Rational Causes

The Concept of Preference in the Social Sciences

Till Grüne

London School of Economics

Ph.D. Thesis

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

Von dem, was der Mensch sein sollte, wissen auch die Besten nichts Zuverlässiges. Von dem, was er ist, kann man aus jedem etwas lernen.

As to what man should be, even the wisest know little of certainty. As to what he is, it is from the behaviour of everybody that he learns.

Georg Christoph Lichtenberg

Abstract

The concept of preference is used in the social sciences to explain and predict behaviour. This thesis investigates the conditions the preference concept has to satisfy in order to operate as *explanans*. First, it defends the naturalistic position that preferences are causes of behaviour. More specifically, it is argued that preferences are programming properties that are themselves not causally efficacious, but causally relevant in that they realise efficacious properties. Further, the argument that the allegedly intentional nature of preferences poses a problem to such a causal relevance is rejected. Second, methodologies of preference attribution are discussed. The methodology of introspection in its current form is rejected, as well as the Radical Behaviourists' proposal to avoid mental properties altogether. Instead, it is argued that preferences are theoretical concepts. Third, a framework is provided that connects preferences over prospects of different degrees of abstraction. Such a framework allows to attribute specific preferences on the basis of observed actions and derive from these specific preferences more abstract preferences which are employed in the explanation and prediction of behaviour. Fourth, this thesis develops a model of preference change. It is specified under which conditions the inconsistency of an agent's behaviour with the preferences previously assigned to her should be interpreted as a preference change. The model then takes those behavioural observations and predicts how the preferences must have been changed in order to retain consistency. Principles guiding such a change are specified and operationalised, and the ensuing model is compared to existing ones.

Contents

List of Figures

Introduction

The concept of preference enjoys widespread use in the social sciences. It is found in most disciplines of microeconomics, in particular consumer choice theory and welfare economics, in decision theory, game theory, and rational choice theory; the notion as well appears in the empirical methods of sociology and in some anthropological investigations. Preference is one of the central notions of the social sciences.

Despite this widespread usage in the sciences, it is subject to conceptual controversy. The lack of conceptual clarity is detectable in textbooks and lectures, in research seminars and in professional publications alike. In this thesis, I wish to discuss four areas where I find confusion and disagreement prevalent. First, there are ontological issues like the causal relevance of preferences and their relation to neurophysiological or physical properties. Second, there are methodological questions about the attribution of preferences to agents; whether introspection can or must be used, whether preferences can be derived from behaviour and whether the constraints imposed on them make any theory employing the preference notion a normative discipline. Third, once preferences are attributed by a specific method, the issue arises how different types of preference are interrelated. Preferences might be between different concrete options the agent faces, or between highly abstract aspects. Which sorts of restrictions and consistency requirements hold between these different preference types is a very controversial issue. Fourth, there is the issue of preference dynamics. Preferences are attributed to an agent at a particular point in time. To employ them at later times to explain or predict the agent's actions requires an assumption about the preferences' behaviour over time. These four issues will be discussed in the following four chapters.

In chapter one, I develop and defend a naturalistic notion of mental properties that are identified by their causal role in bringing about behaviour. I argue firstly that facts can be the relata of a causal relation and that hence the realization of a mental property can be such a relation between facts. Further, I presented a concept of mental properties as internal factors interacting with external factors in the production of behaviour. The question then arises how mental properties can be causes at all. I argue that they are not causally efficacious, but causally relevant in that they realise efficacious properties. By defending this position, I reject both type-reductionism and epiphenomenalism, while maintaining a naturalistic understanding of mental properties.

One of the most fundamental anti-naturalistic attacks on such a position is the claim that mental properties are intentional properties, and that their semantic content makes them different in kind from physical properties. If this attack was valid, then the social sciences would face fundamentally distinct objects and had to employ fundamentally distinct methods. I defend the naturalistic position by drawing an analogy between measurement in physics and measuring mental properties, and by showing how the propositional content of mental properties arises in their measurement, without any significance for their ontological status. What this chapter establishes is that the *explanandum* of the social sciences – i.e. human behaviour and, derived from that, the emergence and persistence of social institutions – is the effect of the *explanans* – i.e. mental properties of the agent.

In chapter two, I discuss how the *explanans* as the cause of the *explanandum* is identified – that is, how mental properties are attributed to agents. I argue that a methodology that identifies mental properties as causes of behaviour through a methodology of constant conjunctions is not sufficient for the social sciences. In particular, I will criticise introspection as a methodology in this direction as insufficient. Further, I will caution against a wholesome rejection of mental properties as explanans, as scientific behaviourists did in response to the shortcomings of introspective psychology.

Instead, I will argue that social science methodology has to attribute mental properties as theoretical terms. Frank Ramsey's and David Lewis' discussion of how theoretical terms acquire meaning

will serve as a basis here. However, in contrast to Lewis' account, I will caution against the employment of 'folk psychological platitudes' to fill such a theory with content. In its place, I suggest to adopt a particular theory only as a conjecture. This raises the problem how to select any candidate from the infinite set of possible conjectures. Contrary to some claims, I will dispute any *apriori* arguments, and admit that mental properties are epistemically indeterminate.

In chapter three, I discuss how different types of preference are interrelated. Preferences can be between different concrete options the agent faces (which I call *worlds*), and between highly abstract aspects (which I call *prospects*). Preferences over worlds, I will argue, are the only ones that can be derived from observed behaviour; but preferences over prospects are necessary for the explanation and prediction of behaviour. If the outcomes over which preferences are defined are too specific, then the theory is empty – one cannot explain any situations with it that are not exact repetitions of past choices. Thus a *principle of equivalence* is needed that connects preferences over worlds with preferences over prospects. I will provide such a principle that is based on a model of causal beliefs, and that I think is acceptable on the basis of very weak plausibility considerations.

The fourth chapter offers a model of preference change. It is constructed following the general structure of models of epistemic change. It distinguishes between the externally caused direct change of a single preference, and the collateral change that the system has to undergo in order to accommodate the direct change and remain consistent. The satisfaction of these two principles of accommodation and consistency still allow for a multitude of possible results, which need to be narrowed down by additional principles. The two further principles discussed here are that of conservatism and of entrenchment. Three preference change operators are constructed that satisfy some important properties.

The concept of preferences discussed in this thesis is specifically geared to its function in the explanation and prediction of behaviour. It is this function for which preferences are most relevant in the

social sciences. But preferences are used for other purposes as well – in models of practical reasoning, and in theories of individual and collective welfare. I have deliberately restricted myself to not include these functions in one and the same theory of preference. To the contrary: the general conviction I express in this thesis is that past attempts at a unified theory of preference have led to many of the confusions still prevalent in the social sciences; and that therefore we should not strive for one theory of preference, but for many theories of preference, conditional on their functions. This thesis, then, strives to provide an account of preferences in their explanatory and predictive function.

Chapter 1

Mental Properties

1.1 Introduction

In a paper that 'opened the eyes of so many of us',[1] Donald Davidson argued that action explanations require not only reasons but also causes. The reasons for an action are the mental properties attributed to an agent through a successful rationalization. To yield an explanation, such a rationalization must first of all be true – the agent must indeed hold those reasons – and secondly, the action must appear reasonable in the light of the reasons attributed. Those are plausible conditions, but the question is whether they are sufficient for explanation – a question Davidson clearly denies:

> Something essential has clearly been left out, for a person can have a reason for an action, and perform the action, and yet this reason not be the reason why he did it. Central to the relation between a reason and an action it explains is the idea that the agent performed the action because he had the reason. Of course we can include this idea too in justification; but then the notion of justification becomes as dark as the notion of reason until we can account for the force of that 'because'. (Davidson 1963, 9)

Let's imagine Hans who is visiting his grandmother. Hans is generally not a very altruistic chap, and we know that what he

[1] Isaac Levi introducing the author in a symposium on Arthur Danto's work NYC XX.09.2002

loves most about his grandmother is her money. So, in general, Hans' occasional visits to his grandmother were explained (both by him, if he chose to be honest, and by those who knew him) by his desire to win her favour and hence be placed prominently in her will. But on one earlier occasion, Hans had reported that – to his own surprise – he went solely out of fondness for the old lady.

Now, how do we explain Hans' current visit? Why did he go this time? It could be out of greed, or out of love, or both – but rationalization cannot tell us which of the three it was. All of them are plausible – Hans has affection both for his grandmother and for her money – and both times the action is reasonable in the light of the reasons given. But it might nevertheless be the case that Hans acted out of *one* reason only, and hence rationalization cannot be a sufficient condition for action explanation. What is needed is an account of how reasons have a causal effect, and how these reasons which are also causes provide a sufficient explanation of action. This is what this chapter tries to do.

how do reasons become causes?

However, this chapter will not follow Davidson's argument in asking how reasons can *also* be causes. Davidson's approach is to take mental properties primarily as reasons, which are *by their nature* under-determined by physical properties, and which *by their nature* have to be identified by interpretation. This leads to a fundamental inconsistency with his objective of making reasons also causes (Antony 1989).

Instead, in this chapter I will develop a notion of mental properties that does not rely much on the notion of reason. Mental properties - beliefs and desires – are identified by the causal role they play in the production of behaviour, and they have this role because they pick out a physical property by which they are realised. I then argue why they can be realised by physical properties despite the fact that these mental properties are usually depicted as having semantic content. Given this, I argue that mental properties are causally relevant for action – relevant in the sense that they pick out the causes of behaviour. And that should be sufficient for explanation: we learn something about the causal history of an action if we are provided with a mental property of the agent.

1.2 The Concept of Causation

Mental properties in this chapter will be defined in terms of causal roles. Thus, their instantiation (or at least the instantiation of a property that they realise) will feature as the relata of causation. This requires a specific understanding of causation – namely, whether the relata of causation are events or facts. Events are particulars like persons or things. They can be described as individuals instantiating certain properties, but no description will ever fully exhaust them; they have a 'secret life' (Steward 1997) beyond any description. Facts, on the other hand, function as truth-makers of propositions or sentences. They are as Ramsey puts it, 'existential propositions, asserting the existence of an event of a certain sort' (Ramsey 1927). Facts report the instantiation of properties, but they do not themselves instantiate properties. They are different from events, because they are not particulars. So even though the event 'the death of Caesar' might be close to the fact 'that Caesar died', it is easy to distinguish the two by ascribing them a further property and checking for their identity: while 'Caesar's death' might have been bloody, treacherous or a deed of dedication to republicanism, and still remain the same event, the fact 'that Caesar died' can have none of these properties, and the fact 'that Caesar dies bloodily' is a different fact altogether.

Thus the distinction between facts and events assumes importance for the identification of mental properties in terms of their characteristic causal role. As it will become clear in this chapter, any specific mental property will function as a necessary element of a jointly sufficient but non-necessary cause. To have a craving for sweets, e.g., is part of sufficient cause of eating the chocolate bar in front of you – other ingredients of that sufficient cause will be the belief that chocolate is sweet, that the brown object in front of you is chocolate, as well as the absence of commitments to a low-calorie diet or to asceticism. It will not be, however, the only sufficient cause of eating the chocolate – a desire to take it away from your siblings, and the appropriate beliefs, would be another. Understanding causes in this Millian way provides the explanation of action on the basis of mental properties with an enormous flexibility. Mental properties are attributed holistically, as I will discuss in chapter 2. But once attributed, they can be individuated, and

then different combinations of these ascribed properties will render different explanations or predictions.

For this understanding of mental causation, nevertheless, causation must be of facts. If causation were between events, the cause of an action would be one and only one event each time – a particular that instantiated various properties. The non-analysability of events then would therefore prevent the possibility of a combinatorial framework of mental properties to sufficient causes of actions, and thus would greatly reduce the explanatory power of the whole project.

There has been considerable opposition to the position that causation must be of facts. The most widely cited one is Davidson's employment of the famous 'slingshot argument', originally designed to show that there cannot be more than one fact, to the question of the relata of causation. Davidson's argument that causation cannot be between facts proceeds with three assumptions: (a) that the connective 'causes' in the sentence 'C causes E', where C and E are facts, cannot be truth-functional. If '· causes ·' were truth-functional, then the substitution of a relatum by another one with the same truth-value would preserve the truth of the sentence. For example, let 'the streets are wet today because it rained' be a true sentence. Further, it is true that whenever it rains in London, then individual transport increases by 20%. Thus, if '· causes ·' was truth-functional, 'it rained' would be substitutable by 'individual transport increased by 20%' *salva veritate*. But this would yield the sentence 'the streets are wet today because individual transport increased by 20%', which might well be wrong. Thus '· causes ·' is not truth-functional.

However, it is plausible to assume that (b) a logically equivalent sentence can be substituted for a sentence flanking the 'because' *salva veritate*. Logical equivalence of two sentences means that they have the same truth-value in *all* models. For example, we can use the substitution 'the streets are wet because the heavens opened' *salva veritate* as long as 'the heavens opened' is true in every instance in which it rains and vice versa. Last, it seems plausible that (c) the truth of a sentence 'a cause b' is preserved under substitution of coextensive singular terms. In Davidson's own words:

> If Smith's death was caused by the fall from the ladder and Smith was the first man to land on the moon, then

the fall from the ladder was the cause of the death of the first man to land on the moon. (Davidson 1967, 152)

The slingshot argument itself goes as follows. If causation is between facts, then the sentential relatum C in 'C causes E' is logically equivalent to the sentence 'all x such that $x = x$ are *identical* to all x such that $x = x$ and C (formally: all $x : (x = x) =$ all $x : (x = x \wedge C)$), as it retains its truth value in all and only those cases where C is true. Then, according to (b), C in 'C causes E' can be substituted by this logically equivalent sentence such that:

[all $x : (x = x) =$ all $x : (x = x \wedge C)$] causes E

But according to assumption (c), we can now replace the singular term 'all $x : (x = x \wedge C)$' with the singular term 'all $x : (x = x \wedge X)$', because they refer to the same objects: those objects which are identical to themselves. Neither C nor X are quantificationally bound by x and can thus freely exchanged in the term without changing the reference. Thus our sentence can be changed to

[all $x : (x = x) =$ all $x : (x = x \wedge X)$] causes E

But then by assumption (b) it follows from the above sentence that 'X causes E'. Any sentence that has (contingently) the same truth value as C can by this method replace C in 'C causes E'. Thus it seems that the 'causes' in 'C causes E' after all *is* truth-functional, which contradicts assumption (a). Given the correctness of all three assumptions, Davidson sees as the only solution to this paradox to conclude that causation cannot be between facts.

As Mellor (1987) has demonstrated, we do not have to accept the conclusion, as the assumptions can be shown to be wrong. Mellor focuses on Davidson's assumption (c), and points out a group of cases, where the substitution of one singular term by another co-extensive one does falsify a causal statement.[2] This group consists of facts that state a contingent identity, as 'the F = the G', where one of these singular terms is replaced by the other one, thus creating a necessary identity. Such a necessary identity then turns out to falsify the causal statement, as necessary facts cannot be either causes or effects. For example, take the statement concerning a mountaineering accident involving more than one climber:

[2]Searle(1995) similarly criticises the slingshot argument as based on false presuppositions, but instead focuses on assumption (b)

Don's fall is the first fall because Don's rope is the weakest rope

Given the truth of this statement, both 'Don's fall is the first fall' and 'Don's rope is the weakest rope' are true. Thus, 'Don's fall' and 'the first fall' are co-extensive events, and 'Don's rope' and 'the weakest rope' refer to the same thing. Hence it follows from (c), that they can be substituted for each other, yielding, amongst others, the following results:

Don's fall is Don's fall because his rope is the weakest

Don's fall is the first fall because his rope is his rope

But these causal statements are clearly false; which falsifies assumption (c). Thus, Davidson's slingshot argument is rejected: causation does not need to be fully truth-functional, when it appears between facts. Causation can therefore be between facts, and it should be, as Mellor further points out. While it seems possible that the sentential relata of 'causes' or 'because' often refer to events, they cannot always do so. Causes are often expressed as *negative facts*, as in: 'he did not die, because he secured himself'. To translate such a sentence into an event requires the existence of a non-event, as 'his survival' (or 'his non-death', if you like), which would lead to an awkward ontology. In particular, non-events do not make true all sentences that should be related to them. While 'He dies slowly' and 'He died instantly' both imply 'He died', and hence are made true or false by the event of his death, nothing like that holds for non-events. 'He did not die slowly' and 'He did not die instantly' do not imply 'He did not die', hence they cannot be made true or false by the event of his non-death. It therefore seems preferable – or even metaphysically mandatory – to employ facts instead of events in causal statements.

