely, PoC_1 and PoC_2[32]. However, they do not affect the PoS. According to the PoS, both the interpreter and the interpreted agent respond to the environment and form mental states in the same way, when subject to the same circumstances. The PoS does not exclude the possibility that the interpreter's and the interpreted agent's responses may lead them to having mostly false beliefs about the surrounding environment. That is, it does not exclude the possibility that the interpreter and the interpreted agent are brains-in-a-vat. Their first-order beliefs may as well be mostly false. However, if the interpreter and the interpreted agent respond to the environment and form mental states in the same way, then PoS_1 may nonetheless lead to true second-order beliefs, that is, true beliefs about (another individual's) beliefs, and, more generally, true beliefs about (another individual's) mental states. In other words, PoS_1 may nonetheless sanction correct interpretations.

Quite independently from whether or not the second premise is defensible, however, there is a decisive reason to hold that the transcendental strategy is not capable of delivering the wanted result. The objection was firstly presented by Stroud and decisively challenges the possibility that any 'strong' transcendental argument can achieve radical anti-sceptical results. Recall the general structure of transcendental arguments. The first premise claims that we possess the ability to Φ, where Φ is typically a fact about thought or experience. However, the sceptic may object that, since Φ is a fact about psychological reality, it remains within the realm of things that we take, or believe, to be the case. Thereby, the sceptic may resist the conclusion that X is true and, instead, simply accept the conclusion that we take, or believe, X to be true. Stroud's main idea is that one cannot start from a fact about psychological reality and arrive at a conclusion about how the world is, independently of our mind. More precisely, it is possible to reach a conclusion about how the world is from a premise about psychological reality only if one is willing to embrace a controversial idealist position, according to which how the world is depends on certain features of our mind. Otherwise, the conclusion must be another claim about psychological reality, namely,

[31] See LEPORE, E. and K. LUDWIG, [2005], pp. 200-202.
[32] These objections apply, *mutatis mutandis*, to the principle of good.

that we take, or believe, certain things to be true or that certain things seem to us to be true. The upshot is that no 'strong' transcendental argument can be successful.

Stroud's objection is particularly clear when applied to the case under consideration. The difficulty in finding an independent proof that correct interpretation is possible suggests interpreting the first premise of the transcendental argument as saying only that we take, or believe, correct interpretation to be possible, on the grounds that we frequently reach intersubjective agreement about the interpretation of people's behaviour. However, once we adopt this reading, the only conclusion that we can draw is a conclusion about our psychological reality. For instance, consider the principle of truth. Suppose Davidson is correct insofar as he maintains that optimal agreement requires the interpreter to assume that most of the agent's beliefs are true. This shows that, for any interpreter *i* and interpreted agent *j*, intersubjectively agreed interpretation requires *i* to *represent j*'s beliefs as mostly true. As Stroud claims, however, "the admitted necessity of finding largely true beliefs among the beliefs one attributes does not imply that the beliefs one attributes are in fact largely true"[33]. That is, even if the interpreter is bound to take, or believe, that the agent has mostly true beliefs in order to ascribe mental states to him, it does not follow that most of the agent's beliefs are really true. More generally, a 'strong' transcendental argument can show only that we take, or believe, either the PoC or the PoS to be true; not that either the PoC or the PoS is true.

7. A 'modest' transcendental argument

Although 'strong' transcendental arguments fail, Stroud argues, in general, that a more 'modest' transcendental argument may still help us reaching conclusions of anti-sceptical significance. For the purpose of this thesis, it is worth considering two questions. First, is a 'modest' transcendental argument sound, when applied to the case under consideration? Second, what are the implications for the problem of IUCs? I shall attempt to answer the first question in this section and the second question in the next section.

In general, a modest transcendental argument has the following form.

(1') We take, or believe, ourselves to posses the ability to Φ;

(2') We could not take, or believe, ourselves to posses the ability to Φ, without taking, or believing, X to be true

[33] See STROUD, B. [2000], p. 186.

(3') Hence, we take, or believe, *X* to be true.

In the case under consideration, the argument becomes:

(1') We take, or believe, ourselves to posses the ability to correctly interpret each other;
(2') We could not take, or believe, ourselves to posses the ability to interpret each other correctly, without taking, or believing, either the PoC or the PoS to be true;
(3') Hence, we take, or believe, that either the PoC or the PoS is true.

Typically, we take, or believe, ourselves to posses the ability to correctly interpret each other on the ground that we frequently reach intersubjective agreement *of the right kind* in the interpretation of other people's behaviour. The qualification is necessary, because what makes us believe that correct interpretation is possible is not just any sort of intersubjective agreement, but justifiable intersubjective agreement, that is, one reached on the basis of appropriate evidence or methods. If this is the case, premise (2') can be understood as saying that we could not reach *valid* intersubjective agreement in interpretation without taking either the PoC or the PoS to be true. In other words, since what we take, or believe, to be correct interpretation is interpretation that is validly intersubjectively agreed upon, we can say that our taking either the PoC or the PoS to be true is necessary for such an intersubjective agreement to be possible. By substituting appropriately, this argument has the following implication for the problem of IUCs:

(3.1') If we take, or believe, either the PoC or the PoS to be true, then we take, or believe, the interpreter's and the interpreted agent's preferences to be comparable in terms of strength.

I shall take for granted the first premise by conceding that valid intersubjective agreement in interpretation is frequent, as I have done in chapter 3. The interesting premise is the second. Clearly, its assessment depends on whether one adopts the PoC or the PoS as the central interpretive principle. In a recent paper, Stroud has tried to defend premise (2') by considering PoC_1 and, essentially, by rehearsing Davidson's position in the modified context of a 'modest' transcendental argument. However, we have seen above that there are independent reasons to prefer the PoS to the PoC. Therefore, in this section, I shall examine premise (2') by considering PoS_1.

It is tempting to argue for (3') by means of an inference to the best explanation type of argument. Accordingly, the fact that we take, or believe, PoS_1 to be true is what best explains why we reach valid intersubjective agreement in interpretation. However, an inference to the best explanation argument allows for the possibility that, although we actually take, or believe, PoS_1 to be true, we could nonetheless reach valid intersubjective agreement in a different way. By contrast, (2') claims that taking, or believing, PoS_1 to be true is the only way to reach intersubjective agreement. This means denying that the conjunction of the following claims can be true: (i) for any three individuals *k*, *h* and *i*, *k* and *h* reach valid intersubjective agreement about the interpretation of *i*'s behaviour; (ii) for any three individuals *k*, *h* and *i*, it is not the case that *k* and *h* take, or believe, that, were they subject to the same circumstances as *i*, they would recognise the same facts as providing (either believed or normative) reasons and would accordingly form the same attitudes as *i*.

Let us see whether or not a possible world where both (i) and (ii) are true is conceivable. Suppose that both *k* and *h* takes *i* to be similar to them *qua* being with a mind and mental states. Furthermore, suppose that both *k* and *h* takes *i* to be radically different from them as far as his response to the environment and his mental state formation are concerned. For instance, *i* is such that, everything else being the same, he recognises the fact that it is raining outside to be a reason to go to Paris during Christmas, he forms the intention of playing football on the basis of such a reason and reads a book about IUCs on the basis of such an intention. Finally, suppose that *k* and *h* can validly reach intersubjective agreement about the interpretation of *i*'s behaviour because they possess a manual, or a 'theory of mind', about *i*, which describes the infinite reason-relations that *i* may recognise and the infinite types of mental interactions that his mental states may instantiate so as to produce observable behaviour. Thus, both *k* and *h* can infer *i*'s mental states and interpret his behaviour on the basis of the relevant evidence, through the use of this book.