Having established that causation is between facts, I can now discuss mental properties as composite causes of actions. Further, I can neglect arguments to the effect that mental properties (or their instantiation: mental states) cannot be the causes of behaviour because they are not events (for example, see Hornsby 1993). Even if there is a difference in kind between causation as between events and between facts (as Steward 1997 argues), facts (and thus mental states) can be causes of behaviour.

1.3 Internal Factors and System Behaviour

In this section I will present a concept of mental properties that bears similarities to Dretske's (1988) account. The first step will be to specify what an internal factor of a system is, on the basis of a notion of a system and its behaviour, and then build the concept of mental property from there. I start with the claim that a system *behaves* when a factor *internal* to the system contributed to a change in some of its *systemic* properties.

Any collection of components that is perceived as producing a unified outcome of some sort is a system. A property is *systemic* if it is attributed to the whole system, not only to a part of it. For example, the position of her index finger is a property of an agent, not to her finger alone, as it specifies the position of the finger in relation to the agent's physical appearance in its entirety; the liquid content in one of the finger's cells, on the other hand, is a property usually attributed to the cell, not to the agent. Therefore, bending the index finger is a candidate for the agent's behaviour, while osmosis on the cell's wall is generally not. The attribution of properties to a system or to its parts of course leaves many borderline cases ambiguous – internal organs, for example, can to some extent be said to have a life of their own – but a crude distinction will suffice here.

? _ Don't we need voluntariness / conscious control?

A property internal to the system is considered an *internal factor*, if it contributes to a change in one or more of the system's properties distinct from it. Systems of sufficient complexity *behave* if they exhibit property changes that are caused by at least one internal factor. Internal here means that the factor in question is the system or any part of it. Growing fingernails, sweating, losing hair or getting pregnant is therefore human behaviour, just as is writing a poem, speaking to one's child or making love. Similarly, machines and plants do things: jets fly to Bangkok, vacuum cleaners pick up dust, and flowers bloom and shed their leaves. What are excluded are systemic property changes that are caused by external factors only. Having one's hair ripped out, falling down the stairs or being flown to Bangkok thus does not fall under the category of behaviour.

N.B.

The internal/external divide of course depends on the definition of the borders of a system and is therefore crude. For example, certain things happen inside of a system and yet are not internal

to it, like the aneurysm in a human brain or the electrode (with micro-battery attached) in a rat's spine. Further, many systems are too primitive to really have any internal/external divide. For example, a feather describes a series of very complex movements in the wind, which are partly determined by its mass and shape. When we speak of its mass and its shape as its internal factors and thus of the feather's behaviour, we don't mean to say that the feather *does* anything. The internal/external divide in these cases just breaks down due to lack of complexity in the system:

> There is no difference, as far as I can tell, between what happens to an electron in a magnetic field and what an electron *does* in a magnetic field. There definitely is a difference between what happens to an animal placed in water and what it does when placed in water. (Dretske 1988, 11, his italics)

Albeit behaviour is characterised by at least one internal factor causing a systemic property change, the behaviour is not brought about solely by those internal factors. Internal factors are, as discussed above, individually necessary elements of a jointly sufficient but not necessary cause. In many cases, external factors are further elements of the joint cause. Two types of external factors can be distinguished. First, *triggering* factors cause behaviour by affecting internal factors which in turn bring about behaviour. Second, *facilitating* factors cause behaviour by accomplishing the causal efficacy of internal factors. An example for a triggering factor is the tap just below a person's uppermost knee: it causes a complex chain of internal factor changes (neurons firing, muscles contracting) which then causes the person to jerk her leg. An example for a facilitating factor is the presence of air, allowing a person to speak. While in a vacuum, the speaker would perform a series of internal factor changes: vocal cord contractions, mouth movements, etc., without any speech resulting. Only the presence of air allows the transmission of sound waves and thus of speech.

Behaviour is thus an event characterised by a change of systemic properties caused by at least one internal factor. The *description* of behaviour varies with the degree to which external facilitating factors are included. At one extreme, behaviour can be described with internal factors only: as internal factors causing other internal

factors to change. Provided the system is intact, such a description is always possible. For example, we can describe speech behaviour without invoking any facilitating factors like the presence of air: instead of describing the production of sound-waves, we can describe the vocal cord contractions, breathing rhythm, tongue movements and mouth positions that would produce speech under appropriate conditions.

However, descriptions of this sort are found in common parlance only for relatively simple kinds of behaviour: sweating, smiling or farting are examples. Most behaviour descriptions of everyday language aim at a level that requires external facilitating factors for the behaviour's realization. Like a Russian doll, the description of behaviour consists of many layers, according to the degree to which external facilitating factors are included. At the core is the description by purely internal factors; at the outside is a description that hardly seems to be behaviour of a single agent at all. This is illustrated in the following series of descriptions of the same behaviour: her neurons fired, her muscles contracted; her vocal strings vibrated, her mouth opened; she uttered a number of meaningful sounds in front of her; she declared war on the neighbours; she started the conflict; she unleashed a nuclear holocaust; she brought about the end of the human race.

Most behaviour descriptions lie between those extremes. They require the presence of some external facilitating factors, but at the same time try to ensure that at least one internal factor of the system is still a necessary element – if it was not, the phenomenon could not be described as behaviour of that system.

External descriptions of behaviour (i.e. those that require external facilitating factors) are broad and less precise than the small-scale internal descriptions, as they cover the many possible ways in which those external effects can be brought about by internal means. 'She boarded a plane', for example, does not specify in precisely what ways the passenger observed her environment, or how exactly she moved her legs. Instead, the external description covers all the internal factors which are appropriate for the external effect to be reached – external behaviour descriptions therefore often have a teleological form.

The teleological description of behaviour carries the combined effect of internal and external factors beyond the boundaries of

the system in question. Instead of describing the complex property changes internal of the system, it describes the goal at which these changes are aimed. We don't need to say that Fritz registered a static shadow on his retina of a particular size and a particular angle of aberration between his left and right retina, that he raised his stretched arm to the height of his eyes, and that he bent his finger – we can say that Fritz shot Franz with a gun. The killing, of course, is contingent upon a number of external causes – the object Frits holding in his hand being a pistol, the cartridge being live, no obstacle between Fritz and Franz, Franz' position in relation to his shadow – all of these external causal relation are assumed in the description.

The internal factors, however, often get associated with the goal of the behaviour, even if one or more of the external factors are not present. In this case people often say that she tried to speak, board the plane, etc. signifying that while the internal factors were active, the absence of some external facilitating factor prevented the realization of the action. (Absence of the right internal factors in relation to 'trying', 'attempting' is a more difficult notion which I will not discuss here).

We now have all the building blocks we need for mental properties. Human beings are usually able to identify what is human or animal behaviour or not – the absence of external causes that could fully explain some human or animal locomotion lets them conclude that some internal factor must have contributed, and hence that they witness behaviour. The internal factor(s) at work then can be identified according to the specific behaviour that they help producing: they are identified by their causal role. But external factors also contribute – hence internal factors are identified according to their causal role, conditional on certain external triggering and facilitating factors. The interrelations between internal factors, triggering and facilitating external factors and behaviour is illustrated in figure 1.1.

One can think of properties thus individuated as dispositions to behave.[3]

[3]I have said earlier that behaviour includes a very broad range of phenomena; many of which we would not usually include under actions. However, in the following, I will use behaviour loosely for those phenomena that we often categorize as actions. The attempt to distinguish actions from behaviour more general, by identifying the former as guided by intentional attitudes, is not helpful, I think. The way I have discussed mental properties (whose intentional quality yet remains to be investigated) shows that we start with the identification

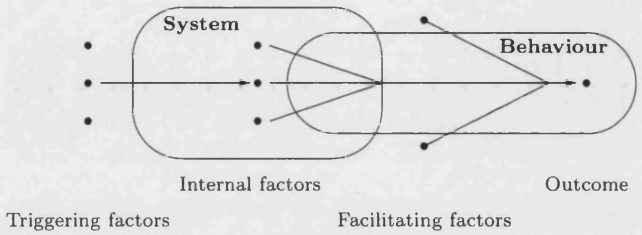

Figure 1.1: Causes of Behaviour

Each behaviour observation then leads to the attribution of a specific internal disposition that causes the observed behaviour in interaction with facilitating factors. However, complexities arise when under identical external factors, different behaviour is exhibited at different times. Under these conditions, it is economical to compartmentalise internal factors further. Instead of ascribing one internal factor for each behaviour, a group of internal factors is identified as the individually insufficient but necessary elements of a jointly sufficient but non-necessary cluster of internal and external facilitating factors.[4] The individuation of these internal factors is guided by consideration of *simplicity* and *comprehensiveness*: as many kinds of behaviour should be covered with as few as possible internal factors. Additionally, the individuation is influenced by the number of triggering factors, and by the number of different behaviours under identical facilitating factors. A more comprehensive discussion of the individuation of internal factor, and in particular of the standard practice to distinguish desires and beliefs, will follow in chapter 2.

These remarks about individuation of internal factors could be read as a plea for a purely instrumental notion of internal factors. I disagree with such a reading. What is needed here is a notion

of behaviour and then identify the internal factors that cause it. However we do this, the border between action and non-action behaviour will remain fuzzy, with many phenomena falling in both categories. Instead of attempting such a distinction, I will continue using 'behaviour' *simpliciter* and implicitly assume that we are talking about sufficiently complex behaviour that merits an explanation with reference to mental properties.

[4]The discussion of causal factors as INUS conditions goes back to Mackie (1980). However, I do not intend to employ the INUS concept as a reduction of causes, but rather as their representation.

of internal factors (as behavioural disposition) which are causally relevant for behaviour. This requirements puts certain constrains on internal factors that render them with ontological substance.

> To speak of an object having a dispositional property entails that the object is in some non-dispositional state or that it has some property which is responsible for the object manifesting certain behaviour in certain circumstances. (Armstrong 1968, 86)

This raises the question of how functional (dispositional) properties are realised, and the related issue of how the identification of functional properties can in any way provide a causal history that explains the agent's behaviour. This is the topic of the next section.

1.4 Mental Causation

Mental properties, as defined above, are distinct from physical properties, in the sense that they are not part of the terminology of physics. Nevertheless, mental properties are employed to individuate the causes of behaviour – hence mental properties, albeit distinct from physical properties, are claimed to witness physical effects. At the face of it, this distinction collides with the assumption that physics is complete; an assumption which claims that:

> All physical events are determined, or have their chances determined, by prior physical events according to physical laws. (Papineau 1993, 16)

If this assumption were true, then all the causes of behaviour must be describable as physical properties. Under this conclusion, it seems wrong to claim that instantiations of mental properties cause behaviour; unless one also claims that mental properties are nothing but physical properties, or that behaviour suffers constant overdetermination from two types of causes. The dilemma the naturalist philosopher of mind finds herself in is expressed by Moser:

> Either the domain of the mental will be causally impotent. . . . or physical categories will not have a monopoly on causal explainers. (Moser 1994, 21)

If one wants to remain within a naturalist position – which assumes that behaviour has a physical component, that physics is complete and that there are no regular cases of overdetermination – then one cannot admit Moser's second option. Instead, I will grasp the first horn of Moser's dilemma.

To do so without falling either for epiphenomenalism or reductionism, I have to argue that mental states can be causally relevant to behaviour without being causally efficacious, and that therefore mental properties can be meaningfully differentiated from physical properties without becoming causal competitors. I will first argue for a claim that is already implicit in the functional definition of mental properties given above: that mental properties are supervenient on physical properties. I will then limit that claim so that type reductionism of the mental to the physical is excluded; and last I will argue for a realization relation between the mental and the physical such that mental properties do come out as more than mere effects of physical properties, and thus reject an epiphenomenal conclusion. This, I hope, will suffice to see that mental properties can be considered *causally relevant* for behaviour.

According to the functionalist approach endorsed here, to have a mental property means having a physical property that typically occupies a particular causal role, given certain sensory inputs and other mental properties. The mental supervenes on the physical in the sense that there cannot be a difference between the mental properties of a system without there being a difference between the physical properties of a system. As the functionalist account presupposes that the mental is realised through the physical, and this presupposition entails supervenience, the functionalist therefore presupposes supervenience.

The supervenience of the mental on the physical derives from two premises: the *completeness of physics*, and the so-called *manifestability of the mental*.[5] This latter premise assumes that if two systems are different in some mental property, there must be a possibility of this difference manifesting itself in some physical property *inside* of the systems.[6] Roughly, mental differences must be in principle

[5] Compare McGinn 1982: 29; Papineau 1993, 18.

[6] This claim is stronger than Papineau's, who allows for manifestation outside of the system. This is necessary for his account of the broad content of propositional attitudes. For my differing treatment of broad content, see section 3.

detectable in physical terminology, while magical powers and supernatural sensory abilities (divine revelation, telepathy) are excluded.

From completeness follows that identical physical causes yield identical physical effects; and from manifestation follows that mental difference requires physical difference. Thus, completeness and manifestation together imply that physical identity prohibits mental differences, and hence imply supervenience of the mental on the physical.

Supervenience, however, is too wide a notion to cope with mental causation adequately. After all, both the epiphenomenalist and the reductionist approach to mental states are compatible with supervenience of the mental. The epiphenomenalist on the one hand claims that the physical states which cause behaviour also cause the respective mental state, while this state itself does not cause anything. As the mental state according to this view is an effect of the physical state (or at least of the collection of all physical states of the system) it can change only with its causes, and hence supervenes on them. The reductionist on the other hand claims the identity between physical and mental properties (state *types*), which trivially implies supervenience. Hence in both views supervenience is satisfied, but mental states are causally irrelevant.

The reductionist argues for the identity of mental and physical properties. If a mental property M is identical to a physical property P, it follows that for any instantiation of M through an object o, the identical physical property is instantiated; i.e. from the identity of P and M follows $Mo \Leftrightarrow Po$. Such a biconditional then serves as a bridge law for a reduction of any theory involving property M to physics. But there is a problem. Motivations are ascribed to a range of living creatures, many of which do not share the same brain structure as humans. People say that horses want to flee fire; that dogs want to signal to their masters and that frogs want to still their hunger. If type identity held, however, for each different brain structure a different mental state would apply. Assuming that humans, horses and frogs differ in their brain structure, a horse's desire to escape a fire would be a different mental property altogether than a human's desire to do the same. Furthermore, fictional creatures like the 'Terminator' of recent Hollywood productions are made out of entirely different materials, but their environment treats them like

fellow creatures and assigns similar mental properties to them as to humans. If one imagined being confronted with such creatures, whether constructed by humans, or arriving from a different time or planet, would it be wrong to attempt such attributions in the face of behaviour structurally similar to humans? At the least, it is not clear what purpose such a differentiation would serve but the upholding of type physicalism.[7] Finally, beyond the differences in species or between terrestrial and extraterrestrial, it is not even clear that human beings instantiate their mental states in any uniform manner. So far, neurophysiology has identified little more than areas of the brain that show more activity when the agent is aroused, angered or concentrates on a given task. But whether identical physical processes take place in the agent's brain whenever the agent is in a specific mental state is by no means uncontroversial.

It therefore seems plausible to admit that mental states can be realised in a multitude of physical ways. According to this multiple realisability assumption, the relation between instantiations of M and P are weaker than reductionists assume. All that has to hold is that any instantiation of a mental property necessitates a physical property: $Mo \Rightarrow \exists P_i : P_i o$. The reductionists' biconditional, on the other hand, does not hold for any P_i, and hence the identity between any P_i and M must be false.

Papineau (1993, 38), however, doubts the credibility of multiple realisability. If functionally identified mental properties could be realised in different ways, he says, it would be incredible that they all led to the same behavioural output:

> I am not accusing the functionalist picture of inconsistency, but only of incredibility. The difficulty I am concerned with arises when some mental state S, which mediates between physical input R and physical output T, is realised by a range of different physical states P_i. The puzzle is: why do all the different P_is which result from R all nevertheless yield the common effect T? Now, it is possible that every such P_i should just happen to yield T . . . . However, if this were so, it would be the kind of coincidence that cries out for explanation. (Papineau 1993, 35-36)

[7] Compare Putnam 1967; Fodor 1974.

This argument seems puzzling – after all, all the different P_is yield the same effect T because they were individuated by their potential to yield T. All the P_is are simply of the same *psychological type.* That this unity does not establish the unity on the physical level – that a psychological type does not necessarily establish a physical type – is exactly the point of autonomous mental property explanation that I want to defend here.

Naturalism that rejects the possibility of multiple realisation Papineau-style begs the question of what constitutes mental properties. It presupposes without justification that it is the physical level that determines the unifying and distinguishing criteria for mental states – in other words, it requires mental properties to be natural kinds. Only on this basis can people conclude that certain mental properties are only superficially similar, while being distinguished in their deep (physical) structure:

> . . . in spite of their superficial similarity when viewed top-down, human psychology and Martian psychology are independent of each other in causal/explanatory structure and sources of evidence and must be considered, to all intends and purposes, distinct sciences. (Kim 1998: 122)

The line of attack against this argument is to point out the severe limitations of requiring mental properties to be natural kinds. Behaviour is an event of purely physical manifestation; however, the possibility of describing this event in merely physical terms, derived from this point is largely theoretical. When we want to explain why the agent reached out for the ice-cream, or why the car crash occurred, we talk about events which themselves do not have the unity of natural kinds.

> Someone equipped only with the vocabulary and explanatory resources of physics would not hit either on car crashes or on reachings for ice-creams as types of physical event which required uniform physical explanations. And so once again, I am left wondering what the puzzle is supposed to be. In so far as there is a question, it is a question that cannot even be posed, let alone answered, unless we permit ourselves the resources of a richer vocabulary and explanatory scheme than mere physics is able to offer. (Steward 1996, 672)

To require mental properties to be homogeneous from a physical perspective, as Papineau requires against functionalism, is to eliminate exactly those resources. The moral high ground of the naturalist then seems to stand against the possibility of social science. It is however not necessary to reject naturalism; rather, it suffices to reject the ill-understood concept of natural kinds as a requirement for mental properties.

> It is perfectly consistent to accept the Kripke-Putnam account of semantics of natural kind terms like 'gold' and 'water' while rejecting such an account for mental terms. The notion of a natural kind is not the most luminous of notions; but I do not think that we should be bothered if we are required to say that pains, like poisons and mousetraps, are not a natural kind, and lack a scientifically determinable essence. (Shoemaker 1984, 283)

To the contrary: that mental properties are not natural kinds, but supervene on physical properties, is the justification of the social sciences autonomy from the natural sciences while maintaining a naturalistic position.

A second reductionist response claims that even though mental properties might be multiply realised, it is nevertheless possible to reduce them to the disjunction of these physical properties. The reply of the anti-reductionists here is twofold: first, they doubt that such a disjunction can be considered a real property. This defense is rather weak, as it seems difficult to clarify what exactly a 'real property' is. A second defense is more fruitful. It hinges on the principal openness of the above thought experiments. If mental properties are multiply realisable at all, i.e. if different physical properties can take on the causal role that specifies the mental property, then it is always possible that there are more realisers than we know so far. Under these conditions it is not possible to form a disjunction of physical properties which could serve as a reductive basis – all that such a disjunction could do is to summarise the current state of scientific knowledge.

Thus, despite the fact that mental properties supervene on physical ones, it is impossible to establish a type identity between the mental and the physical – reductionism of mental properties therefore fails.

From mind-body supervenience, however, follows a minimal relation between mental and physical properties. Mental properties are not realised by the same physical property all the time; but each time a mental property is be realised, it must be realised by some physical fact (an instantiation of a physical property). Papineau formulates this realization relation as follows:

> In order for a mental . . . type M to be realised by an instance of some physical type P, M needs to be a second-order property, the property of having some property which satisfies certain requirements R. And then M will be realised by P in some individual X if and only if this instance of P satisfies requirements R. In such a case we can say X satisfies M in virtue of satisfying P. (Papineau 1993, 25)

The desire to drink milk then is the *secondary property* of having a (possibly undetermined) physical property that under specific conditions makes one drink white liquids out of suitably shaped containers. To identify this secondary property, one does not have to determine what physical properties are involved here; it is enough that there are some, and that they are the realisers of the causal role in question.

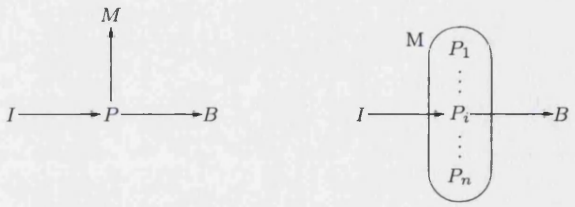

Figure 1.2: Epiphenomenalism vs. Token Identity

A realization relation of the described sort prevents both a possible overdetermination of behaviour, and it prohibits any epiphenomenal conclusion. It prevents overdetermination by chaining M to a causally efficacious physical event, thus excluding the possibility that M and P become causal competitors. It prohibits the epiphenomenal conclusion – not the letter, but its spirit – by admitting that a mental event is not causally efficacious as the epiphenomenalist claims, but nevertheless *causally relevant* in that it realises

a causally efficacious event. The difference between the two interpretations is illustrated in figure 1.2. While the epiphenomenalist claims that the mental property is nothing but a causal by-product of the causally efficacious physical property, the position defended here claims that the mental property is the category of a set of physical properties that all realize the behaviour in question. In this sense, the causally relevant property *programs* for causally efficacious ones in the following sense:

> A useful metaphor for describing the role of the property is to say that its realization programs for the appearance of the productive property and, under a certain description, for the event produced. The analogy is with a computer program which ensures that certain things will happen (. . . though all the work of producing those things goes on at a lower, mechanical level. (Jackson and Pettit 1990a, 114)

Programming properties relevant for an effect, although not involved in the causal chain, provide a causal history of that effect. The epiphenomenalist claim that mental properties are causally inert thus stands undefeated; but the bearing this result has on the possibility of explanation of behaviour with recourse to those mental properties is greatly changed.

> How is a particular psychological set causally relevant to an agent's doing something . . . given that the action is produced (...) – without leaving anything to be explained – by a certain complex of neurophysiological states? The programme model suggests that the psychological set will be causally relevant so far as its realization in an agent of that kind makes it more probable than it would otherwise have been . . . that there will be a neurophysiological configuration present – maybe this, maybe that – that is sufficient to produce the required behaviour. (Pettit 2002, 181)

Mental properties program for the causes of behaviour, and they therefore can be employed in causal explanation of behaviour: By finding out about what the agent wanted, we learned something about the causal history of some events. Additionally, they are –

globally speaking – much better at this type of explanation than anything else so far available. They master multiple realisability, which hampers progress in neurophysiological explanations (and will possibly make it impossible in principle); they allow one to explain behaviour in the terminology in which it is described, and not in the overtly reductive terminology of physics. It therefore allows one to focus on the macrostructure of behaviour, instead of the causal (and thus physical) microstructure. Finally, mental events come equipped with an easy (if somewhat rough) epistemology, as I will discuss in chapter 2.

1.5 Mental Content

Mental properties are often presented as propositional attitudes. In particular, beliefs and desires, the mental properties which are most important for explanations in the social sciences, are considered to have propositional content: we say that the agent believed that it was raining and she desired that she stayed dry when we explain why she took the umbrella. This impression is supported by the formal treatment of mental properties: as the belief that p, $bel(p)$, or the desire that q, $des(q)$, or as the preference relation pPq between propositions p, q.

Many philosophers therefore think of mental properties as relations between people and propositions: my belief that the sun is shining is a special kind of relation between me and the proposition 'The sun is shining'; your desire to stay at home on such a beautiful day like this is a relation between you and the proposition 'I stay home on Mon, 1st of March 2004'. To claim that mental properties are propositional attitudes then becomes an ontological claim: namely that semantic content in the form of a proposition is part of the mental property. Colloquially, as mental properties are internal factors, it is claimed that propositions must be 'in the head'. The claim that mental properties are relations between people and propositions is called the 'Relational Thesis'.[8]

Whatever we think propositions to be exactly, they are semantic content – that is, they are what sentences of the same meaning express. But objects of such a sort do not have either location in

[8] For example in Fodor (1987).

time or space: propositions are abstract objects. The Relation Thesis therefore implies that mental states (as instantiations of mental properties) are very specific kinds of objects; in particular, that they are fundamentally different from physical states. This claim – albeit of course not cast in terms of propositions – can be found in the writings of the 19th century German philosopher and psychologist Franz Brentano.

> Every mental phenomenon is characterised by what the Scholastics . . . called the intentional (...) inexistence of an object, and what we might call (...) reference to a content, direction toward an object (...), or immanent objectivity. Every mental phenomena includes something as object within itself, although they do not all do so in the same way. In presentation something is presented, in judgment something affirmed or denied, in love loved, in hate hated, in desire desired and so on. This intentional in-existence is characteristic exclusively to mental phenomena. No physical phenomenon exhibits anything like that. We can, therefore, define mental phenomena by saying that they are those phenomena which contain an object intentionally within themselves.' (Brentano 1995, 88-89)

Propositions – the abstract objects Brentano calls content of direction toward an object – do not have any physical correlate. The Relational Thesis implies that mental properties are relations between persons and these abstract objects; hence there cannot be a physical or neurophysiological correlate of mental states either. As causation is between physical phenomena only, mental properties, understood as relations between people and propositions, cannot be causes of anything, and in particular not causes of actions.

This is a far-reaching result. If the social sciences deal with mental properties, and the natural sciences with physical phenomena, then it follows from Brentano's distinction that the social sciences and the natural sciences are fundamentally *distinct in their objects.* But if intentional properties are fundamentally different from physical properties, and cannot be causes, then nothing can be explained by referring to them. Thus it would follow that the natural and the social sciences have to be *different in their methods* as well. While

the natural sciences explain, the social sciences have to adopt some other practice, maybe *Verstehen*, maybe something else. The characterization of mental properties as intentional states contributes to the deep-reaching divide between *Geistes-* and *Naturwissenschaften.*

I think that the thus constructed divide between the social and the natural sciences on these grounds is unwarranted. I think that we *can* explain in the social sciences, and that this explanation refers to mental properties as the *causes* of behaviour.

There are two ways to defend such a naturalistic account of mental properties. One can either expand the supervenience-base of the mental – to include all ontological constituents of the representational content – or one can argue that the individuation through the representational content is ontologically irrelevant. That is, one drops the assumption that mental properties have semantic content in any essential way and thus frees oneself from the burden to show how this semantic content could be reducible to physical events.

I think the attempt to extend the supervenience-base is not fruitful. First, there is the question whether it can be done at all. Propositions are abstract objects. Thus it is not clear whether they have physical realisers at all; or if they do, then it is an open question what these physical realisers are. Second, if it can be done, then the result will be an extremely uneconomical ontology. To answer the qualms of the anti-naturalists, one had to point out what parts of the realisers of a mental property were the realisers of the semantic content, a rather dooming task it seems to me.

Instead, I will argue for a naturalistic account of mental properties by denying that intentional states are *genuine* relations between people and abstract objects. Any such argument against the ontological relevance of propositional content has to show how mental properties acquire their semantic content – which they undoubtedly have – without being ontologically committed to them. The argument presented here relies on an analogy between the ascription of semantic content to mental properties and the ascription of numbers to physical objects in (numerical) measurement procedures.[9] To clarify this analogy, I will first rehearse some basics of the theory of measurement.

[9]This idea has been aired by a number of authors without proper elaboration, for example Churchland 1979, 100-107; Davidson 1989, 57; Field 1980, 114; Stalnaker 1984, 7. The only elaboration of this idea that I know of is found in Matthews 1994.

1.5.1 Measurement Theory

Measurement theory explains why and how numbers can be employed to represent a property; more precisely, how a scale – a mathematical structure consisting of a set of numbers and a transformation rule – can be meaningfully applied to a set of objects. A *scale* is an ordered triple $\langle U, N, f \rangle$ with U being an empirical relational system, N a full numerical relational system, and f a function that maps U homomorphically onto a subsystem of N. An empirical relation system consists of a domain specification and one or more relations over that domain; for example, the set of all extended objects O which are ordered according to their weight R_w forms an empirical relational system $\langle O, R_w \rangle$; the set of extended objects ordered according to their weight and their density R_d forms another empirical relation $\langle O, R_w, R_d \rangle$. A full numerical relational system consists of the set of all real numbers and one or more relations $S_1 ... S_n$ defined over it – mathematical relations like, for example, equality or calculus operations. A subsystem of a relational system consists of a subset of the domain of the system with the same relations defined over it. We say that a numerical subsystem preserves properties and relations of an empirical system if the function f maps the elements of the empirical domain $a, b, c \ldots \in O$ onto the numerical domain N in such a way that

$$aR_ib \text{ if and only if } f(a)S_jf(b)$$

If f satisfies this requirement, we speak of N as a homomorphic image of U. To prove that there is such a scale with a homomorphic mapping is to prove a *representation theorem*; its result is that the property or relation that was identified in the empirical relational system is indeed preserved in the numerical subsystem.

'Measuring an object' thus is a two-step procedure. First, the empirical structure of an object is identified: its weight, length or temperature. The empirical structure determines the *dimension* of measurement. For each dimension, the homomorphic image of the empirical structure allows the measurement of *magnitude*. The assignment of numbers to objects is thus only part of the measurement process. Numerical assignments across dimensions are obviously not comparable.

From the existence of a homomorphic image of an empirical system it does not necessarily follow that this numerical subsystem is the only possible representational space. To say that a stone has a weight of 11 pounds, attributes a property to the stone – that of having a certain mass. But we can express exactly the same fact by saying that the stone weighs 5 kilogram, or 175.57 ounces. There is – to some degree – an arbitrary choice of unit in expressing the same fact about the object numerically. The theory of measurement discusses this peculiar relation between objects and their properties on the one hand, and numbers and their operations on the other.

When we express weight in kilos or pounds, we represent a feature of this property in a numerical space. Such a representation preserves a certain structure of that property (the 'greater than' relation, its additivity or substractability, for example), and makes it more convenient to handle. But there are many kinds of representations that all preserve the same structural features of that property: pounds and kilos are just two examples from an infinite set of possible representations. We then say that in these cases the scales measuring this property are *equivalent transformations* of each other.

Where scales are equivalent transformations of each other, the empirical domain stays the same, but the mapping and the numerical subdomain change: let $\langle U, N, f\rangle$ be the Celsius scale and $\langle U, N, g\rangle$ the Fahrenheit scale. Then $g = F \circ f$, that is, g can be generated through a transformation function F from f. We say that $\langle U, N, f\rangle$ is unique up to transformation F, and thus categorise scales according to the kind of transformation function they allow. Absolute scales only allow identity transformation; ratio scales (e.g. weight, length) allow for transformation by a coefficient; interval scales (e.g. temperature) allow for positive linear transformations; ordinal scales allow for monotone transformations.

How is the uniqueness of a scale determined? If all elements of a scale are known, the type of scale can be proven:

> To determine its uniqueness, we need to know the scale, which means we need to know both an empirical relation system and a full numerical relation system. From the knowledge of a scale, we can, at least theoretically, infer precisely what the uniqueness properties of the numerical assignments are. (Suppes and Zinnes 1967, 15)

But the exact structure of the empirical system is not known *a priori*. Instead, the form of the empirical system is conjectured, measurement scales established on that basis and the results of the measurement employed in explanatory or predictive theories. Measurement thus proceeds by trial-and-error: different measurement scales are applied to a system – i.e. measurements are taken – and the measurements are compared for consistency.