In order to defend the thesis that (2') is true, one needs to deny that this example constitute a genuine possibility. The most likely candidate for rejection is the assumption that *k* and *h* can reach valid intersubjective agreement about the interpretation of *i*'s behaviour. In turn, the most likely explanation of why this is impossible is based on the unintelligible character of such an interpretation. The idea is that *k* and *h* cannot take the interpretation of *i*'s behaviour to be correct because, in some sense, it is unintelligible. In what sense exactly? After all, their interpretation is sanctioned by the available evidence and by knowledge of the relevant causal relations

between the environment, *i*'s mental states and *i*'s behaviour. As such, their intersubjective agreement seems to be validly reached.

However, something more seems to be required for validity and, thereby, for taking an interpretation to be correct. An additional requirement is that the concepts in terms of which the interpretation is formulated are applied appropriately. For instance, in the previous example, *k* and *h* interpret *i* as recognising the fact that it is raining outside to be a reason to go to Paris during Christmas. At first sight, however, the fact that it is raining outside can hardly count as a reason to go to Paris during Christmas, at least if we use the concept of 'reason' appropriately. The application of the concept of 'reason' to express the relation between the fact that it is raining outside and the action of going to Paris during Christmas would be appropriate only under the hypothesis that *i* were mistaken in several of his beliefs and/or evaluations. For instance, *i* may associate rain with Paris because he spent an amusing day in Paris under the rain, with his wife, few years before. The combination of his past experience and the fact that it is raining outside generates in *i* a desire to go to Paris and the mistaken beliefs that the fact that it is raining outside is a reason to go there. In turn, *i*'s mistaken belief would make the fact that it is raining outside an operative reason to go to Paris during Christmas. However, this case is excluded by stipulation. By assumption, *k* and *h* know all the circumstances to which *i* is subject, including his personal history. By assumption, *i* does not make any evaluative and/or non-evaluative mistake. Everything else is normal. Nothing else can explain why *i* recognises the fact that it is raining outside as a reason to go to Paris during Christmas, except, perhaps, that this is how *i* uses the concept of 'reason'.

Such a radical diversity sets a limit to the intelligibility of interpretation. If no plausible explanation of why *i* recognises a reason-relation between two apparently unrelated facts is possible, *i*'s mental behaviour appears mysterious. Although the manual indicates that this is how *i*'s mind works, the intelligibility of the interpretation at the personal level remains seriously compromised. In the light of this feature, we may conjecture that *k*'s and *h*'s most likely reaction is to judge both their interpretation and the book on which it is based to be mistaken. As a consequence, *k*'s and *h*'s most likely reaction is to look for an alternative interpretation of *i*'s behaviour, where the concept of 'reason' is used in accordance with their standards of appropriateness. The important point is that these standards seem to be fixed, *ceteris paribus*, by the conditions of application that they recognise. In turn, such conditions are determined by what they would recognise to be an (either believed or normative) reason in similar circumstances, that is, they are determined by the principle of recognition. Since embracing the

principle of recognition is what guarantees the intelligibility of interpretation and, in turn, intelligibility is one of the necessary conditions for taking an interpretation to be correct, it follows that embracing the principle of recognition is necessarily required for taking an interpretation to be correct.

A similar idea can be expressed with respect to the principle of rationality. In the previous example, if no plausible explanation of why *i* forms the intention of playing football on the basis of the fact that rain outside is a reason to go to Paris during Christmas, the interpretation remains unintelligible. Once again, *k*'s and *h*'s most likely reaction is to look for an alternative interpretation of *i*'s behaviour, where mental state formation obeys the standards of appropriateness that they recognise. Yet, once they embrace their own concept of 'reason', it appears that, *ceteris paribus*, those standards are fixed by how their own mind would interact on the basis of the (either believed or normative) reasons that they recognise, that is, they are fixed by the principle of rationality. The result is that embracing the principle of rationality is necessarily required for the intelligibility of interpretation and, thereby, for taking an interpretation to be correct.

If the remarks in this section are correct, it follows that a 'modest' transcendental argument applied to the case of interpretation can be vindicated.

8. The epistemological problem of IUCs re-considered

The 'modest' transcendental argument that we have examined raises one important objection: the appeal to a vague notion like intelligibility cannot, by its very nature, provide a conclusive reason to accept the thesis that we take, or believe, preferences to be interpersonally comparable in terms of strength. I think that this point should be granted. Nevertheless, for the sake of the argument, I shall assume that the 'modest' transcendental argument is genuinely sound. In this section, I shall focus on the question of what implications this argument has for the problem of IUCs.

In the light of (3.1'), the 'modest' transcendental argument assures a transcendental invulnerability to the belief that preferences are interpersonally comparable. The invulnerability is due to the fact that we could not reach intersubjective agreement about the interpretation of other people's behaviour without taking such a belief to be true. At first sight, however, the fact that we could not but assume comparability in order to reach intersubjective agreement does not have any interesting epistemological consequences. After all, what we are bound to do is a descriptive matter, while

knowledge and justification are evaluative matters. On the one hand, the fact that we are bound to take, or believe, preferences to be interpersonally comparable from the start does not imply that preferences are really interpersonally comparable from the start. On the other hand, the fact that we are bound to take, or believe, preferences to be interpersonally comparable from the start does not imply that our beliefs about how different people's preferences compare in terms of strength can be (scientifically) justified, even if the former belief provides the basis for the latter kind of belief. The upshot is that the 'modest' transcendental argument appears to have no implications for the epistemological problem of IUCs.

However, this conclusion is too quick. Indeed, several authors have argued, in general, that a 'modest' transcendental argument has interesting anti-sceptical implications. According to Stern, the crucial distinction to keep in mind is the one between the "epistemic sceptic" and the "justificatory sceptic" [34]. The former is the fictitious individual who doubts whether we can have any knowledge at all. The latter is the fictitious individual who doubts whether we can have any justified beliefs at all. According to Stern, the 'strong' transcendental argument is directed towards the "epistemic sceptic", while the 'modest' transcendental argument is directed towards the "justificatory sceptic". Given that all 'strong' transcendental arguments fail, the doubt about whether we can have genuine (scientific) knowledge remains. Nevertheless, a 'modest' transcendental argument can guarantee the possibility of having (scientifically) justified beliefs about certain subject matters. In the case under consideration, the doubt remains about whether or not preferences are interpersonally comparable in terms of strength and – ultimately – about whether or not we can have (scientific) knowledge of ICs of preference strength. However, if Stern is correct and if the 'modest' transcendental argument is successful, it is nonetheless possible to have (scientifically) justified ICs of preference strength. Let us examine how.

The 'modest' transcendental argument implies that, if different interpreters are bound to take, or believe, preferences to be comparable from the start, then, *ceteris paribus*, they form the same beliefs about how different individuals' preferences compare in terms of strength. However, there are two things that the 'modest' transcendental argument cannot do. First, it cannot show that our belief that preferences are interpersonally comparable is evidence that they really are. Second, it cannot show that the PoS is reliable for making ICs of preference strength. Therefore, the problem is that a 'modest' transcendental argument cannot show, by itself, that it is possible to have

[34] See STERN, R. [1999], p. 42.

(scientifically) justified ICs of preference strength both within an evidentialist framework and within a reliabilist framework. However, a 'modest' transcendental argument can show that we can have (scientifically) justified ICs if it is combined with a coherentist theory of epistemic justification. As we have seen in chapter 1, this is a theory about the structure of justification. Its central tenet is that each belief is justified in terms of other beliefs, so that justification is simply a function of the relationship between various beliefs.

First, let us consider the evidentialist framework. Suppose that an individual *k* wants to compare the preference strengths of two individuals *i* and *j*. Suppose also that the evidence concerning their preferences is entirely identical. Finally, suppose that *k* is an incurable sceptic about ICs of preference strength. As a consequence, *k* believes that: (i) the evidence concerning *i*'s and *j*'s preferences is the same; (ii) for any two individuals *i* and *j*, if the evidence concerning their preferences is the same, they have the same preference strengths; (iii) *i* and *j* do not have the same preference strengths. If consistency is the mark of justification, then at least one of *k*'s beliefs is not justified, because (i), (ii) and (iii) do not form a consistent set. This does not tell us yet which of his beliefs *k* should revise. However, things change if the 'modest' transcendental argument seen above is brought into play. Necessarily, *k* takes, or believes, (ii) to be true in order to reach valid intersubjective agreement. Thus, a coherentist theory of epistemic justification requires *k* to revise either (i) or (iii). If, as it seems plausible, it is possible to justify (i) on independent grounds, (iii) remains the only unjustified belief. Thus, a coherentist theory of epistemic justification requires *k* to believe its opposite, i.e. (iii'), according to which *i* and *j* have the same preference strengths. In other words, if (i) can be justified on independent grounds and if *k* is bound to believe (ii), (iii') is the only justified ICs of preference strength that *k* can make. The result is that, when combined with a coherentist theory of epistemic justification, the 'modest' transcendental argument shows that ICs can be (scientifically) justified within an evidentialist framework[35]. Indeed, if the empirical evidence determines (i) and if the 'modest' transcendental arguments assures invulnerability to (ii), then the coherentist theory of justification *determines* (iii') in accordance with acceptable standards.

Second, let us consider a reliabilist framework. Consider the previous example. If we embrace a coherentist theory of epistemic justification, consistency becomes the mark

[35] The same is true if one embrace a deontologist theory of justification. If one of *k*'s epistemic duties is to maximise the consistency of his beliefs, then, *ceteris paribus*, he is deontologically justified if he believes (iii'), but not if he believes (iii). From an epistemic point of view, given that he is bound to believe (ii), *k* cannot do any better, in terms of justification, than in the case expressed by (iii').

of justification. In particular, consistency shows that the individual's doxastic attitudes are reliably acquired and, thereby, that they are justified. However, (i), (ii) and (iii) do not form a consistent set of beliefs. This suggests that at least one of these beliefs is not reliably acquired and, therefore, is unjustified. Once again, if the 'modest' transcendental argument is sound, necessarily, *k* takes, or believes, (ii) to be true in order to reach valid intersubjective agreement. As *k* cannot revise (ii), the only possibility to have consistent beliefs is by revising either (i) or (iii). Once again, if it is possible to show on independent ground that (i) is reliably acquired, justification requires *k* to believe (iii'), as (iii') is the only belief that makes the set consistent. It follows that, when combined with a coherentist theory of justification, the 'modest' transcendental argument shows that ICs can be (scientifically) justified also within a reliabilist framework. Indeed, the evidence about the reliable acquisition of (i), the transcendental invulnerability of (ii) and a coherentist theory of justification jointly *determines* (iii') in accordance with acceptable standards.

9. Conclusion

In this chapter, I considered whether or not we can solve the problem of IUCs by appealing to a transcendental argument. I examined two kinds of transcendental argument: a 'strong' and a 'modest' transcendental argument. The 'strong' transcendental argument attempts to show that different people's preferences are interpersonally comparable. One of the consequences is that, if the relevant evidence is correct, it is possible to have (scientifically) justified ICs of preference strength. However, I argued that such an argument succumbs to the objection that is generally made against transcendental arguments of a 'strong' form, namely, that, *ceteris paribus*, it is not possible to infer a conclusion about how a mind-independent world is from a premise about our psychological reality.

The 'modest' transcendental argument attempts to show only that we take, or believe, different people's preferences to be interpersonally comparable. At first sight, even if successful, this argument does not seem to bring any result of anti-sceptical significance. After all, even if we are bound to assume that different people's preferences are interpersonally comparable from the start, we cannot assume either that different people's preferences are really interpersonally comparable or that it is possible to have (scientifically) justified ICs of preference strength. However, I argued that the 'modest' transcendental argument can offer a positive solution to the problem of IUCs if

it is combined with a coherentist theory about the structure of justification, both within an evidentialist and within a reliabilist framework. Indeed, if, as it seems plausible, beliefs concerning the evidence about two individuals' behaviour can be independently justified, then the 'modest' transcendental argument and a coherentist theory of justification uniquely determine ICs of preference strength.

CONCLUSION

We started our analysis of the problem of IUCs by making two platitudinous observations. On the one hand, we noticed that, in everyday life, we not only ascribe preferences with a specific content to other people and to ourselves, but we also compare them in terms of strength. A remarkable fact is that we typically make ICs of preference strength with relatively little difficulty. In particular, we often do not find inter-personal comparisons of preferences more difficult than intra-personal comparisons, that is, of comparisons involving our own preferences.

On the other hand, we noticed that the ease with which we compare preferences in everyday life contrasts with the difficulties that ICs of preference strength pose at the theoretical level. In the framework provided by orthodox economics, the problem presents certain characteristic features. Typically, choice behaviour is considered the only admissible evidence for the ascription of preferences. Moreover, preferences are supposed to satisfy a relatively rich set of (both substantive and technical) conditions, which allows us to determine and represent them by means of numerical functions with different uniqueness features. In particular, when preferences satisfy the axioms of expected utility theory, they can be represented by a cardinal utility function, which assigns a measure of preferential strength to each of the options in the preference domain.

The problem of IUCs arises at this stage. Although choice behaviour is sufficient for measuring each individual's preferences, it is not sufficient for determining ICs of different individuals' utilities. For each individual, the measurement is relative to the best and the worst options in his preference ranking. However, choice behavioural evidence is not enough to tell whether or not different people hold the options at the top and at the bottom of their rankings with identical preferential strength. As a consequence, it is not possible to determine, on the basis of choice behaviour, whether or not different people's preference strengths are really the same when they have the same numerical value. Equivalently, on the basis of choice behaviour, it is not possible to claim that different people's utilities are co-scaled.

As IUCs appear to be underdetermined by choice behavioural evidence, the most natural reaction is to extend the set of admissible evidence beyond choice behaviour. In this thesis, we have seen various ways in which this can be done. For instance, one can gather information about latency or probability of choice (e.g. Waldner), verbal expressions (e.g. Harsanyi), expressive reactions (e.g. Weintraub), facial expressions, body temperature and

other proxies (e.g. List)[1] and use them as additional evidence for preference ascription and IUCs. Unfortunately, as shown by various authors and, in particular, by List[2], even such a broader set of empirical evidence is insufficient to determine IUCs. The upshot is that IUCs are empirically meaningless in a very robust sense. This result poses the following crucial epistemological challenge: can we have knowledge, and in particular, scientific knowledge, of how different people's preferences compare in terms of strength? Or, at least, can we have scientifically justified beliefs about how different people's preferences compare in terms of strength? In this thesis, I have tried to present and discuss alternative ways of addressing these issues. In this conclusion I shall pursue two goals. First, I want to summarise the results of the previous analysis. Second, I want to examine where these results leave us and, in particular, whether or not the epistemological problem of ICs of preference strength remains a serious challenge.

As we have seen, although the empirical meaninglessness threatens the possibility of having scientifically justified ICs of preference strength, it does not exclude it completely. Empirical meaningfulness is, at best, only a sufficient condition for scientific justification. Other non-empirical considerations may help break the underdetermination by the empirical evidence and potentially lead to scientifically justified beliefs about how different people's preferences compare in terms of strength. Within an evidentialist framework, this means that other non-empirical considerations might be used as evidence to determine IUCs. By contrast, within a reliabilist framework, this means that other non-empirical considerations might be reliable guides for the determination of IUCs.