The numerical relational subsystems employed in such measurement scales are not necessarily identical to subsets of the real number system. To prove that an empirical system is homomorphic to a numerical one does not mean that the empirical system is homomorphic to the real number system. A real number system contains amongst others the operations of addition and multiplication – that is, three-place relations – as well as the binary relation 'less than'. Relations of empirical systems measured with a numerical system can be considerable weaker, in that they lack e.g. addition or multiplication. For example, the empirical system measured with Moh's hardness scale does not contain any relation corresponding to the addition operation, but only the 'less than'-relation. That a numerical relational system of a given scale is isomorphic to the real number system is therefore an additional result that is not necessarily born out in the representation theorem itself.

Hence there are many relations definable over the numerical domain, which do not have any correspondence to any property in the empirical system. Strictly speaking (as in the above paragraph) these properties or relations are therefore not part of the numerical system which is a homomorphic image of the empirical system in question. This does nevertheless not prohibit the application of those properties and relations to the numerical system, as long as these applications are meaningful:

> This does not mean (...) that manipulations of the numbers in the domain of a given numerical system – to infer facts about the elements in the domain of the corresponding empirical system – must involve only those relations in the given numerical system. Relations neither contained in a given numerical system nor having a direct correspondence in the related empirical system may nevertheless be used. There are, of course, certain limitations imposed upon the manipulations of the numbers of a numerical

> system, but these limitations relate to certain criteria of meaning of individual sentences rather than to those relations contained in a numerical system. (Suppes and Zinnes 1967, 8-9)

Thus we can employ all arithmetic operations on any numerical system in statements about the empirical system, independently of the question whether these correspond to any empirical operations, as long as the operations are meaningful. Suppes and Zinnes (1967, 64) give two examples of this:

> (i) The ratio of the maximum temperature today (t_n) to the maximum temperature yesterday (t_{n-1}) is 1.1.
>
> (ii) The ratio of the difference between today's and yesterday's maximum temperature (t_n and t_{n-1}) to the difference between today's and tomorrow's maximum temperature (t_n and t_{n+1}) will be 0.95.

The truth of statement (i) depends crucially on which of the equivalent scales are used for the measurement of t_n and t_{n-1}. Using the Fahrenheit scale, with t_n=110 and t_{n-1}=100, the statement is true; using the same temperatures measured in Celsius, t_n=43.3 and t_{n-1}=37.8, the ratio is 1.15, hence the statement is false.

The truth of statement (ii), on the other hand, is robust under all positive linear transformation from t_n to $t_n^\star$, as can be easily shown:

$$\frac{t_n - t_{n-1}}{t_n - t_{n+1}} = \frac{(\alpha t_n^\star + \beta) - (\alpha t_{n-1}^\star + \beta)}{(\alpha t_n^\star + \beta) - (\alpha t_{n+1}^\star + \beta)} = \frac{t_n^\star - t_{n-1}^\star}{t_n^\star - t_{n+1}^\star}$$

The meaningfulness of a statement involving operation on the numerical system of a measurement scale thus depends on the form of its admissible transformations, not on the operations defined on the empirical system:

> A numerical statement is meaningful if and only if its truth (or falsity) is constant under admissible scale transformations of any of its numerical assignments, that is, any of its numerical functions expressing the result of measurement. (Suppes and Zinnes 1967, 66)

Meaningfulness, besides representation and uniqueness, thus becomes a third criterion of any measurement scale.

The measurement process takes the relations and properties of the empirical domain as given. In no way does the measurement explain these structures; nor does the measurement require that the measurer is acquainted with them. People measure temperature without understanding thermodynamics; in fact, the measurement of temperature, i.e. the construction of the Fahrenheit and Celsius scales antedated the advent of modern thermodynamics. The successful application of measurement procedures to the contrary often tells us something about the structure of the phenomena measured – but we have to be careful what exactly we can import from the homomorphic numerical system back into the empirical system.

Measurement theory thus clarifies how the representation of specific properties of an empirical domain through numbers is justifiable. It further clarifies that a successful measurement does not necessarily exhaust the empirical system, that many measurement scales are unique only up to admissible transformations in scale, that the representing numbers may have operations defined over them which are not found in the empirical domain, and that the properties represented need not be independently known to perform the measurement.

1.5.2 Measuring Mental Properties

Even in the early discussions of measurement theory, one finds considerations that the fundamentals of the theory – the idea of perseverance of a system's property when represented by another scale - can be applied not only to numbers, but as well to 'persons, attitude statements, or sounds' (Suppes and Zinnes 1967, 7). In this section, I want to argue that the ascription of propositional attitudes is analogous to the ascription of numerical measures.

It first needs to be noted that this analogy is completely distinct from methodological consideration. How something is measured – i.e. what sort of practices are involved – is not the question that measurement theory asks. This methodological question will be addressed in chapter 2. What is discussed here are the preconditions for the possibility of such measurement practices. It focuses on the logic of measurement, but it will provide crucial hints about the ontological nature of what is measured.

The first point of importance here is that numerically measurable

properties, like length or temperature, are not intentional properties in any ontologically relevant sense, despite their formal appearance. Properties represented in numerical measurement are preserved, without the numbers representing those properties having to be part of the property in any way.

> . . . talk of how much things weigh is relational: it relates objects to numbers, and so to one another. But no one supposes the numbers are in any sense intrinsic to the objects that have weight, or are somehow "part" of them. What are basic are certain relations amongst objects; we conveniently keep track of these relations by assigning numbers to the objects and remembering how these relations among the objects are reflected in the numbers. (Davidson 1989, 59)

Formally, a statement of length of a particular object states a relational predicate: the predicate that relates the object with the length indexical. However, nobody claims that physical properties are therefore intentional properties, in the sense that an object which has this property must in any way contain that number. Such a claim would be wholly absurd, as it would render all measurable properties of physical objects non-physical properties, as they would contain an abstract entity – a number. If this was admitted against all odds, properties like length, weight or temperature would not be causally relevant in the naturalistic position defended here.

Further, the existence of equivalent scales makes the idea of weight or length or temperature as a relation between a thing and a number entirely implausible. There just isn't any one number that we could single out as the one the object is related to, and to relate them to infinitely many ones doesn't make sense either. Instead we have to conclude that having a temperature of x degrees is not a genuine relation between objects and numbers at all. Rather, temperature is a property intrinsic to objects, and it is represented in numerical space.

Thus having a weight, length, temperature, etc. are not intentional properties, although they are represented with the help of abstract objects, viz. numbers. The problem that seemingly befalls intentional properties, that they cannot be causes of anything, therefore does not apply to them. And of course it must not, as otherwise physics could not explain.

Why now are attributions of propositional attitudes similar to numerical measurements? I want to discuss two aspects of this alleged analogy. First, in both cases a property is represented by abstract objects. Second, on both cases the numerical scales have equivalent transformations.

The property of weight can be specified as having a particular causal role: maybe the capacity to tip the balance of a scale, or to tire a man or a mule after some hours of travel, etc. This causal role is explicable – in today's physics – as the atomic consistence of the object having this property, and the gravitational forces acting upon it. Nevertheless, we stick with the causal role most of the time: we say that the roll flattened the Tarmac because it had great weight; without any reference to atomic structures or gravitational forces. However, we often need to refine their causal role by specifying that it was heavier than something else – in order to explain, say, that something tipped the balance – or that it was so much heavier than something else – for example to explain the occasion that one object flew x feet further than another object. To express these more specific roles conveniently, we represent the property of weight in numerical space.

The same applies to mental properties relevant in the social sciences. Desires and beliefs are specified by their causal role between sensory input and behavioural output. Roughly speaking, beliefs are shaped by input conditional on other beliefs, and desires shape output conditional on other desires and beliefs. Neurophysiology just begins to give us an inkling how these causal roles are realised in physical terms; but for explanations of behaviour it is enough to cite those beliefs and desires without any reference to their underlying realisers. Of course these causal roles usually need to be refined: beliefs by the conditions that make them true, and desires by the conditions that satisfy them. These conditions, expressed as propositions, then index the properties defined by their causal roles.

From this very simplistic analogy, one can already derive that physical and mental properties alike can be said to be represented by abstract objects. The story for mental properties, however, is slightly more complex than that. I already argued in section 1.3 that mental properties need to be compartmentalised, at least into beliefs and desires. Now, both desires and beliefs are further individualised by the propositions that specify their causal roles: by the truth

conditions of beliefs and the satisfaction conditions of desires. Like particle movement in classical mechanics is represented by direction and magnitude, so does a proposition represent the 'direction' of the mental property. In both cases, nevertheless, it would be a mistake to think of the particle or the mental property as 'containing' the direction or the proposition.

A closer look at the proposition-dimension, however, reveals that propositions alone are not sufficient to represent the dimension of mental properties. Consequently, Matthews (1994) proposes a three-dimensional representation space for mental properties. Each point in this space is represented by the ordered triple $\langle a_i, \langle s_j, r_k \rangle\rangle$, consisting of an attitude type a_i, and what he calls a designated proposition $\langle s_j, r_k \rangle$, where r_k is a Russelian proposition and s_j a sentence type, a token of which in a particular context serves to designate that Russelian proposition.

To use the ordered tuple $\langle a_i, p_i \rangle$ will not yield a sufficiently rich individualisation of mental properties. People often fail to identify facts like 'Honiara is far from London' with 'The capital of the Solomon-Islands is far from London'; they therefore might believe the one without believing the other. Such failures of substitutivity show the need to individuate mental properties finer than propositions alone can do. Nevertheless, it is not possible to simply turn propositional attitudes into sentential attitudes. Sentences do not always specify a proposition uniquely: sentences with *deictic* terms, for example, specify a proposition contingent upon the specific context in which this sentence is uttered. Thus the ordered tuple $\langle a_i, s_j \rangle$ will not be fine grained enough for the necessary individuation of mental properties either: my belief today that this glass is dirty is clearly distinct from my belief yesterday that this glass is dirty, but a sentential representation alone cannot distinguish them.[10] Although neither propositions nor sentences alone suffice for the individuation of mental properties, the combination of the two will. Thus, mental properties are represented in a three dimensional space, with each point designated by the ordered triple $\langle a_i, \langle s_j, r_k \rangle\rangle$. Theories of mental properties that are represented by propositions and sentences can be found, for example, in Davidson's *unified decision theory.*[11]

[10] That sentences are too fine-grained, on the other hand, is not an argument against $\langle a_i, s_j \rangle$, as this overt individuation can be tackled by finding the appropriate transformation function between equivalent scales.

[11] 'A radical theory of decision must include a theory of interpretation and cannot presuppose

But mental properties are not only distinguished in kind, according to propositions and sentences. Like particle movement that consists in a direction and a magnitude, mental property kinds consist in a propositional/sentential dimension and an intensity measure. I will not discuss the measurement of probability and desirability indices much in this thesis, but two things are worth mentioning here. First, it is now obvious that any such numerical magnitude should not be expected to be found 'contained' in the mental property. Second, these numerical measures are based on the structure of mental property kinds and the propositional/sentential dimension; it would therefore be a mistake to identify mental properties with the numerical measure of 'desirability' or 'utility' alone.

Two objections against the interpretation of propositional content as a product of measuring mental properties need to be met. First, it has been argued that the logical inference relations over the propositional space are justified only if those relations are homomorphic images of relations over mental properties themselves; and that therefore mental properties must have a semantic component. Second, it has been argued that mental properties are genuine relations between people and propositions, because the propositions do not exhibit the same transformational properties as genuine scales do. I will argue against these claims in turn.

The propositions by which mental properties are represented are usually assumed to adhere to propositional logic. These logical inference relations over the representational space, it is claimed, are relations that similarly exist over mental properties. They have to be, it is argued, for otherwise their validity over the representational space would be unwarranted. If this claim is correct – that is, if there the homomorphism would cover this relation as well – then we would be back at Brentano's problem. Mental properties were measured with propositions and sentences because they themselves hold a semantic content. Indeed, this line of argument has led some to the conclusion that there must be a 'mental logic' or 'mental language'.

But as measurement theory makes clear, certain relations might be defined over the representational space, without them necessarily being found in the respective empirical space. It is thus simply not a priori decidable whether one can read certain relations from the

it' (Davidson 1974a, 147).

representational system back into the empirical system. Instead, it is an empirical question:

> The point to emphasise here is simply that it is unclear what empirical significance, if any, is to be attached to the fact that our representations of certain empirically related propositional attitudes stand in inferential relations to each other. (Matthews 1994, 141)

The claim that the success of representing mental properties is due to these properties having an intrinsic semantic component is warranted as little as the claim that the numbers attributed in weight measurements are in any way part of the property of having mass, or length, or temperature. Brentano's problem – that mental property had to be fundamentally distinct from physical properties due to their intentional content – thus does not occur for this approach.

> From a measurement-theoretic perspective, it would be very surprising indeed to discover that one could read back into the empirical system all the properties and relations of the representational system. (Matthews 1994, 138)

Instead of reading logical inference relations back into the empirical system, they are relations which only apply to the representation space. Whether they do apply, then, is a question of their meaningfulness, which in turn is determined by the transformation properties of the applicable scales.

The Relational Thesis that lies at the heart of Brentano's problem claimed that mental properties are genuine relations between people and propositions. I rejected the Relational Thesis on the grounds that the relation is just a seeming one, like other seeming relations occurring in measurement, and that this appearance does not have any ontological significance. Some philosophers have argued that this argument is based on a pseudo-analogy. While it is correct to point out that some measurement do indeed not represent genuine relations, they claim, it is wrong to declare all predicates with indexicals as designating non-relational properties. The difference has to be drawn, so they claim, according the admissible transformations of the scales involved.

> Take being a son: each son is related to one particular father. But it makes no sense to say that token sons are 'indexed' by some father or other, which one depending on the choice of a unit. Of course not: there is no analogue of a unit here. It is because being a son is a genuine relational property, while having a temperature isn't, that unit-relativity is essential to the pseudo-relationality of temperature, and inapplicable to the real relationality of being a son. (Crane 1990, 228)

To the contrary, I will argue that admissible transformations are found in the sentential and the propositional dimensions of the representational space. Transformations in the sentential dimension are admissible wherever substitutivity of sentences within attitude types implies identity of mental properties. This is of course ultimately a question of what determines the causal role – many examples show that beyond a certain threshold, complexity in the way content is expressed might have a causal role. In the general case, people master the representational space and are able to identify differently expressed attitude contents. Similarly, transformations in the propositional dimension are admissible if causal roles stay constant over differences in propositional content. To illustrate this, I will employ a rather outlandish – in a literal sense – thought experiment: Putnam's Twin World example.

Suppose another planet in the universe (called 'Twin Earth') that is exactly alike to earth but for one fact: on Twin Earth there is no water, that is, no H_2O. Instead an observably and causally indistinguishable substance XYZ runs in Twin Earth's rivers, comes out of Twin Earthian's taps, etc. As everything else on Twin Earth is the same as on earth, people speak Twin English, which is the same as our English except that they use 'water' to refer to XYZ. Now an astronaut from earth lands on Twin earth, and she desires to drink water. She goes to the nearest river, drinks a few handfuls of XYZ, and then tells a Twin Earthian next to her: 'Hmmm, the water tastes really good'. The Twin Earthian agrees, and proceeds to drink XYZ himself.

Under the functional position defended here, the astronaut reasonably presumes that the Twin Earthian has the same desire and the same belief as her: viz. they both want to drink water and they both believe that the water tastes really good. After all, she cannot

tell any difference between the causal roles of her mental properties to his. But if an omniscient observer distinguished mental properties by their propositional content, the astronaut would have a belief and a desire – about H_2O – different from the Twin Earthian, whose belief and desire are about XYZ.