The first strategy that we have examined is based on an inference to the best explanation type of argument. In general, the idea is that a theory, or an assumption, is justified if it offers, or contributes to offering, the best explanation of a certain phenomenon. In turn, the criteria for individuating the best explanation typically include pragmatic considerations, such as explanatory power, simplicity, or parsimony. In the case of IUCs, the argument is that we are justified in assuming that different people's utilities are co-scaled insofar as this provides the best explanation of their (comparative) behaviour. Different authors emphasise different pragmatic virtues as mostly relevant for the problem under consideration. For instance, Harsanyi claims that the assumption that different people's utilities are on the same scale is "the least arbitrary hypothesis"[3], at least when all the empirical evidence is the same across individuals; whereas Waldner and List emphasise, respectively, the simplicity and parsimony of such an assumption. In this thesis, however, I argued that this

[1] See HARSANYI, J. [1955] and [1977], LIST, C. [2003], WALDNER, I. [1972], WEINTRAUB, R. [1998].
[2] See LIST, C. [2003].
[3] See HARSANYI, J. [1955] and [1977].

strategy fails. The assumption that different people's utilities are co-scaled is not pragmatically advantageous in any of the senses listed above. Indeed, it does not add anything to the explanation of individual behaviour and it does not make a theory including it either more parsimonious or simpler than a theory that does not include it. Ultimately, this means that this strategy fails to demonstrate that IUCs can be scientifically justified.

The second strategy pursues a nativist approach. An argument of this kind was first offered by Goldman in the context of a sub-personal explanation of our mindreading capacity. Goldman argues that the problem of comparing the intensity of different people's mental states is a particular case of the problem of ascribing mental states to other people. According to Goldman, mindreading consists in simulating, or replicating, the working of another individual's mind. The simulator, first, asks himself what mental states he would have if he were subject to the initial mental states of the simulated agent; then, on the basis of the result of such an introspective exercise, he ascribes – by analogy – these mental states to the simulated agent[4]. Clearly, such a mental ascription is justified to the extent that the simulator's and the simulated agent's mind-systems are similar in certain relevant respects. Predictive success at mindreading offers some evidence that this is indeed the case. However, even if the belief about how different people's preferences compare in terms of strength is formed by using the same information-processing mechanisms, predictive success offers only *prima facie* evidence that ICs of preference strength too are justified. In fact, the interpersonal similarity required in order to have meaningful ICs of preference strength is higher than the one required in order to have reliable behavioural predictions. Goldman fills this gap by arguing that the assumption of interpersonal psychological similarity should be considered an innate feature of the mind.

In chapter 3, I extended this strategy to the other main approach to mindreading, namely, Theory-Theory. According to it, mindreading is performed by means of a 'theory' about other people's mind, which the mindreader more or less tacitly possesses. Predictive success offers some evidence that this 'theory' represents very closely the working of other individuals' mind systems. However, the closeness of the theory representation required in order to have meaningful ICs of preference strength is higher than the one required in order to have reliable behavioural predictions. Once again, one can fill the gap by arguing that this is an innate feature of the mind.

The nativist strategy then claims that the assumption that different people's utilities are co-scaled is justified if ICs of preference strength are performed through innate mechanisms that are either hyper-similar across individuals or very closely representative

[4] See GOLDMAN, A. [1989].

of the working of other individuals' mind-systems. In this thesis, I argued that this strategy fails because it reduces to an inference to the best explanation kind of argument in certain crucial respects. The most recent literature interprets the claim that a cognitive capacity or mechanism is innate as the claim that such a capacity or mechanism is a psychological primitive. In turn, the literature suggests that we have a reason to take a cognitive capacity or mechanism to be a psychological primitive only if, on the one hand, it is part of a correct psychological explanation of human behaviour, and, on the other hand, its acquisition cannot be explained by any theory at the psychological level, but only by a theory at a lower level. Proponents of psychological primitivism suggest identifying 'the correct explanation' by reference to the 'best explanation' in our possession. It thus seems that the second strategy collapses into the first strategy examined in this thesis in at least one crucial respect. There is indeed an important difference between the two. The first strategy tries to solve the problem of IUCs by claiming that we are justified in assuming that different people's utilities are co-scaled because *this* assumption is part of the best explanation of their behaviour. By contrast, the second strategy pursues a more indirect route and holds that we are justified in assuming that different people's utilities are co-scaled because the assumption of *innate interpersonal similarity* of different people's mind-systems (either in the Simulation Theory or in the Theory Theory form) is part of the best explanation of their behaviour. Despite this difference, the same objections made against the first strategy apply to the second as well. In this thesis, I argued that the best explanation of how people make ICs of preference strength merely requires that people take, or believe, the assumption of interpersonal similarity (either in the Simulation Theory or in the Theory Theory form) to be true, but not that the assumption is really true.

In the wake of the failure of these strategies, some authors attempt to contrast the sceptical challenge about IUCs by offering more radical 'in principle' solutions, which take the form of 'possibility' arguments. Their primary goal is to show that different people's preference strengths are comparable. Their secondary goal is to show that it is possible, in principle but not by means of empirical or pragmatic considerations alone, to have (scientific) knowledge of, or, at least, (scientifically) justified beliefs about how different people's preferences compare in terms of strength. In this thesis, I considered both 'possibility' arguments made in the context of an economic-oriented analysis and one 'possibility' argument made in the context of a more philosophy-oriented analysis. Within the former set, the first argument that I considered is based on Broome's work on personal goodness. It claims that, if individual preferences are independent from personal identity, then it is conceptually possible to construe a universal preference scale, provided that each

person shares two possible lives with at least another person, that is, each person belongs to at least one overlapping pair, and that everyone is connected to everyone else by a chain of overlapping pairs. If this is the case, different people's preferences are interpersonally comparable in terms of strength. Within the same set, I considered two arguments that are based on a functionalist understanding of the nature of preferences. According to both of them, it is conceptually possible to identify two points that play the same causal role in different people's minds. If this is the case, functionalism allows us to conclude that preferences are interpersonally comparable in terms of strength. The former argument claims that these points are given by the most preferred and by the least preferred option, respectively. The latter argument – offered by Bradley – claims that one point is given by the ethically neutral prospect, while the other is given by the total desirability of all prospects in the preference domain. Within the latter set of 'possibility' arguments, I considered a 'strong' transcendental argument. Its goal is to demonstrate that different people's preferences are interpersonally comparable from the start on the grounds that, necessarily, correct interpretation requires interpersonal comparability.

These arguments face different challenges. However, the main objection that can be raised against them is a common one: they do not to show that different individuals respond to the environment and form mental states in the same way. This is a recurrent theme across this thesis. In fact, the nativist strategy too similarly fails to show that different people's mind-systems are identical, in certain crucial respects. Therefore, this emerges as the main factor preventing the possibility of having scientific knowledge of, or, at least, scientifically justified beliefs about, how different people's preferences compare in terms of strength.

The previous analysis suggests that, perhaps, we should stop wondering whether or not different people's mind-systems are *really* identical, or hyper-similar, and rather starting from the claim that we do *take* them to be so. This remark is what motivates exploring a more 'modest' transcendental solution to the problem of IUCs. The objection that is generally made against 'strong' transcendental arguments is that it is not possible to infer a conclusion about a mind-independent world from a premise about our psychological reality. In the case under consideration, the objection is that it is not possible to infer a conclusion about the interpersonal comparability of preference strengths from a premise about the (supposed) correctness of interpretation. The goal of a 'modest' transcendental argument is precisely to demonstrate that, necessarily, interpretation requires the interpreter only to *take*, or believe that, different people's preferences are interpersonally comparable from the start. In this thesis, I tried to defend a 'modest' transcendental argument by insisting that its acceptance is required by the very intelligibility of interpretation.