This distinction, however, seems wholly implausible to me. None of the actions of either the astronaut or the Twin Eartian are in any way affected by this distinction. It is empirically completely empty and should thus be discarded. Therefore, there are cases where different propositions can index one and the same mental property. The causal roles that mental properties play are not uniquely identified by their satisfaction conditions. Something like equivalent transformations therefore exist over the representational space of propositions. The measurement-theoretic approach dissolves this apparent problem by establishing a transformation rule between those scales that measure mental properties with the same causal roles: now, having a desire to drink H_2O is equivalent to having a desire to drink XYZ in exactly the same fashion as one and the same object has a weight of one pound or 454 grams.[12] By analogy to the numerical case, then, mental properties should not be considered genuine relations between people and abstract objects. Thus, Brentano's problem does not apply, and mental properties supervene on physical properties.

1.6 Conclusion

In this chapter, I have developed and defended a notion of mental properties as – potentially heterogeneous – collections of physical properties that are identified by their unifying causal role in bringing about behaviour. I have argued firstly that facts can be the relata of the relation '· causes ·', and hence that the realization of a mental property can be such a relatum. Further, I presented a concept of mental properties as internal factors as insufficient but necessary parts of a collection of internal and external – triggering and facilitating – factors that is sufficient but unnecessary in

[12]Thus the measurement theorist does not have to retreat into banality and claim that the representational space consists only of *nominal* scales, as for example claimed here: 'We are entirely free to use any predicate we like for ascribing any property . . . we do not need justification by proving the existence of suitable empirical and numerical relational systems or by demonstrating corresponding representation theorems' (Beckermann 1996, 11).

the production of behaviour. The question then arose how mental properties could be causes at all. I argued that they are not causally efficacious, but causally relevant in that they program for efficacious properties. This programming notion, as I explained, rejects both type-reductionism and epiphenomenalism, while maintaining a naturalistic understanding of mental properties. One of the most fundamental anti-naturalistic attacks on such a position is the claim that mental properties are intentional properties, and that their semantic content makes them different in kind from physical properties. If this attack was valid, then the social sciences would face fundamentally distinct objects and had to employ fundamentally distinct methods. To defend the naturalistic position, I therefore argued against the claim that mental properties are intentional properties by showing how the propositional content of them arises in their measurement, without any significance for their ontological status.

Chapter 2

The Methodology of Mental Property Attribution

2.1 Introduction

In the last chapter, I argued that mental properties are specified as causal roles in the interplay between sensory input and behavioural output of a sufficiently complex system. These causal roles are realised by the physical properties mental properties program for; thus mental properties are significant in action explanation as they name causes of behaviour. Mental properties therefore have a potential function in those social sciences that strive to explain.

However, this conclusion says little about the methodological problems that the social sciences face. All that has been clarified is that the *explanandum* of the social sciences – human behaviour and derived from that the emergence and persistence of social institutions – is the effect of the *explanans* – mental properties of the agent. How the *explanans* as the cause of this effect is identified – that is, how mental predicates are meaningfully attributed to agents – remains an open question.

In this chapter, I will argue that a methodology that determines the meaning of mental predicates through a methodology of constant conjunctions is not sufficient for the social sciences. In particular, I will criticise introspection as a methodology in this direction as insufficient. Further, I will caution against a wholesome rejection of mental properties as *explanans*, as scientific behaviourists did in response to the shortcomings of introspective psychology.

Instead, I will argue that social science methodology has to attribute mental properties as theoretical terms. I will start with a caution: one cannot hope that a theory of these terms can be directly derived from our folk psychological practices. Frank Ramsey's and David Lewis' discussion of how theoretical terms acquire meaning will serve as a basis here: the meaning of mental predicates is determined through a theory, with the evidence exclusively based on behavioural observations.

The approach that I advocate is to take an appropriately designed theory and employ it as a conjectured hypothesis. The attempt is to subsume all observed behaviour under a theory as simple as possible: with as little theoretical terms as possible and as strong restrictions over them as possible. If one succeeds in achieving a satisfactory fit, we have assigned meaning to the theoretical terms *and* learned something about the agent.

However, this raises the problem how to select any candidate from the infinite set of possible conjectures. Contrary to some claims, I will dispute any *apriori* arguments, and admit that mental properties are epistemically indeterminate. The view of cognitive theories presented here is consequentially instrumental to some degree.

2.2 Insufficiency of Causal Methodology

One of the necessary (but not sufficient) conditions for causal inference is information about constant conjunction (or, expressed in more contemporary terminology, information about probabilistic correlation). To secure this information, it is necessary that the purported cause and effect can be observed *independently* of each other.[1] If one wants to derive one's methodology of mental property ascriptions directly from their ontological characterisation as (programming) causes of behaviour, one has to proceed in two steps. First, a set of possible mental properties of the agent has to be identified, which are observable independently of her actions; and second, it has to be shown that a member of such a set is correlated to a specific action of hers. This correlation then could function as the basic evidence for a causal methodology of mental state ascriptions.

[1] By observing purported causes and effects independently I mean that it is logically possible that each realization of a property can appear without the other.

The claim that such properties are independently observable was put forward in most of the 19th century writings of psychology. It claimed that the occurrences in our minds were susceptible to *introspection*; that we could observe or listen to – in analogy to our sense organs – ourselves thinking, feeling or wanting. Although this view is not prevalent anymore in today's psychology,[2] it can still be found in many a social scientist's working assumptions. Occasionally (as witnessed by myself), teachers still explain the utility function with reference to *pleasure* as an introspectable quantity. Further, many microeconomic textbooks start their curriculum with consumer theory that takes preference as a primitive notion (or, at least, they offer such an account as one option besides the revealed preference account – compare Varian 1992; MasCollel et al. 1995). This approach however, by abdicating from any account of a preference-ascribing methodology, suggest that these preferences can be readily found in economic agents – or as it is often put, that although we cannot introspect the magnitude of our desires, the introspection of our preference orderings is unproblematic. Last, empirical studies of consumption patterns often use questionnaires. Information asked in those questionnaires includes the evaluation of quality aspects.[3] These techniques implicitly assume that the individual has privileged access to such information – and thus assumes the possibility of introspection.

In the following section I will rehearse some arguments to the effect that introspection is both incomplete and fallible. From this I conclude that introspection cannot yield a secure basis of a methodology of preference ascriptions. Some of these arguments were first put forward by partisans of Scientific Behaviourism in the first half of the 20th century. Although I share their critical perspective, I find fault with their own constructive program of replacing mental properties in the explanation of behaviour with observable stimuli and stimuli-response regularities. Scientific Behaviourism, I will argue

[2]However, for a cautious advocacy of a return to introspection in psychology, see for example Liebermann (1979).

[3]The subdiscipline of Consumer and Market Research relevant here is *attitude research*. Typical question techniques include the rating scale, by which the respondent is asked to indicate her position on a dimension of opinion. Scales are often calibrated numerically ('give a score out of ten'), diagrammatically ('Indicate in which area you feel most represented') or semantically ('rate on a field between very good and unacceptable'). For further discussion, see Worcester 1986, p. 127-142. Other techniques involve the ascription of adjective to concepts, or vice versa, or a choice between alternative attitudinal positions.

in sub-section 2.2.2, suffers from a vicious circularity in its attempt to explain human behaviour.

I will conclude from this section that mental properties are necessary for the explanation of human behaviour, although these mental properties cannot be observed independently.

2.2.1 Incompleteness and Fallibility of Introspection

Even if we cannot observe other people's mental properties, we have, introspective psychology claims, direct access to our own beliefs, desires, feelings, etc. If this was true, introspective data can be obtained either by truthful reports of the introspecting agent; or, by assuming a basic similarity of the human mind, a psychologist can obtain data through empathy. But introspection is troubled by at least three problems: it does not satisfy inter-subjective agreement, it is incomplete and fallible. I will argue for these three claims in turn.

The objection that introspection does not achieve inter-subjective agreement arises from methodological concerns. For progress in a scientific discipline, the data base chosen should allow the greatest possible degree of agreement and communication amongst observers. Reports about the physical world satisfy the criterion of inter-subjective agreement to a higher degree than introspective reports do. Scientists can look at the same physical object and compare their observations. If disagreement about a certain observation prevails, the physical object can be subjected to various measurement procedures. Mental occurrences do not share these properties. Strictly speaking, no-one can introspect the same mental occurrence as anyone else. All that we can do is to put ourselves in an identical position as someone else, and report our introspection. Further, there is no alternative measurement procedure that could facilitate or even replace introspection: no meter, no microscope, and no spectrograph. For these reasons, although it might exist, introspection can be rejected on purely methodological grounds as too unreliable a technique to be granted scientific status.

But there are metaphysical claims to the opposite of this rejection. Introspection is said to be the privileged access of the individual to her mental occurrences: unmediated, immanent and direct.

Thus, although it does not satisfy inter-subjective agreement, introspection is claimed to be the most reliable method to obtain information about mental occurrences. This claim finds its strongest proponent in Cartesian psychology, which proposes that mental properties like beliefs or desires are *essentially* conscious states, and that this characteristic of consciousness guarantees the introspectability of any such property. Descartes himself took being 'conscious' to refer to an allegedly intimate source of knowledge about one's own mental occurrences. For example, he says that 'to be conscious is both to think and to reflect on one's thought' (Descartes 1991, 335), and elsewhere he had defined 'thought' as 'all the operations of the will, the intellect, the imagination and the senses' (Descartes 1984, 113), and reflection means the mechanism that forms thoughts about thoughts. Thus, Descartes treats everything mental as introspectively accessible:

> As to the fact that there can be nothing in the mind, in so far as it is a thinking thing, of which it is not aware, this seems to me to be self-evident. For there is nothing that we can understand to be in the mind, regarded this way, that is not a thought or dependent on a thought. If it were not a thought or dependent on a thought it would not belong to the mind *qua* thinking thing; and we cannot have any thought of which we are not aware at the very moment when it is in us. In view of this I do not doubt that the mind begins to think as soon as it is implanted in the body of an infant, and that it is immediately aware of its thoughts, even though it does not remember this afterwards because the impressions of these thoughts do not remain in the memory. (Descartes 1984, 171-172)

This claim allows him to conclude that introspection is *necessarily* infallible, providing one with the complete picture of the contents of one's mind. Perhaps even more thoroughly than Descartes, Locke identifies the mental with the introspectively conscious, claiming that it is unintelligible 'that any thing thinks without being conscious of it, or perceiving, that it does so' (Locke 1975, 115) because thinking consists in being conscious that one thinks.

The Cartesian position has been attacked from two angles. On the one hand, it has been argued that introspection is *incomplete*

– that one might have a mental property without being aware of it. On the other hand, it has been argued that introspection is *fallible* – that one can introspect a mental property without actually having it. I will present these two lines of attack here in turn, and thus argue that introspection is not the reliable instrument it is sometimes claimed to be.

Doubts about the completeness of introspection were sometimes raised even in the heyday of philosophical introspection, the 17th and 18th century. Descartes himself seemed to have not been free of such doubts, for example when he says that 'many people do not know what they believe, since believing something and knowing that one believes it are different acts of thinking, and the one often occurs without the other' (Descartes 1985, 122). Similarly, Hume admits that

> 'tis certain, there are certain calm desires and tendencies, which, tho' they be real passions, produce little emotion in the mind, and are more known by their effects than by their immediate feeling or sensation. (Hume 2000, 417)

Despite these admissions, no clear line was drawn between having a mental property and being aware of that mental state by any of the quoted thinkers. Instead, those cases cited were treated as pathological by them. Many later psychologists followed them in this view and expressed their confidence that profound training and exactness in method would remedy these cases[4]

More recent discussions made clear that to have a mental property without being aware of it is far from pathological, but is rather a systematic condition of the human mind. Many of Conan Doyle's stories draw their dramatic drive from the fact that the story's characters did perceive important evidence without being conscious of it. It then takes a mastermind like Sherlock Holmes to identify the importance of some possible evidence for the others (and of course in particular of the loyal Dr. Watson) to become aware of their perception. So for example in *Silver Blaze*: ' "Is there any point to which you would wish to draw my attention?" – "To the curious incident of the dog in the night-time." – "The dog did nothing in the night-time." – "That was the curious incident", remarked Sherlock

[4]For such a view, see James 1890, Otis 1920, Bentley 1926, Pennington and Finan 1940.

Holmes'. Holmes' conversational partner believed that the dog was silent the night the crime was committed, but he would not have expressed this as one of his beliefs about the facts of that night – to him, the dog did 'nothing'.

Adding fact to fiction, Dretske (1993) offers a couple of self-experiments for the reader. The reader is asked to view all the elements of two relatively complex two-dimensional figures. Those two figures differ in some minute detail. Now Dretske focuses on those readers who could *not* tell the difference. Assuming that indeed they saw the figure in its entirety, the readers experienced different visual stimuli, and thus, according to Dretske's terminology, had different *state* conscious experiences for each figure. However, they were not able to tell the difference between the two figures – thus they were not conscious *of* the difference in their experience, they were not *fact* conscious of the difference, in Dretske's terminology. Like Sherlock Holmes, they might at some later point identify the differences *from memory* (state consciousness is a condition of the brain, not the eye!), but at the moment they are state conscious without being fact conscious of it.

> There can be conscious differences in a person's experience of the world – and in this sense, conscious features of his experience – of which that person is not conscious. . . . It follows, therefore, that what makes a mental state conscious cannot be our consciousness of it. If we have conscious experiences, beliefs, desires, and fears, it cannot be our introspective awareness of them that makes them conscious. (Dretske 1993, 278-79)

Introspective awareness thus is not necessary for being in a conscious state (having a mental property). Consciousness does *not* imply introspectability, as Descartes or Locke claimed. The distinction between state consciousness and fact consciousness thus opens the conceptual door for the possible incompleteness of introspection.

The empirical findings of many psychological experiments confirm this possibility. Particularly relevant for the question of completeness are experiments that show the inability of test persons to register and hence to report a changed evaluation or motive state. Experiments in the cognitive dissonance tradition, and in particular in the area of insufficient-justification and attribution manipulation,

show these results with great clarity.[5] Bem and McConnell (1970) asked subjects to write an essay opposing their own views and then report their views afterwards. Subjects bribed or coerced into assuming the counter-position showed no change of their evaluation in their post-essay reports; subjects who were given insufficient justification for writing the essay, or who were manipulated into believing that they had a free choice, showed a significant shift of their evaluation towards the position assumed in the essay. Bem and McConnell then asked the individuals in the latter group what their evaluations had been one week earlier, before they received the task to write the essay. Subjects reported throughout that their attitudes had not changed, and that they were the same as before writing the essay; while control subjects had no problems reporting their previous attitudes with great accuracy. Thus subjects apparently changed their attitudes in the absence of any subjective experience of change.

A related experiment by Goethals and Reckman (1973) supports this conclusion. High school students changed their opinions on a policy issue in the face of persuasive counterarguments from an authoritative person, but later reported that the post-discussion opinion had been their opinion throughout. Again it seems that subjects changed their attitude in a cognitive process – understanding and accepting a position that reportedly was not their own – without being able to introspect and report this change themselves.

Nisbett and Schachter (1966) followed up on these phenomena and confronted subjects with their change in behaviour. They asked subjects to take a series of electric shocks of steadily increasing magnitude. Before administering the shocks, half the group were given placebo pills which purportedly produced heart palpilations, breath irregularities, hand tremor, and butterflies in the stomach – symptoms most often reported as accompanying the experience of electric shocks. It was postulated that those who took the pill would attribute the symptoms of receiving the shocks to the pill, and thus would endure more severe electric shocks. And so they did: subjects who took the pill endured four times as much amperage as those who did not. In the debriefing interview, pill-attribution subjects were asked with increasing precision about the effect of the pill. Most of them denied that their belief in the pill's effects bore any signifi-

[5]Nisbett and Wilson (1977) survey five other literatures significant in this context. I will confine myself to discussing the following three experiments for their simplicity, clarity and convincing results.

cance to their ability to take the shocks; instead, they claimed that all their attention was consumed by shocks themselves. Only 3 out of 12 subjects reported having attributed their symptoms during the administration of the shocks to the pill. When the experimenter finally revealed the design of the experiment, including the postulated effect of the placebo, most subjects found the postulated effect of the pill plausible, but non-applicable in their own case.