Admittedly, the nature of this defence is such as to leave margin for vagueness and, thereby, disagreement. Moreover, the alleged fact that we are bound to take preferences to be interpersonally comparable from the start does not have, by itself, clear implications for the epistemic evaluation of ICs of preference strength. In other words, even if successful, a 'modest' transcendental argument needs to be complemented by additional considerations. In this thesis, I suggested that the transcendental strategy shows that ICs of preference strength can, at least, be scientifically justified if it is combined with a coherentist theory about the structure of justification, both within an evidentialist and a reliabilist framework. If coherentism is defensible, it is possible to show that scepticism does not get off the ground, although it clearly remains a logical possibility. The main defect of this solution is that the acceptance of the coherentist assumption is very controversial and questionable. Therefore, one may read this thesis as achieving a disjunctive result: either it provides a positive argument for the possibility of having (scientifically) justified IUCs, if coherentism is true, or it provides an argument by elimination, to the effect that none of the existing solutions allow for the possibility of having (scientifically) justified IUCs.

At this stage, we are finally in a position to ask whether or not the epistemological problem of ICs of preference strength remains a pressing challenge. How threatening is this problem, if it is mainly due to the fact that it is not possible to show that the assumption of interpersonal similarity between people's mind-systems is sound? After all, the problem seems to massively generalize. Let me explain why. Suppose we consider two material bodies belonging to the same natural kind. Their internal workings are as unobservable as the internal workings of different individuals' mind-systems. Typically, in the former case, we assume that the internal structure is functionally identical when the empirical evidence is the same. However, one might argue that the empirical evidence is not really sufficient to justify this assumption and leaves room for the sceptical hypothesis that the two bodies' internal structure may be different. If such scepticism is a serious possibility, it brings some interesting implications to the fore. For each item, we can identify the internal elements of their functionally defined structure on the basis of the causal role that they play. Notice that the very notion of causal role cannot but be defined with respect to a background theory about how the elements of the system causally interact. In order to compare the functional properties of the items, we must individuate cases where the properties under consideration play the same casual role. However, this requires us to make an assumption of internal similarity. If this assumption is not justified, the results of our comparison are not justified either. Let us consider an example. Suppose we want to measure the temperature of two objects, belonging to the same natural kind. Suppose that the empirical evidence about both

of them is completely identical. Typically, we would be prepared to ascribe the same temperature to these objects. However, if we have no reason to assume that their internal structure is the same, this conclusion is unjustified. The problem is that this might be true for the measurement of all functionally defined properties!

If we want to preserve a certain significance to the problem of IUCs, we need to offer some reasons as to why the case concerning the mind invites a more serious scepticism than the case concerning material bodies. Here I shall simply sketch some of them. The first reason is that there is no 'third' object to which the measurement of preference strength can be related. The temperature of two objects belonging to the same natural kind is supposed to be comparable because it is possible to determine two points that these objects have in common, namely, the water's freezing point and the water's boiling point. These points provide a common reference with respect to which the temperature of the objects can be measured and co-scaled. Notice, however, that commonality is defined with respect to a third object, i.e. water, which belongs to a different natural kind. No analogue object exists for ICs of preference strength. Even when preferences are defined in functionalist terms, the parallel between the measurement of preference strength and temperature breaks down in some crucial respects. For instance, consider Bradley's solution. Although the ethically neutral prospect is supposed to play the role of water, the two cases are not perfectly analogous. In fact, the ethically neutral prospect is one of the items included in the preference domain of each individual. By contrast, water is a different object altogether. Perhaps, this may not be a problem. After all, the comparability of beliefs is equally based on the individuation of two propositions, namely, the necessary and the impossible propositions, with respect to which different people's beliefs are supposed to play the same causal role. If the reasoning is sound in the latter case, it must be sound also in the case of preferences.

The second reason is that, even if it is conceptually possible to identify cases with respect to which different people's preferences play the same causal role, it is not possible to gain epistemic access to such cases. In other words, even if it is possible to solve the conceptual problem of whether or not preference strengths are interpersonally comparable, it is not possible to solve the epistemological problem of whether or not we can have (scientific) knowledge or (scientifically) justified beliefs about how different people's preferences compare in terms of strength. If Bradley's analysis is correct, we can have, at most, (scientifically) justified ICs of preference strength with respect to a significant zero-line, but not (scientifically) justified ICs of preference strength levels or differences.

The third reason has to due with the difficulty concerning the characterisation of the mind-body relation. Most of the methods used to measure temperature rely on measuring some physical property of a material body, which varies with temperature. The relation between physical properties and functionally defined properties here is straightforward. However, this is not the case for mental properties. Their relation with the underlying neurophysiological properties is a particularly controversial issue. As a consequence, inferences from the physical to the mental level are still regarded with suspicion.

Finally, and related to the previous, the fourth reason why the problem of IUCs remains an interesting challenge is given by the difficulty in addressing the question about the nature of mental states. In particular, the debate remains open about whether or not, and to what extent, we should embrace some sort of psychological realism about mental states or we should rather favour some anti-realist account.

As we have seen in various places, the problem of IUCs is not independent from the solutions given to these issues. This explains why the assumption of interpersonal similarity and, more generally, the comparison of different people's preference strengths keep raising sceptical doubts. Nonetheless, I hope that this thesis may contribute to the debate by clarifying the nature of the problem, putting forward some new solutions and suggesting possible directions for future analysis.

BIBLIOGRAPHY

ADAM, M. "Two Notions of Scientific Justification", *Synthese*, 158 (2007), pp. 93-108.

ARNESON, R. J. "Equality and Equal Opportunity for Welfare", *Philosophical Studies*, 56 (1989), pp. 77-93.

ARROW, K. J. *Social Choice and Individual Values*, 2nd ed., New York: Wiley, 1963 (1951, 1st ed.).

"Extended Sympathy and the Possibility of Social Choice", *The American Economic Review*, 67 (1977), pp. 219-225.

BARON-COHEN, S. *Mindblindness: An Essay on Autism and Theory of Mind*, Cambridge, Mass.: MIT Press, 1995.

BARRY, B. *Theories of Justice*, Berkeley: University of California Press, 1989.

BLOCK, N. "Troubles with Functionalism", in BLOCK, N. (ed.) *Readings in Philosophy of Psychology*, Vol. 1, Cambridge, Mass.: Harvard University Press, 1980.

BRADLEY, R. "Impartiality in *Weighing Lives*", *Philosophical Books*, 48 (2007a), pp. 292-302.

"Comparing Evaluations", Unpublished Manuscript, 2007b.

BROOME, J. "A cause of preference is not an object of preference", *Social Choice and Welfare*, (10), 1993, pp. 57-68.

Ethics out of Economics, Cambridge: Cambridge University Press, 1999.

Weighing Lives, Oxford: Blackwell, 2004.

"Reply to Bradley and McCarthy", *Philosophical Books*, 48 (2007), pp. 320-328.

"Can There Be a Preference-Based Utilitarianism?" in M. SALLES and J. WEYMARK (eds.), *Justice, Political Liberalism and Utilitarianism: Themes from Harsanyi and Rawls*, Cambridge: Cambridge University Press, forthcoming.

CARRUTHERS, P, and P., SMITH (eds.) *Theories of Theories of Mind*, Cambridge: Cambridge University Press, 1996.