The experiments clearly show the severe limits of introspection. Although subjects in each case must have been (state) conscious of the relevant stimuli – as they were verbally reported in each case – and although their behaviour significantly changed because of these stimuli, they were not able to report this change, even upon direct questioning. The most plausible answer for these test results is that the subjects were state conscious of the stimuli in such a way that the stimuli were causally efficacious, but that the subjects were not fact-conscious of the determining influence of these stimuli. This in turn seems most likely the case because the subjects did not have the introspective facility to gain such a fact consciousness, or that there was something that blocked this introspective facility for example in high-dissonance contexts. Therefore, introspection is at best incomplete.

This insight stood at the beginning of modern decision theory. Ramsey rejected the measurement of an agent's beliefs from introspection,

> for the beliefs which we hold most strongly are often accompanied by practically no feelings at all. (Ramsey 1926, 65)

Instead, he confined himself to a derivation from actual and hypothetical choices. Considering this fortunate start, it is surprising how much introspective residue is still found in the social sciences of today.

It could still be argued that despite its limitations, introspection may be epistemologically reliable within its proper domain. In other words, as soon as a subject has introspective access to her thoughts, then her verbal report, if truthful, constitutes reliable data. Even more, if there is any introspective data, then it could still count as the best possible data, if one agrees with those philosophers who thought that the individual has privileged access to her own mind.

Descartes for example argues that the mind is 'better known' than the body (in his Second Meditation). Locke claims that our knowledge of 'things without us' is 'not altogether so certain, as our intuitive knowledge' (Locke 1975, 631).

The thus claimed infallibility (if incompleteness) of introspection was first challenged by Kant, who pointed out that introspection

> . . . represents to consciousness even our own selves only as we appear to ourselves, not as we are in ourselves. (Kant 1929, B153)

Kant thinks of introspection as an 'inner sense'. In introspecting, the mental occurrences of the agent appear to the agent herself; the agent senses herself thinking. But if there is something that is sensed and something that is sensing involved in introspection, then there are two separate entities with a relation that is fault-prone. Armstrong, following Kant, holds the introspecting state and the state being introspected to be 'distinct existences'. He draws a mechanical analogy to the awareness of mental states: the introspecting mind acts like a scanner scanning itself.

> It is clear that the operation of scanning and the situation scanned must be "distinct existences". A machine can scan itself only in the same sense that a man can eat himself. There must remain an absolute distinction between the eater and the eaten: mouth and hand, say. Equally, there must be an absolute distinction between the scanner and the scanned . . . the registering will have to be something logically distinct from the featured that are registered. (Armstrong 1969, 107)

If this analogy holds for introspection, introspected states are distinct from introspecting states. The relation then certainly cannot be one of necessity – a mechanism correlating one's thoughts with one's thoughts about them is breakable, or even manipulable. The understanding of introspection as an 'inner sense' implies that the relation between mental property and introspection is contingent, and prone to error.

The possibility of errors is again substantiated by psychological experiments in the cognitive dissonance tradition. Particularly interesting in this context are those experiments that show how subjects report the influence of non-effective stimulus factors. Nisbett

and Wilson (1977) report three such experiments. In the first, subjects were given a selection from a novel that described a situation of strong emotional impact. Some subjects read the whole selection, others read parts from which passages had been deleted. From the subjects' reports, there was no difference in the emotional effect of the full selection or the reduced ones. However, confronted with the deleted passages and asked about the relevance of the passages on the emotional impact of the selection, a large majority of the subjects reported that the passage had an influence (those who got the reduced version reported that it would have had an influence).

In the second experiment, subjects were asked to view and rate a documentary film according to three dimensions – how interesting they thought it was, how much they thought other people would be affected by it, and how sympathetic they thought the main character to be. Half the viewers were exposed to a distracting noise while watching the film. Although the rating from the noise-exposed group and the control group showed no significant differences, a majority of the subjects exposed to the noise reported that it had an impact on their rating.

In the third experiment, subjects were asked to predict the magnitude of shock they would take in experiments on the effects of intense electric shocks. Half of the subjects were given an 'assurance' that shocks would not yield 'any permanent damage'; the other half was not given such an assurance. The inclusion or exclusion of the assurance did not have a significant effect on the prediction. Nevertheless, a majority of subjects reported that it did.

The experimental results show that introspection is both incomplete and sometimes fallacious – and this 'sometimes' is sufficiently often to worry even from a methodological point of view. Together with the initial methodological concern, one can therefore conclude that introspection, as it is understood today, is a method too unreliable to be employed for preference attribution. It does not satisfy inter-subjective agreement; it is incomplete and often fallacious. As long as the systematic causes for incompleteness and error are not identified, and methods developed to circumvent them, introspection cannot be the method to provide us with independent observations of mental properties. To use correlations between introspected mental properties and actions for causal inference is too unreliable to be

adopted in a science of behaviour.[6]

Further, I believe that introspection is the only candidate that promised to provide direct and independent observation of mental properties. With this option gone, we have to abandon our hopes for a theory of behaviour that is based on probabilistic correlation between mental properties and behaviour altogether. From this result, psychology can draw two very different conclusions. The one is to insist on a causal methodology based on correlations between observable stimuli and behaviour and negate the theoretical role of mental properties in general. This is the route the psychological mainstream took in the first half of the 20th century with scientific behaviourism. As it will become clear in the next sub-section, I think that this was the wrong choice. The other conclusion is to abandon a methodology based on correlations and reconstruct mental properties as theoretical concepts. This is the way cognitive science made popular from the '70s onwards, and I will argue for it in the second section of this chapter.

2.2.2 The Circularity of Behaviourism

In the last section I argued that psychology and the social sciences couldn't hope to employ introspective information as their database. The exclusion of introspective data shifts the methodological focus onto behavioural observation itself (which includes verbal behaviour, which is to be distinguished from verbal accounts of introspection in the way it is interpreted). Insofar as the critique of introspection leads to the conclusion that all data employed in psychology and the social sciences must be behavioural – or at least, publicly observable – data, this critique is in accord with a *methodological* behaviourist position.

Such a position, however, needs to be clearly delineated from the claim that psychological theory should eschew discussion of inferred mental processes and mechanisms altogether, which was the essence

[6] In a very interesting monograph Ericsson and Simon argue that 'verbal behaviour is to be accounted for in the same way as any other behaviour . . . by developing and testing an information-processing model of how information is accessed and verbalised in response to stimuli' (Ericsson and Simon 1993, 62) This proposal might possibly make introspective data fruitful again: by employing to it the same cognitive theory of behaviour that will be discussed below, and thus deriving mental properties from it. This approach puts introspection from its head on its feet: it treats introspective data through the lens of the theory, instead of granting it privileged access, and trying to construct the theory from it.

of *scientific* or radical behaviourism.[7] In the following, I will show that the version of scientific behaviourism as defended by B.F. Skinner is untenable as a methodology of human action explanation.

Skinner contended that scientific psychology ought to be concerned only with the formulation of laws relating observables such as stimuli and responses; not with unobservable mental processes and mechanisms such as attention, intention, memory and motivation.

> The objection to inner states is not that they do not exist, but that they are not relevant in a functional analysis. We cannot account for the behaviour of any system while staying wholly inside it; eventually we must turn to forces operating upon the organism from without. Unless there is a weak spot in our causal chain. . . , the first [sensory input] and the third [behavioural output] links must be lawfully related. If we must always go back beyond the second link [mental states] for prediction and control, we may avoid many tiresome and exhausting digressions by examining the third link as a function of the first. (Skinner 1953, p. 42)

This quotation shows what is fundamentally right with scientific behaviourism: its quest for simplicity in theory and for a data base that – through public observability – maximises the potential for intersubjective agreement between skilled observers. Both criteria are universally acknowledged conditions for the acceptability of a scientific theory. First, the theory accepted should be the simplest among those that are formally capable of accounting for the data. Second, the data that the theory accounts for, and by which it is appraised, should be only data that skilled observers can agree upon. Both criteria seem to be exemplarily satisfied in some paradigmatic cases of animal behaviour. For example, when a rat learns to press a lever to obtain food, Skinnerians explain this by an appeal to the 'law of effect'. This law says that a response - emitted in the presence of a stimulus and followed by a reinforcer – increases the probability of that response in the presence of that stimulus. Another strong

[7]The third major strand of behaviourism, *analytic* behaviourism, is a *theory of meaning* of mental properties and does not have a direct bearing on the methodological discussion followed here.

case for the application of the law of effect appears in situations driven by instinct, as beautifully described by Wooldridge:

> When the time comes for egg laying the wasp Sphex builds a burrow for the purpose and seeks out a cricket which she stings in such a way as to paralyze but not kill it. She drags the cricket into her burrow, lays her eggs alongside, closes the burrow, then flies away, never to return. In due course, the eggs hatch and the wasp grubs feed off the paralyzed cricket, which has not decayed, having been kept in the wasp equivalent of deep freeze. To the human mind, such an elaborately organised and seemingly purposeful routine conveys a convincing flavor of logic and thoughtfulness – until more details are examined. For example, the wasp's routine is to bring the paralyzed cricket to the burrow, leave it on the threshold, go inside to see that all is well, emerge, and then drag the cricket in. If, while the wasp is inside making her preliminary inspection the cricket is moved a few inches away, the wasp, on emerging from the burrow, will bring the cricket back to the threshold, but not inside, and will then repeat the preparatory procedure of entering the burrow to see that everything is all right. If again the cricket is removed a few inches while the wasp is inside, once again the wasp will move the cricket up to the threshold and re-enter the burrow for a final check. The wasp never thinks of pulling the cricket straight in. On one occasion, this procedure was repeated forty times, always with the same result. (Wooldridge 1963, 82; quoted from Dennett 1978)

In such a case, the regularity of the behaviour is so deeply rooted, that the two above criteria dictate the application of the law of the effect. In fact, it would be wrong to interpose any 'inner states' here. If the wasp's behaviour was explained on the basis of her desire to drag her victim into the burrow, it would become inexplicable why she would not learn from experience – that is, why she did not end all that cricket-magic after a certain number of repetitions with a courageous push right over the threshold into the final destination. Instead, her seemingly erratic behaviour is explicable by a (possibly genetic) conditioning that has her perform the individual steps in accord with the respective stimuli (the positions of the cricket).

but there still be desire to do the actions performed. (compare the pornography example).

The problems with scientific behaviourism begin where its proponents claim that this kind of Baconian explanation – subsuming observations under generalizations about lawful regularities – is *all* that psychology should aim for.

> We change the relative strength of responses by differential reinforcement of alternative courses of action: we do not change something called a preference. We change the probability of an act by changing a condition of deprivation or aversive stimulation; we do not change a need. We reinforce behaviour in particular ways; we do not give a person a purpose or an intention. (Skinner 1971, 94)

That this methodological confinement is too tight to work for the explanation of human behaviour can be shown in at least three instances: scientific behaviourism cannot account for many kinds of learning, it suffers from an infinite regress and it cannot explain why systems behave differently under identical stimuli environments.

First, the law of effect has problems to correctly model learning in animals and humans. The wasps of the above example does not learn, hence the application of the law of effect is safe. But there are animals that learn without showing the postulated stimulus-response-reinforcer cycle. For example, the juvenile white-crowned sparrow is incapable of singing during the critical learning period; all it does is listen.[8] Similarly, nestling buntings, when they are too young to fly, learn to identify north as the centre of rotation of the night sky; they do so by watching the movements of the stars and inferring their centre of rotation. Months later, they use this knowledge to keep oriented during the night-time portions of their migratory flights.[9] Human learning behaviour – for example from textbooks – seems to fall in similar patterns. Most learning depends on the perception over time of a systematic stimulus relationship. Whatever response the animal or human may or may not make during the period when the relationship is perceived is of little relevance to the learning that occurs. The lack of a response and hence of a potential reinforcer poses the problem for behaviourism of how to distinguish the acquisition of new stimulus-response mechanisms

[8] For a discussion how white-crowned sparrows learn songs, see Nelson and Marler, 1994.

[9] For a series of wonderful experiments, involving young buntings and a planetarium, see Emlen 1972.

from the disregard of any such stimuli. While a cognitivist approach can explain those learning phenomena as the acquisition of beliefs – and hence of behavioural dispositions – the behaviourist either has to find some form of reward mechanism, or has to deny that any learning is happening here at all.

This problem is exacerbated in explaining human behaviour. Humans are regularly faced with situations they never encountered before - never even envisaged before. However, they have the ability to behave in a systematic and reasoned manner, that is in accord with general patterns of human behaviour. Confronted with a threat in a stick-up (Dennett's example), most people will stay quiet and hand over their wallet, even though they might never have been in such a situation before.

> It is perfectly clear that what experience has taught me is that if I want to save my skin, and believe I am being threatened, I should do what I believe my threatener wants me to do. But of course Skinner cannot permit this intentional formulation at all, for in ascribing wants and beliefs it would presuppose my rationality. He must insist that the "threat stimuli" I now encounter are similar in some crucial but undescribed respect to some stimuli encountered in my past. . . He is positing an external *virtus dormativa*. He has no record of any earlier experiences of this sort, but infers their existence, and moreover endows them with an automatically theory-satisfying quality. (Dennett 1978, 67)

To explain systematic behavioural patterns like this, the behaviourist must revert to an antecedent that does not satisfy the criterion of observability at all. Instead of having prior evidence for a particular conditioning, the behaviourist now postulates such a conditioning because he needs it for the required explanation.

Secondly, the Skinnerian 'law of effect' is very vague in its scope of application. It does not really clarify how similar the experienced stimulus and enforcer have to be to a new situation to elicit the response detailed in the law. It is clear that in Pavlov's famous conditioning experiment, an untrained dog will not exhibit salivation, even if exposed to the bell sound under laboratory conditions. But even a trained dog conditioned on the sound of the bell might

not salivate when immersed in water, or when exposed to a bitch in heat. Similarly, the behaviourist might insist that when confronted by muggers, the subject reacts as in other situations with a 'threat stimulus'. But how do we identify the scope of these claims? How do we determine that situation X is still determined by a 'food stimulus' or a threat stimulus', while situation Y is already outside of its scope, obeying different laws of effect? For laws of effect to be operational, the behaviourist has to apply to particularly strong *ceteris paribus* conditions.

> In characterizing a man's behaviour in terms of frequency [of bits of behaviour in response to stimuli], we assume certain standard conditions: he must be able to execute and repeat a given act, and other behaviour must not interfere appreciably. . . . We eliminate, or at least hold constant, any condition which encourages behaviour which competes with the behaviour we are to study. (Skinner 1953, 42)

The laboured attempt to formulate the *ceteris paribus* conditions in terms of behaviour shows its fatal shortcomings. First, it is unclear how *behaviour* could interfere with behaviour. It seems already here that a concept of consistency is needed, and that such a concept cannot be defined over stimuli or behaviour. A similar point can be made about the concept of 'competing behaviour'. Second, the 'conditions which encourage behaviour' cannot be reduced to the stimuli which stand at the beginning of the causal chain. One stimulus can give rise to very many different reactions. Any attempt to purify one reaction as *the* only one involves determining the presence or absence of further stimuli, which in turn need specific conditions to play their 'purified' role. The attempt to eschew the mental property level by referring to observable behaviour only requires a never-ending qualification of background conditions and hence leads to an infinite regress.

Thirdly, behaviourism suffers from the fact that systems – and in particular humans – behave differently in identical stimuli environments. In Balzac's *Old Goriot*, for example, the manipulator Vautrin offers the poor student Eugène a cash credit, which the latter rejects. But moments later, after Vautrin had implied that he would ask for a high interest, Eugène is ready to take the money. Why? Eugène is faced with the same stimulus environment: the

money he so urgently needs in Vautrin's hand. But the latter's demand of a high interest is not a stimulus that would make Eugène take the money – unless for a complex chain of beliefs and desires that made him believe he could save his honour by taking a usurious credit from Vaudrin, but not by taking a chanty. If such scenarios are admitted at all, then the explanation of different actions under identical present stimuli must revert to differences in the past of the system – differences that could give rise to different perception, interpretations and expectations. But once that is admitted, the boundaries of scientific behaviourism are crossed and we are in the realm of the cognitive, as will become clear in the next section. Scientific Behaviourism cannot account for some kinds of learning, it suffers from an infinite regress and it cannot explain why systems behave differently under identical stimuli environments. It therefore does not provide a workable methodology of action explanation.

Skinnerian Behaviourism set itself and everybody else a very high-minded goal: to allow psychology into the realm of science only with the highest credentials of simplicity and inter-subjective agreement. Unfortunately, they threw out the baby with the bathwater. In the face of the problems discussed, it seems little more than highhanded to claim that

> . . . psychologists have unwittingly been analysing contingencies of reinforcements, the very contingencies responsible for the behaviour mistakenly attributed to an internal originator. (Skinner 1990, 1209)

Even though it is – from a scientific standpoint – undoubtedly desirable to explain behaviour in terms of stimuli and reinforcers alone, this method has failed for the above reasons for most cases of human behaviour.

For all those cases, there is an explanation involving intentional mental states, and this explanation should be applied at least in all those cases where behaviourism fails. Skinner, as Dennett points out, fails to see the distinction between explaining and explaining away, and thus he thinks that behaviourist methodology has to *replace* a cognitive one. That position merely leads to the rejection of most of psychology as non-scientific, without any operational methodology. Instead, cognitive explanations co-exist with $S - R$

explanations (where those are possible at all) – they are explanation of higher complexity with lesser possibility to test, but they are explanations nevertheless.

2.2.3 Saving Behaviourism?

Behaviourists have tried to defend their position against the above arguments by making two concessions. First, they have expanded the $S-R$ relationship to a *functional relationship* between the environment and the behaviour. The behavioural variable B is now seen as a function of the intensity vector of the environmental influences I : $B = F(I)$. F is interpreted as a description of the functional stimulus for a response B (Zuriff 1976, Gibson 1950, Ullmann 1980). Second, the past environmental stimuli affecting the agent are now included in the vector I (Rachlin 1977c). Behaviourist explanation then operates as follows:

> A stimulus like water may now be threatening to a person due to its past interaction with the person. The quality of being threatening to a person can be viewed as a current dispositional property of the stimulus acquired through its membership in a stimulus class which previously interacted with that person. . . . being a threatening stimulus does not require a representation of the stimulus inside the threatened person to explain why only this person is threatened by the stimulus. Thus, for the behaviourist, the behaviour of the hydrophobic is a function of a number of environmental variables, and one of these is the environment's history of interaction with this person. (Zuriff 1985, 165)

An agent's behaviour is now explained by reference to present as well as past stimuli affecting the agent. But unlike Skinner's program, there is no *prior* evidence that past stimuli have conditioned the agent. In fact, no such conditioning is on the record at all, as no stimulus-response-reinforcer relation was observed. Rather, the influence of past stimuli as *potential* conditioning factors is *hypothesised.* As we find the agent to act overly fearsome in proximity to water, we diagnose the agent to be hydrophobic. We then search for past incidents that could have conditioned the agent to perceive

water as threatening. Given that the search is successful, the behaviourist finds the functional relationship between a stimulus class and the behaviour supported. What is missing from this type of behaviourism is the prior identification of conditioning stimuli through a reinforcer: e.g. an accident when swimming or boating at early age, pressure from swimming instructors or parents, vivid imaginary story about monsters or big fish underwater, etc. that resulted in the compulsive avoidance of water already observed before the current phenomenon is to be explained. Instead of the rigid Skinnerian stimulus-response-reinforcer regularity the behaviourist now invokes stimuli that are potential and plausible conditioners, without any independent evidence available for this role.

This watered-down version of behaviourism faces two conceptual problems. First, the connection between past stimuli and present behaviour is unclear. Skinner requires a direct 'causal chain' between behaviour and 'forces operating upon the organism from without' (Skinner 1953. 42). For the newer version of behaviourism, the question arises: how can a past stimulus (no longer present) be a cause for a present behaviour? To clean up this chain conceptually, something has to be interposed that creates continuity between stimulus and behaviour – a state or a property, at least a dispositional one. Second, the ex post hypothesis is not based on any selection criteria. Agents are exposed to an enormous number of stimuli throughout their lives. What are the scientist's criteria in selecting the causally relevant stimuli? Which of them are potential and plausible conditioners? To stay with the example, there are many people who experience some threatening situation involving water in their early lives; however, only a few of them develop hydrophobia. So what made the respective stimulus condition some people to avoid water, and others not? Within the framework of conditioning, the new-version behaviourist is unable to answer this question. As long as the behaviourist insists on the direct causal chain from stimulus to behaviour, it remains unclear how some members of the set of all stimuli of an agent's history can obtain causal relevance.

The two problems that a watered-down behaviourism faces – the continuity of the causal chain and the selection of relevant stimuli – are solved by the re-introduction of mental properties into the prediction and explanation of behaviour. Mental properties are acquired at a point in time, retained over a period, and they become

efficacious when at a later time specific conditions are satisfied. Thus they do provide the continuous chain between a stimulus and a later response. Further, mental properties are assigned under the regime of a theory that postulates specific relations between stimuli, mental properties and behaviour. As will become clearer in the discussion below, the theory and the mental properties governed by it cannot be disassociated. Hence for all mental properties assigned, a selection rule comes attached to them. Whether the mental property assignment is correct of course remains to be tested against reality, just as the behaviourist's claim about a functional relationship between stimuli input and response output; because of the interconnection between assignment and selection, however, the former is much clearer in its claims than the latter.

The nature of mental properties employed in this fashion was discussed in chapter 1. The possible justification for assigning particular configurations of these mental properties will be discussed in the next section.

2.3 A Cognitive Theory of Behaviour

A modified version of behaviourism, conceding that behaviour is a function of past and present stimuli, faced the problems of how to model this claim causally, and how to select the causally relevant stimuli. These two problems can be resolved by reintroducing mental properties into the explanation of behaviour. A stimulus at t_1 caused the agent to believe p; this belief is a disposition to perform act q at t_2, given certain conditions. Thus the causal chain runs from stimulus to belief and from belief (plus further conditions) to behaviour. Further, mental properties are hypothesised *ex post* just as the behaviourist hypothesises the past stimulus. But mental properties are highly modularised. They are distinguished in kind (at least two: beliefs and desires) and are individuated by their content. A mental property is ascribed not on the basis of the (repeatable) response in one highly specific situation alone – as the behaviourist does – rather, it is ascribed on the basis on a variety of situations, in each of which it plays a causal role.

In this section I will discuss the approach of identifying mental properties as theoretical concepts. I will present the origin of this account – Sellar's myth – first, and show that the common-sense

interpretation of the theory implied here is not warranted. Instead, I will argue for the Ramsey-Lewis account of theoretical terms. To show that the use of theoretical terms construed this way is common beyond the social sciences, I will give two examples from the natural sciences, and only then discuss the application to mental properties and cognitive theories of behaviour.

2.3.1 A New Type of Mental Properties

The mental properties reintroduced are quite different from those employed in 19th century psychology. They respond to the need to organise connections of higher complexity between past and present stimuli and behaviour; but otherwise, they accept many of the methodological positions the behaviourists argued for. First, they are assigned exclusively on the basis of behavioural evidence (including verbal *behaviour*). Second, they are assigned as the *causes* of that behaviour. Third, their assignment – in whatever form – does not say anything about the internal representation of these causes. In particular, the complexities of mental-material dualism do not arise – mental properties, as discussed in section 1.4, program for physiological or physical causes. Further, even though mental properties might be ascribed as intentional attitudes (properties directed to a semantic object), this ascription does not imply that they are internally represented as such. As argued in section 1.5.2, their intentional quality might arise solely in the process of their measurement.

The justification for this new type of mental property ascription is that it is driven by a theory that allows better explanation and prediction of behaviour than a behaviourist position. This theory postulates a particular mechanism for how mental properties interact in bringing about behaviour; and in accord with this mechanism, the theory attributes mental properties to individuals such that their past and present stimulus experience and behaviour is fitted best under its pattern. It is the need for this pattern, which allows for more complex connection than direct relations between stimuli and responses, which gave rise to the new concept of mental properties. Sellars illustrates this idea nicely by 'making a myth. . . or, to give it an air of up-to-date respectability, by writing a piece of science fiction – anthropological science fiction' (Sellars 1956, 309). This

myth suggests a development of cognitive theory in three stages. First, in the prehistory of psychology, humans spoke in a purely behaviouristic idiom – all their descriptions and explanations of their and others' behaviour were in terms of 'public properties of public objects located in space and enduring in time'. In the second stage, this idiom is found wanting, and is consequently replaced by a more complex idiom that includes mental properties:

> In the attempt to account for the fact that his fellow men behave intelligently not only when their conduct is threaded on a string of overt verbal episodes – that is to say, as we would put it, when they "think out loud" – but also when no detectable verbal output is present, Jones develops a theory according to which overt utterances are but the culmination of a process which begins with certain inner episodes. And let us suppose that this model for these episodes which initiate the events which culminate in overt verbal behaviour is that of overt verbal behaviour itself. In other words, using the language of the model, the theory is to the effect that overt verbal behaviour is the culmination of a process that begins with inner speech' (Sellars 1956, 317-18)

In Sellars' myth, Jones invents a new *stance* towards his fellow humans: he interprets them as intelligent beings, instead of seeing their behaviour – in a proto-Skinnerian perspective – determined by present stimuli. He does *not* do so out of humanistic motives, because he wants to bestow an air of freedom and dignity upon them; rather, he *has* to do so because the simplistic stimulus-response explanations do not work. By postulating intelligent human behaviour, modelled as semantic internal episodes, Jones provides himself with a richer explanatory tool.

Sellars' myth then continues to a third stage, where Jones' new theory has proven successful and hence has been adopted by many of his fellow human beings. In that situation, the theory that was meant to ascribe mental properties to others is employed by Jones and his compatriots to ascribe mental properties to themselves as well – on the basis of their own behavioural evidence. This development leads to a shift in the theory's role. While in stage two it is the prediction and explanation of others' behaviour that required

the ascription of mental states to them, in stage three the ascription itself comes into prominence. 'What began as a language with a purely theoretical use has gained a reporting role' (Sellars 1956, 320). This ascriptive function of the theory has mislead many to believe that they indeed do have privileged access to their own mental properties.

As I showed in section'2.2.1, there are no convincing arguments for such a privileged access, but there is lots of experimental evidence against it. But the finding of Nisbett and Wilson (1977) and others not only show that the purported introspection is rather blind; they also show that the purported introspective access leads to *systematically wrong* results. This requires explanation. If people *systematically* report wrong stimuli, motives or cognitive processes, then there must be some mechanism that provides these reports, but which is not connected to introspection. Sellars' myth provides an explanation for this other mechanism: a theory of the deliberating mind that arises out of the need to explain other people's behaviour more adequately (while of course open to mistakes), which finally is transformed into a theory of self-ascribed mental properties.[10]

Because of its purportedly wide-spread use and deeply ingrained character, the theory of Sellar's myth has been alternately baptised 'folk theory' or 'common sense psychology'. Both of these names betray two crucial but possibly unwarranted assumptions. The first is that Sellar's 'theory' can indeed be seen as a body of folk psychological laws. If that was so, the explanatory and predictive device that Jones uses could be made explicit – it could be expressed in sentences using only well-defined terminology – and without further ado be used as a scientific theory. The second assumption is that all theories satisfying Jones' needs in stage two of the myth must be of the same kind, if not identical, and that this unified theory must be by and large true.

The truth of both claims would have important consequences for the methodology of behavioural explanation. If the first claim

[10] In the cases discussed by Nisbett and Wilson, the systematicity of the errors becomes explicable only once we turn our backs on introspection and see that people ascribe mental properties to themselves with the help of a theory. In the cases discussed, people had the wrong theory, but for a good reason: dissonance reduction. Faced with insufficient justification for a task at hand, they attributed themselves mental properties that provided that justification. Faced with attribution manipulation, they attributed to themselves those mental properties the simplest theory would attribute.

is true, a theory of behaviour based on mental properties could readily adopt the *form* of folk psychology; if the second was true, a theory that could claim folk theory as its basis would have a good *justification.* Decision theorists have banked on both claims. A good example of this is David Lewis, who suggests:

> Collect all the platitudes you can think of regarding the causal relations of mental states, sensory stimuli, and motor responses. Perhaps we can think of them as having the form: 'When someone is in so-and-so combination of mental states and receives sensory stimuli of so-and-so a kind, he tends with so-and-so probability to be caused thereby to go into so-and-so mental states and produce so-and-so motor responses.' Add also all the platitudes to the effect that one mental state falls under another. . . include only platitudes which are common knowledge amongst us. (Lewis 1972, 212)

Lewis assumes that folk theoretical 'platitudes' are of the form of lawlike generalizations that capture causal relationships. This is a widespread assumption in the philosophical discussion of folk theory. In particular, folk theory is often assumed to consist of three different types of these lawlike generalizations; those between stimuli and mental states, those between mental states and those between mental states and behavioural responses, where $\rightarrow$ signifies a causal relation:

(1) $(x)x$ perceives [something]$\rightarrow M_i(x)$
(2) $(x)M_i(x) \wedge M_j(x) \rightarrow M_k(x)$
(3) $(x)M_i(x) \wedge M_j(x) \rightarrow x$ does [something]

An example of type (1) would be 'a person denied food for any length of time will feel hungry'; an example of type (3) would be 'a hungry person's mouth will water at the smell of food'. Examples for type (2) are cited by Churchland:

> $(x)(p)[(x$ fears that $p) \rightarrow (x$ desires that $\neg p)]$
>
> $(x)(p)(q)[((x$ believes that $p) \wedge (x$ believes that (if p then $q))) \rightarrow$ (barring conflicting desires, distractions, etc., x believes that $q)]$
>
> (Churchland 1981, 71)

But despite its pervasiveness in the philosophical literature, the form of folk psychology is all but uncontroversial. I will briefly discuss two competing approaches to point out the difficulty with the assumption of folk psychology's form.

The first is an approach of social psychology, the branch of psychology that investigates folk. Social psychology makes clear that trait attribution constitutes an important part of folk psychologising. Instead of attributing mental states as the effects of stimuli and the causes of behaviour, social psychologists have found that people attribute personality traits to others on the basis of their appearance, behaviour and certain stereotypes, then associate further traits with the attributed traits and explain their behaviour on the basis of the thus accumulated traits.[11] This approach differs from the one commonly assumed by philosophers in two ways. First, the traits attributed are different from the desires and beliefs philosophers typically think about. Traits like 'honest', 'selfish', 'humorous' or 'nervous' comprise a wide field of behavioural dispositions.[12] They are not necessarily inconsistent with belief-desire ascriptions, but it needs to be argued quite specifically which desires and beliefs make up any given trait. As a little reflection will show, this connection between desire-belief and trait attributes is not the common fair of folk psychology: few people can readily say what desires and beliefs they ascribe to a person when they attribute a trait like 'easy-going' to her. Second, the interconnection between traits is not considered to be causal, but of purely associative character: people 'seem to hold that traits "go together" . . . if a person is judged to be talkative, they are also likely to be judged sociable; in contrast, if they are judged to be cautious, they are also likely to be judged silent' (v. Eckhardt 1995, 36). But while it is commonly claimed that a belief that p and a belief that if p then q causes the belief that q, it is not claimed that talkativeness causes sociability or the other way around. The trait ascription theory is thus not a causal theory.

The second dissenting approach comes from within philosophy itself, and its implications for the form of folk psychology are more radical than the trait ascription theory. It claims that people obtain

[11] For a good discussion of this approach in social psychology, see v. Eckhardt 1995, 35.

[12] Rosenberg and Sedlak 1972 asked experimental subjects to describe persons with at least 5 psychological adjectives. The results were ordered by frequency; all traits mentioned here are from this list.

their folk psychological intuition by putting themselves in someone else's position and then use their own decision-making system to simulate the other person's decision making.[13] From such a simulated deliberation, people derive predictions about other people's behaviour, without needing to resort to a set of lawlike generalizations at all.[14] Admittedly, there are some important arguments against the simulation theory. For one, there is some developmental data that the standard theory can explain well, but which poses a problem for the simulation theory (Stich and Nichols 1992, 146-50); while there so far has not been any set of psychological phenomena that posed a serious problem for the standard theory, but would be explicable on the simulation account. Further, the simulation account is in need of some sort of introspection, and I have raised serious doubts about such a possibility in section 2.2.1. However, none of the arguments are conclusive at this point, so that the simulation account is a contender to the more standard understandings of how people obtain folk psychological intuitions.