CHANG, R. (ed.) *Incommensurability, Incomparability, and Practical Reason*, Cambridge, Mass., Harvard University Press, 1997a.

"Introduction", in CHANG, R. (ed.) *Incommensurability, Incomparability, and Practical Reason*, Cambridge, Mass., Harvard University Press, 1997b, pp. 1-34.

CHIPMAN, J. S. and J. C., MOORE "The New Welfare Economics, 1939-1974, *International Economic Review*, 19 (1978), pp. 547-584.

CHOMSKY, N. *Rules and representations*, Oxford: Basil Blackwell, 1980.

COOTER, R. and P., RAPPOPORT "Were the Ordinalists Wrong about Welfare Economics?", *Journal of Economic Literature*, 22 (1984), pp. 507-530.

"Reply to I.M.D. Little's Comment", *Journal of Economic Literature*, 23 (1985), pp. 1189-1191.

COWIE, F. *What's Within? Nativism Reconsidered*, Oxford: Oxford University Press, 1999.

D'ASPREMONT, C. and L., GEVERS "Equity and Informational Basis of Collective Choice", *Review of Economics Studies*, 44 (1977), pp. 199-209.

DAVIDSON, D. *Inquiries into Truth and Interpretation*, Oxford: Oxford University Press, 1984.

"Judging interpersonal interests", in ELSTER, J. and A., HYLLAND (eds.), *Foundations of social choice theory*, Cambridge: Cambridge University Press, 1986, pp. 195-211.

"Interpersonal Comparisons of Values", in DAVIDSON, D. *Problems of Rationality*, Oxford: Oxford University Press, 2004, pp. 59-74.

DAVIES, M. "Interaction without reduction: The relationship between personal and subpersonal levels of description" *Mind and Society*, 1 (2000), pp. 87–105.

DAVIES, M. and T., STONE (eds.) *Folk Psychology: The Theory of Mind Debate*, Oxford: Blackwell, 1995a.

(eds.) *Mental Simulation*, Oxford: Blackwell, 1995b.

"The Mental Simulation Debate: A Progress Report", in CARRUTHERS, P, and P., SMITH (eds.) *Theories of Theories of Mind*, Cambridge: Cambridge University Press, 1996, pp. 119-137.

"Simulation Theory", Entry for *Routledge Encyclopedia of Philosophy Online*, 2000.

"Mental Simulation, Tacit Theory, and the Threat of Collapse", *Philosophical Topics*, 29 (2001), pp. 127-73.

DENNETT, D. *Content and Consciousness*. London: Routledge and Kegan Paul, 1969.

The Intentional Stance, Cambridge, Mass.: MIT Press, 1987.

DWORKIN, R. "What is Equality? Part 1: Equality of Welfare", *Philosophy and Public Affairs*, 10 (1981a), pp. 185-246.

"What is Equality? Part 2: Equality of Resources", *Philosophy and Public Affairs*, 10 (1981b), pp. 283-345.

Sovereign Virtue: The Theory and Practice of Equality, Cambridge, Mass.: Harvard University Press, 2000.

ELSTER, J. and J. E., ROEMER (eds.), *Interpersonal Comparisons of Well-Being*, Cambridge: Cambridge University Press, 1991, pp. 1-16.

FELDMAN, R. and E., CONEE "Evidentialism", *Philosophical Studies*, 48 (1985), pp. 15-34.

FLEURBAEY, M. and J., HAMMOND "Interpersonally Comparable Utility" in BARBERÀ, S. HAMMOND, J. and C., SEIDL (eds.), *Handbook of Utility Theory*, Vol. II, Kluwer Academic, 2004.

FODOR, J. and E., LEPORE "*D.C. Dennett*: Meaning Holism and the Normativity of Intentional Ascription (and a Little More about Davidson)", in FODOR, J. and E., LEPORE *Holism: A Shopper's Guide*, Oxford: Blackwell Publishing, 1992, pp. 137-162.

GALLESE, V. and A. I., GOLDMAN "Mirror neurons and the simulation theory of mindreading", *Trends in Cognitive Science*, 2 (1998), pp. 493-501.

GETTIER, E. "Is Justified True Belief Knowledge?", *Analysis*, 23 (1963), pp. 121-123.

GIBBARD, A. "Interpersonal Comparisons: Preference, Good, and the Intrinsic Reward of a Life", in ELSTER, J. and A., HYLLAND (eds.) *The Foundations of Social Choice Theory,* Cambridge: Cambridge University Press, 1986, pp. 165–193.

GIBSON, R. "Translation, Physics, and Facts of the Matter", in HAHN, L. E. and P. A., SCHILPP (eds.), *The Philosophy of W. V. Quine*, Open Court, La Salle, IL, 1986, pp. 139-157.

GOLDMAN, A. I. "What is Justified Belief?" in PAPPAS, G. S. (ed.), *Justification and Knowledge*, Reidel: Kluwer Academic Publisher, 1979, pp. 1-23.

"Interpretation Psychologized", *Mind and Language*, 4 (1989), pp. 161-185.

"In defense of the simulation theory", *Mind and Language*, 7 (1992), pp. 104-119.

"The Psychology of Folk Psychology", *Behavioral and Brain Sciences*, 16 (1993), pp. 15-28.

"Simulation and Interpersonal Utility", *Ethics*, 4 (1995a), pp. 709-726.

"Empathy, mind, and morals", in DAVIES, M. and T., STONE (eds.) *Mental Simulation: Evaluations and Applications*, Oxford: Blackwell, 1995b, pp.185–208.

"The mentalizing folk", in D. SPERBER (ed.), *Metarepresentations*. Oxford: Oxford University Press, 2000.

"Simulation theory and mental concepts", in Dokic, J. and J., Proust (eds.) *Simulation and Knowledge of Action*, Amsterdam: John Benjamins, 2002, pp. 1-20.

Simulating Minds, Oxford: Oxford University Press, 2006.

GOPNIK, J. and A. N., MELTZOFF *Words, Thoughts and Theories*, Cambridge, Mass.: MIT Press, 1997.

GOPNIK, J. and H., WELLMAN "Why the child's theory of mind really *is* a theory of mind", *Mind and Language*, 7 (1992), pp. 145-171.

"The theory theory", in HIRSCHFILED, L. and S., GELMAN (eds.) *Mapping the Mind: Domain Specificity in Cognition and Culture*, New York: Cambridge University Press, 1994.

GORDON, R. M. "Folk Psychology as Simulation", *Mind and Language*, 1 (1986), pp. 158-171.

"The simulation theory: Objections and misconceptions", *Mind and Language*, 7 (1992), pp. 11–34.

GRANDY, R. "Reference, meaning, and belief, *Journal of Philosophy*. 70 (1973), pp. 439-452.

GRIFFIN, J. *Well-Being: Its Meaning, Measurement, and Moral Importance*, Oxford: Oxford University Press, 1986.

GRIFFITHS, P. "What is innateness?", *Monist,* 85 (2002), pp. 70 – 85.

HAMMOND, P. "Equity, Arrow's Conditions and Rawls' Difference Principle", *Econometrica*, 44 (1976), pp. 793-804.

"Interpersonal comparisons of utility: Why and how they are and should be made", in ELSTER, J. and J. E., ROEMER (eds.), *Interpersonal Comparisons of Well-Being*, Cambridge: Cambridge University Press, 1991, pp. 200-254.

HARSANYI, J. C. "Cardinal Welfare, Individualistic Ethics, and Interpersonal Comparisons of Utility", *The Journal of Political Economy*, 63 (1955), pp. 309-321.

Rational Behaviour and Bargaining Equilibrium in Games and Social Situations, Cambridge: Cambridge University Press, 1977.

"Morality and the theory of rational behaviour", in SEN, A. and B., WILLIAMS (eds.) *Utilitarianism and Beyond*, Cambridge: Cambridge University Press, 1982, pp. 39-62.

HAUSMAN, D. "The Impossibility of Interpersonal Utility Comparisons", *Mind*, 415 (1995), pp. 473-490.

HEAL, J. "Replication and Functionalism", in BUTTERFIELD, J. (ed.) *Language, Mind and Logic*, Cambridge: Cambridge University Press, 1986.

"Simulation vs. theory theory: What is at issue?", in PEACOCKE, C. (ed.), *Objectivity, Simulation and the Unity of Consciousness: Current Issues in the Philosophy of Mind* (Proceedings of the British Academy vol. 83), Oxford: Oxford University Press, 1994, pp. 129–44.

"Co-cognition and off-line simulation: Two ways of understanding the simulation approach", *Mind and Language*, 14 (1998a), 477–98.

"Understanding other minds from the inside", in O'HEAR, A. (ed.), *Contemporary Issues in Philosophy of Mind*, Cambridge: Cambridge University Press, 1998b, pp.83–99.

"Other minds, rationality and analogy", *Proceedings of the Aristotelian Society, Supplementary Volume* 74, 2000, pp. 1–19.

HENNIPMAN, P. "A New Look at the Ordinalist Revolution: Comments on Cooter and Rappoport", *Journal of Economic Literature*, 26 (1988), pp. 80-85.

JEFFREY, R. C. "On Interpersonal Utility Theory", *The Journal of Philosophy*, 68 (1971), pp. 647-656.

"Remarks on Interpersonal Utility Theory", in Stenlund, S. (ed.) *Logical Theory and Semantic Analysis*, Dordrecht: D. Reidel, 1974, pp. 35-44.

The Logic of Decision, 2nd ed., Chicago: University of Chicago Press, 1983 (1965, 1st ed.).

Probability and the art of judgment, Cambridge: Cambridge University Press, 1992.

JEVONS, S. *Theory of Political Economy*, 4th edition, London: Macmillan, 1911 (1871, 1st ed.).

KAGAN, S. *Normative Ethics*, Boulder, Colorado: Westview Press, 1998.

KHALIDI, M. A. "Innate Cognitive Capacities", *Mind and Language*, 22 (2007), pp. 92-115.

KRANTZ, D.H., LUCE, R.D., SUPPES, P. and TVERSKY, A. *Foundations of Measurement: Vol. 1. Additive and polynomial representations*, New York: Academic, 1971.

KOLODNY, N. "Why be rational?", *Mind*, 114 (2005), pp. 509-563.

LEPORE, E. and K. LUDWIG, *Donald Davidson. Meaning, Truth, Language, and Reality*, Oxford: Clarendon Press, 2005.

LESLIE, A. "Pretence and representation: The origins of "theory of mind".", *Psychological Review*, 94 (1987), pp. 412-426.

"Some implications of pretense for mechanisms underlying the child's theory of mind", in ASTINGTON, J. HARRIS, P. and D., OLSON (eds.), *Developing Theories of Minds*, Cambridge: Cambridge University Press, 1988, pp. 19-46.

"Pretending and Believing: Issues in the theory of ToMM.", *Cogniton*, 50 (1994), pp. 211-238.

"How to acquire a representational theory of mind", in SPERBER, D. (ed.), *Metarepresentation: A Multidisciplinary Perspective*, New York: Oxford University Press, 2000, pp. 197-223.

LESLIE, A. and T., GERMAN "Knowledge and ability in "theory of mind": One-eyed overview of a debate", in DAVIES, M. and T., STONE (eds.) *Mental Simulation*, Oxford: Blackwell, 1995, pp. 123-150.

LEWIS, D. "Psychophysical and Theoretical Identifications", *Australasian Journal of Philosophy*, 50 (1972), pp. 249-258.

Philosophical Papers. Vol. 1, Oxford: Oxford University Press, 1986.

LIST, C. "A Note on Introducing a 'Zero-Line' of Welfare as an Escape-Route from Arrow's Theorem", *Pacific Economic Review*, 6 (2001), pp. 223-238.

"Are Interpersonal Comparisons of Utility Indeterminate?", *Erkenntnis*, 58 (2003), pp. 229-260.

LITTLE, I. D. M. *A Critique of Welfare Economics*, 2nd ed., Oxford: Clarendon Press, 1957 (1950, 1st ed.).

"Robert Cooter and Peter Rappoport, 'Were the Ordinalists Wrong about Welfare Economics?': A Comment", *Journal of Economic Literature*, 23 (1985), pp. 1186-1188.

MACKAY, A. F. "Extended Sympathy and Interpersonal Utility Comparisons, *The Journal of Philosophy*, 83 (1986), pp. 305-322.

MCGINN, C. *Knowledge and Reality*, 2nd ed. Oxford: Oxford University Press, 2002 (1999 1st ed.).

MISHAN, E. J. "A Survey of Welfare Economics, 1939-1959", *The Economic Journal*, 70 (1960), pp. 197-265.

MISKIN, E. "A Theorem on Utilitarianism", *Review of Economic Studies*, 45 (1978), pp. 93-96.

MONGIN, P. "The Impartial Observer Theorem of Social Ethics", *Economics and Philosophy*, 17 (2001), pp. 147-179.

NG, Y.-K. "Happiness Surveys: Some Comparability Issues and an Explanatory Survey Based on Just Perceivable Increments, *Social Indicators Research*, 38, 1996, pp. 1-27.

"A Case for Happiness, Cardinalism, and Interpersonal Comparability", *The Economic Journal*, 107 (1997), pp. 1848-1858.

NICHOLS, S., STICH, S. LESLIE, A. and D., KLEIN "Varieties of Off-Line Simulation" in CARRUTHERS, P, and P., SMITH (eds.) *Theories of Theories of Mind*, Cambridge: Cambridge University Press, 1996, pp. 39-74.

NICHOLS, S. and, S., STICH, S. "Rethinking co-cognition: A reply to Heal", *Mind and Language*, 13 (1998) pp. 499–512.

Mindreading: An Integrated Account of Pretence, Self-Awareness, and Understanding Other Minds, Oxford: Oxford University Press, 2003.

NOZICK, R. *Anarchy, State and Utopia*, New York, Basic Books, 1974.

PARFIT, D. *Reasons and Persons*, Oxford: Oxford University Press, 1984.

PEIJNENBURG, J. and R., HÜNNEMAN, "Translations and Theories: On the Difference between Indeterminacy and Underdetermination", *Ratio*, 14 (2001), pp. 18-32.

PERNER, J. *Understanding the Representational Mind*, Cambridge, Mass.: MIT Press, 1991.

"Simulation as explicitation of predication-implicit knowledge about the mind: arguments for a simulation-theory mix", in CARRUTHERS, P, and P., SMITH (eds.) *Theories of Theories of Mind*, Cambridge: Cambridge University Press, 1996, pp. 90-104.

PETTIT, P. "Decision Theory and Folk Psychology", in BACHARACH, M. and S., HURLEY (eds.), *Foundations of Decision Theory*, Oxford: Basil Blackwell, 1991, pp. 147-175.

Rules, Reasons, and Norms, Oxford: Oxford University Press, 2002.

"Preference, Deliberation and Satisfaction", *Royal Institute of Philosophy Supplement*, 81 (2006), pp. 131-154.

PLANTINGA, A. *Warrant and Proper Function*, New York: Oxford University Press, 1993.

RABINOWICZ, W. and J., OSTERBERG "Value Based on Preferences", *Economics and Philosophy*, 12 (1996), pp. 1-27.

RABINOWICZ, W. "Modelling parity and incomparability" in RABINOWICZ W. and T., RØNNOW-RASMUSSEN (eds.) *Patterns of value: essays on formal axiology and value analysis*, Vol. 2. Department of Philosophy, Lund University, Lund, 2004.

RAMSEY, F. P. "Truth and Probability", 1926, reprinted in MELLOR, D. H. (ed.) *Philosophical Papers*, Cambridge: Cambridge University Press, 1990.

RAPPOPORT, P. "Reply to Professor Hennipman", *Journal of Economic Literature*, 26 (1988), pp. 86-91.

RAWLS, J. *A Theory of Justice*, Cambridge, Mass.: Harvard University Press, 1971.

"Social unity and primary goods", in SEN, A. and B., WILLIAMS (eds.) *Utilitarianism and Beyond*, Cambridge: Cambridge University Press, 1982, pp. 159-186.

RAZ, J. *Practical Reason and Norms*, Oxford: Oxford University Press, 1975.

Engaging Reasons, Oxford: Oxford University Press, 1999.

"The role of well-being", *Philosophical Perspectives*, 18 (2004), pp. 269-294.

ROBBINS, L. *An Essay on the Nature and Significance of Economic Science*, London: Macmillan, 1932.

ROBERTS, F. S. *Measurement theory with applications to decision-making, utility, and the social sciences*, Reading, Mass.: Addison-Wesley Advanced Book Program, 1979.

ROBERTS, K. W. S. "Possibility Theorems with Interpersonally Comparable Utility Levels", *Review of Economic Studies*, 47 (1980a), pp. 409-420.

"Interpersonal Comparability and Social Choice Theory", *Review of Economic Studies*", 47 (1980b), pp. 421-439.

"Valued Opinions or Opinionized Values: the Double Aggregation Problem", in BASU, K., PATTANAIK, P. K. and K., SUZUMURA (eds.), *Choice, Welfare and Development: A Festschrift in Honour of Amartya Sen*, Oxford: Clarendon Press, 1995, pp. 141-185.

ROEMER, J. E. *Equality of Opportunity*, Cambridge, Mass.: Harvard University Press, 1998.

SAMUELSON, P. A. *Foundations of Economic Analysis*, Cambridge, Mass.: Harvard University Press, 1947.

SAMUELS, R. "Nativism in cognitive science", *Mind & Language*, 17 (2002), pp. 233 – 265.

SAVAGE, L. J. *The Foundations of Statistics*, New York: Wiley, 1954.

SCANLON, T. M. "The moral basis of interpersonal comparisons", in ELSTER, J. and J. E., ROEMER (eds.), *Interpersonal Comparisons of Well-Being*, Cambridge: Cambridge University Press, 1991, pp. 17-44.

What We Owe To Each Other, Cambridge, Mass.: Harvard University Press, 1998.

"Structural Irrationality", in BRENNAN, G. GOODIN, R. JACKSON, F. and M., SMITH (eds.) *Common Minds*, Oxford: Oxford University Press, forthcoming.

SCHICK, F. "Beyond Utilitarianism", *The Journal of Philosophy*, 68 (1971), pp. 657-666.

SEN, A. K. *Collective Choice and Social Welfare*, San Francisco: Holden-Day, 1970.

On Economic Inequality, 1st ed., Oxford: Oxford University Press, 1973.

"On Weights and Measures: Informational Constraints in Social Welfare Analysis", *Econometrica*, 45 (1977), pp. 53-89.

"Interpersonal Comparisons of Welfare", in BOSKIN, M. J. (ed.) *Economics and Human Welfare: Essays in Honor of Tibor Scitovsky*, New York, Academic Press, 1979a, pp. 183-201.

"Utilitarianism and Welfarism", *The Journal of Philosophy*, 76 (1979b), pp. 463-489.

Commodities and Capabilities, Amsterdam: North-Holland, 1985.

"Capability and Well-being", in NUSSBAUM, M. and A., SEN *The Quality of Life*, Oxford: Oxford University Press, 1993, pp. 30-53.

"The Possibility of Social Choice", *The American Economic Review*, 89 (1999), pp. 349-378.

SOBER, E. "What is the Problem of Simplicity?" in ZELLNER, A., KEUZENKAMP, H. and MCALEER, M. (eds.) *Simplicity, Inference and Modelling: Keeping It Sophisticatedly Simple*, Cambridge: Cambridge University Press, 2001, pp 13-31.

"Parsimony," in SARKAR, S. and PFEIFER, J. (eds.) *The Philosophy of Science. An Encyclopedia,* New York: Routledge, 2003.

STERN, R. "The Goal of Transcendental Arguments," in STERN, R. (ed.), *Transcendental Arguments: Problems and Prospects*, Oxford: Oxford University Press, 1999.

STICH, S. and S., NICHOLS "Folk Psychology: Simulation or Tacit Theory?", *Mind and Language*, 7 (1992), pp. 35-71.

"Second Thoughts on Simulation", in DAVIES, M. and T., STONE (eds.) *Folk Psychology: The Theory of Mind Debate*, Oxford: Blackwell, 1995, pp. 87-108.

"How do minds understand minds? Mental simulation versus tacit theory", in STICH, S. *Deconstructing the Mind*, Oxford: Oxford University Press, 1996, pp. 136–167.

"Cognitive Penetrability, Rationality and Restricted Simulation", *Mind and Language*, 12 (1997), pp. 297-326.

STIGLER, G. J. "The Development of Utility Theory. I", *The Journal of Political Economy*, 58, (1950a), pp. 307-327.

"The Development of Utility Theory. II", *The Journal of Political Economy*, 58, (1950b), pp. 373-396.

STROUD, B. *Understanding Human Knowledge*, Oxford: Oxford University Press, 2002.

SUZUMURA, K. "Interpersonal Comparisons of the Extended Sympathy Type and the Possibility of Social Choice", in ARROW, K. J., SEN, A. and K., SUZUMURA, *Social Choice Re-examined*, Vol. 2, London: Macmillan, 1996, pp. 202-209.

VON NEUMANN, J. and O., MORGENSTERN, *Theory of Games and Economic Behavior* Princeton: Princeton University Press, 1944.

WALDNER, I. "The Empirical Meaningfulness of Interpersonal Utility Comparisons", *The Journal of Philosophy*, 4 (1972), pp. 87-103.

WEINTRAUB, R. "Do Utility Comparisons Pose a Problem?", *Philosophical Studies*, 92 (1998), pp. 307-319.

WEIRICH, P. "Interpersonal Utility in Principles of Social Choice", *Erkenntnis*, 21 (1984), pp. 295-317.

WELLMAN, H. *The Child's Theory of Mind*, Cambridge, Mass.: MIT Press, 1990.

WEYMARK, T. "A reconsideration of the Harsanyi-Sen debate on utilitarianism", in ELSTER, J & J., ROEMER (eds.) *Interpersonal Comparisons of Well-Being*, Cambridge: CUP, 1991.

"Measurement theory and the foundations of utilitarianism", *Social Choice and Welfare*, 25 (2005), pp. 527-555.

WIGGINS, D. "Incommensurability: Four Proposals", in CHANG, R. (ed.) *Incommensurability, Incomparability, and Practical Reason*, Cambridge, Mass., Harvard University Press, 1997, pp. 52-66.

WILLIAMS, B. "Internal and External Reasons", in WILLIAMS, B. *Moral Luck*, Cambridge: Cambridge University Press, 1981.

YASGUR, S. PhD Dissertation, [2008].