As long as these competing accounts are not securely defeated, it cannot be claimed with any justification that folk theory has the form of a set of lawlike generalizations. Decision theorists and other scientists interested in constructing a cognitive theory of behaviour therefore cannot hope to derive the form of their theories directly from folk psychological platitudes.

Even if one cannot hope that folk psychology will provide the form for a scientific theory of behaviour, one might still be optimistic that the folk intuitions people have – when translated into the right form – can serve as the justifying basis of any such theory. Lewis seems to have this second idea in mind when he proposes to employ only 'the platitudes which are common knowledge

[13]To be more exact, this is Alvin Goldman's view, which differs from the position of simulation theory's other main proponent, Robert Gordon. Gordon claims that the explaining observer simulates by trying to adjust for relevant psychological differences between the subject and her and by imagining what the subject would do. Compare Goldman (1992) and Gordon (1986).

[14]It has been claimed that if the simulation approach is true, then 'there is no such thing as folk psychology!', as Stich and Nichols emphatically claim (Stich and Nichols 1992, 125). I think this is an unfortunate terminological move. One should understand folk psychology in a wider sense, as 'a set of attributive, explanatory, and predictive practices [aimed at people's own and others' psychological states and overt behaviour], and . . . a set of notions used in those practices' (v. Eckhardt 1994, 300). Then the 'set of lawlike generalisations' is one possible form of folk psychology; but there are other possible explanatory practices that are called folk psychology without having the form of a proto-scientific theory.

amongst us'. Examples of such platitudes, which in very important methodological approaches have been given foundational status for economics, can be found in Mill, who claims that 'a greater gain is preferred to a smaller one' (Mill 1949, 6.9.3) or in Robbins, who claims that 'individuals can arrange their preferences in an order, and in fact do so' (Robbins 1935, 78). This account presupposes a great deal of agreement about these matters of mind. But what guarantees such a common agreement? Defenders of this uniformity of folk psychology can point to the relative success we have in predicting other people's behaviour. To those who doubt this success they might point out the many activities we only engage in because we can predict with great accuracy the behaviour of others: road traffic, contractual agreements, games of chess, etc. To those who claim that 'folk' accounts of behaviour in these situations is trivially true, they might point out the behaviour of the mentally ill, in particular of paranoid or autistic people and the seriousness of a lack of such a capacity.[15]

This conviction goes as far as to claim that the principles of folk psychological theory represent the fundamental rational principles of human conduct.

> I think everyone does subscribe to these principles, whether he knows it or not. This of course does not imply that no one ever reasons, believes, chooses, or acts contrary to those principles, but only that if someone does go against those principles, he goes against his own principles. (Davidson 1985, 351)

This is a bit fast. While all theories of this sort arise from a common need – to explain and predict human behaviour – they do not all have a common structure and content. To unreflectively speak of 'the' folk psychology is thus misleading.

The argument against a common structure and content of folk psychology starts with pointing out the ostensible lack of a source of this purported agreement. If all mentally healthy people engage in behaviour-explaining and -predicting activities, and I do think that they do, then the question is how they acquired the capacity

[15]It has been argued that autism is the pathological lack of access to folk psychology and hence the inability to understand and predict the actions of others. Compare Baron-Cohen, 1995.

to engage in these activities? Three potential answers offer themselves: they could have been instructed, they were endowed at birth with this capacity, or they could have learned from their own experience. The first option is easily dismissed: Children are certainly not taught by their parents in the way Jones supposedly taught his fellow human beings in Sellars' myth. Few people, after all, could readily produce a summary of folk psychology; so parents would be hardly able to explicitly instruct their children in something that they do not know how to express.

The second, nativist, option is more of an answer to reckon with. In analogy to other theories of innate capacities, it could be argued that the folk psychological capacities are part of the innate endowment bestowed upon every human being at birth.[16] Obviously, it is not the case that the full capacity of folk psychologising is innate – babies are not from birth onwards able to predict and explain others' behaviour. Nativists instead argue that certain aspects of mental capacities are innate, while other aspects that complete these capacities to their operative stage are acquired through empirical learning. The tradition following Chomsky's generative grammar focuses on the innateness of underlying rules; while followers of Fodor's nativism focus on the innateness of concepts. In both cases, the innate structures need 'filling in' by experience. The linguistic experience of a child in an anglophone family is captured in a pattern governed by the rules of generative grammar; this pattern then enables the child to speak English. Fodor's concepts are dormant and require experience of particular sensations as a trigger – the sensation of red, for example, triggering the dormant concept of 'red', 'colour' and 'not-red'. Similarly, it could be claimed that the 'rules' underlying folk psychology – be it the trait associations or the causal relations between mental properties – are innate, but in need of patterns of behaviour to fill them with; and the concepts of 'desire', 'belief', or character traits could be innate but dormant and waiting for experiences that required their employment.

It could be argued that the apparent correlation between physiological birth-defects and the lack of a folk psychological ability (see footnote above) supports the innateness of folk psychological rules and concepts. But this impression is wrong: for such an argument

[16] I do not know of such a position concerning folk psychology in particular; hence I will take the arguments from broader positions of nativism.

it has to be assumed that the defect affects the postulated folk psychological apparatus directly, and not the general abilities to sense other people's behaviour and learn about its regularities. But any such assumption would beg the question and therefore invalidate the argument.

The argument for these positions rather follows the so-called 'poverty of the stimulus' scheme. Nativists point to the discrepancy between our experience and the ideas, concepts, principles, etc. that we eventually come to have, and argue that this discrepancy cannot be fully explained by the empiricists' notions of learning. It follows that we must be contributing something substantial 'from our side' in the construction of knowledge, and this is what is innate. Empirical evidence seems to give some credence to these arguments in the case of linguistic capacities. The grammatical rules that modern linguistics has identified to govern natural languages are too complex and non-obvious to be known by mere casual reflection on one's own linguistic practice. Instead, their application seems to require a very sophisticated technical apparatus; an apparatus whose existence cannot be explained in terms of generic requirements on communication (Compare Chomsky 1988). Further, the linguistic evidence available to children severely underdetermines any theory of grammar. Many different grammars would be compatible with this evidence; hence it is not explicable from the available evidence alone, that children at a very early age learn and employ a highly homogeneous grammar.

The same arguments do not *mutatis mutandis* apply to folk psychological capacities. First of all, folk psychology has not been investigated to an extent that a secure body of folk psychological rules – comparable to the rules of generative grammar – could be presented. Thus the argument from its high complexity fails. Further, folk psychological capacities are acquired over a much longer time of child development than linguistic capacities are. For example, experiments show that children before the age of four lack a concept of belief, or at least lack the capacity to make allowances for false or differing beliefs in other people, while most children from the age of five already master the rules of grammar and meaning fully. Thus, the acquisition process is supported by a much longer exposition to behavioural evidence than the parallel grammar-acquisition by linguistic evidence. Last, while children exhibit a highly homogeneous

knowledge of grammar, the same cannot be said about folk psychology. This difference might just arise from the relative difficulty in observing folk psychological practices in comparison to linguistic practices; but so far no argument for the existence of a particular homogeneous folk practice is made.

The view I want to present here is an alternative to the explicit learning and nativist accounts. In deviation from Sellars' myth, it contends that every mentally healthy human being 'invents' the theory again for him or herself. It seems unproblematic to most of us that children learn about the external world, its properties and the individuation of its objects. We believe that they do so without their parents giving them a lecture in folk physics (which, again, most parents couldn't), and without them having any sort of innate knowledge of the world, its objects and properties. So if children can develop a complex conceptual framework about the external world, why should they not be able to construct a similar framework to understand and predict other people's behaviour, without explicit teaching or tacit internal knowledge? In this sense, everybody is a little Jones, having insight into the necessity of explaining and predicting his fellows' behaviour with a richer apparatus than mere $S - R$ regularities.

This view implies two consequences. First, common agreement is far from guaranteed. The 'learning by doing' approach, with ample adjustments along the way, will secure some sort of convergence, but there are enough loopholes and space for maneuvering to leave everybody with their own peculiar psychological theory. Mutual understanding is secured by partial overlap, not by identity.

Second, any such theory is highly fallible. Even though people will adjust their theories in light of failed predictions, their experience (and their methodology) will be highly domain-specific and thus do not suffice to guarantee a correct theory. Under the account proposed here, people in the majority will hold theories that yield roughly correct predictions which are non-robust and are based on regularities widely off the mark. Anecdotal evidence gives plenty of examples of what sort of funny psychological preconceptions people have about other people.

Thus, we shouldn't *rely* on the form nor the content 'folk intuitions' when constructing a scientific cognitive theory of human

behaviour. The 'commonly agreed platitudes' are not a guaranteed material for this purpose. This is an important result for the methodology of microeconomics, where it has been prominently claimed that the fundamental axioms of a theory of behaviour can be directly taken and justified from our common sense knowledge.

To the contrary, I propose that any such theory is constructed conjecturally. The psychologist might start with what she thinks are her common-sense psychological notions of causal relations between sense impressions, behaviour and mental properties of the subject she wants to ascribe the mental properties to; she further might construct it according to criteria of simplicity and coherence. The conjectured theory is then tested against all observations of an agent's behaviour in specific environments. I will discuss the structure of such a conjectured theory in the next subsection.

2.3.2 Mental Properties as Theoretical Terms

In the last subsection I argued that a cognitive theory of behaviour cannot hope to obtain its form nor its justification from folk psychological intuitions. Instead, the form of the theory will be a collection of lawlike generalizations – regardless of the form of our folk intuitions – and the content of the theory will be conjectured, instead of being based on allegedly correct folk intuitions.

What does the structure of such a theory look like? An answer can be found in Frank Ramsey's account of theories, later to be elaborated by Lewis (1970, 1972) and Balzer et al. (1987). In response to the verificationist program of the 1920s, Ramsey developed his own account of the relation between theoretical and observational terms. His first insight was that useful theories in general used terms that were too complex to be directly reducible to observational concepts.[17] Instead of trivially tying each theoretical term to an observable event or object, a theory takes an observation as evidence for the truth of a conjunction of sentences involving theoretical terms.

[17] Maybe fascinated by the newly introduced Goedel numbers, Ramsey decided to represent the terms of the theoretical and non-theoretical systems as numerical functions, whose arguments were interpreted as instances of space-time coordinates. He then found that the functional representation of the theoretical terms generally required more arguments than the non-theoretical ones (that they had a 'higher degree of multiplicity'), and derived from this that the theoretical terms were not directly reducible to the non-theoretical ones. 'Such an increase of multiplicity', Ramsey claimed, 'is, I think, a *universal* characteristic of useful theories (Ramsey 1929, 122).

The architecture of theories T suggested by Ramsey consists of a theoretical or formal and a T-non-theoretical part.[18] In the theoretical part, the theoretical terms are interlinked via the theory's *axioms*. These axioms restrict the co-existence of theoretical terms, or stipulated the existence of one concept from the presence of another. The whole structure, not the individual terms, is then connected to the T-non-theoretical terms. If the theory structure fits these T-non-theoretical terms, the theoretical terms would obtain their meaning from exactly the relations they were put in; if it didn't fit, the theoretical terms would not have meaning at all. Once the theoretical terms had acquired meaning, they would in addition to that provide information about the systematic behaviour of the phenomena subsumed under the theory. Examples of this form of theory construction can be found in both the social and the natural sciences. I will give two illustrations from chemistry and biology, before discussing theories of behaviour from this perspective.

The theory of chemical bonds proceeds from a number of key observations. Substances can be identified by characteristic properties like colour, smell, density and structure. Bringing them together, substances are sometimes observed to transform in reactions, although closer observation reveals that the total mass of the *reactants* does not change. There is a particular subgroup of reactions, called chemical analysis, where the number of substances after the reaction is higher than before (electrolysis, heating under exclusion of other substances, exposure to highly reactive substances). It has been observed that the analysis of a specific substance always yields the same compounds, and that it always yields them in exactly the same proportions. In those cases where the analysis of *different* substances yields the same compounds, it is found that these compounds exist in the same proportions, or in exact multiples thereof.

Starting with these observations, the theory hypothesises some products of chemical analysis as 'elements' – as the last products of analysis. It then defines the smallest unit of an element as the

[18] Diez correctly points out that the distinctions 'T-non-theoretical/T-theoretical' and 'observational/non-observational' are it not identical – contrary to the claims by Ramsey and Carnap. 'The former is local, relative to theories (a concept may be T-non-theoretical but T'-theoretical), the latter is global (a concept – perhaps at a given time – is observational or it is not, period)'(Diez 2002, 15). I agree with this view and will from now on use 'T-non-theoretical' or 'extra-theoretical' to denote those concepts which the theory is meant to account for.

atom, and it hypothesises a reaction to consist in a process where an atom of one element combines with a specific number (or at the least, with specific numbers) of atoms of another element. The thus described *valence* property of an element was historically measured as the number of hydrogen atoms that would combine with one atom of the element in question.

The unified theory of ionic and covalent bonds, first developed by Gilbert Lewis in 1916, models the valence of atoms as a property of electronic structures and inter-atomic forces. All elements of the periodic table are uniquely characterised by the number of electrons in their atomic structure. Further, each element in each period is characterised by the number of electrons on its outmost shell. The maximum number of electrons on that shell is eight, as found in the noble gases. Now the theory postulates the tendency of each atom to transfer (in case of the ionic bond) or share (in case of the covalent bond) electrons with other atoms in its environment until both have reached the electronic composition characteristic of the nearest noble gas atom in the periodic table (either in its period or the period above). This theory rudiment is the basis for all prevalent explanations of chemical bonding – and thus for chemical reactions – in contemporary chemistry.

The theory constructs a mechanism that allows one to systematically relate a range of pre-reaction substances to a range of post-reaction substances. It does so by postulating a hierarchy between substances, such that some substances ('elements') turn out to be the most primitive, out of which all other substances are composed. It then associates each element's atom with a particular property (electrons in the outmost shell) and then specifies a process between atoms with particular realisations of that property. This process is captured in a set of axioms of the sort 'if an atom with two electrons on its outmost shell meets another atom with six electrons on its outmost shell, then the former transfers its two electrons to the latter'. The theory predicts from this process and information about the pre-reaction situation the existence of certain substances. The substances and their properties in the pre-reaction situation and the ones in the post-reaction situation are the only observable evidence for the theory. The postulated properties only exist – or to put it more precisely, the employed predicates have meaning only – if the theory fits the data. The postulated mechanism itself

is merely conjectured and not directly confirmed by any evidence. Even extremely refined measurement instruments (X-ray diffraction, electronic microscope, scanning tunnelling or the particle chamber) only reveal the existence of atoms or electrons; they do not give any information about the processes involved in chemical bonding.

A second example of a Ramsey-type theory is found in biology. Here, the botanist distinguishes between different species (a species being defined as the group of organisms that can successfully breed offspring) and records differences in properties between members of one species. Exemplars of the common garden pea, for example, bloom either white or purple. It is further observed that the pea plant is dominantly self-pollinating, and that its descendants tend to have the same trait as its parent.

For the explanation of this observation, Mendel proposed that the visible phenotype of the exemplar is determined by a theoretically conjectured genotype. A gene, however, cannot be directly read off from the phenotype. Instead, genes determine a phenotype only in pairs. If gene pairs consist of mixed traits, the gene which determines the phenotype is said to be the dominant gene. A particular pea plant might bloom white, but some of its descendants might bloom violet. In such a case, the pea is said to carry the dominant and the recessive gene.

Genes cannot be directly observed in organisms (and the reducibility of Mendelian genetics to molecular genetics is highly controversial – see Kitcher 1984), nor can they be defined in direct relation to any phenotypical property of an organism. Instead, the genotype of an organism is a theoretical concept, and it is attributed according to a pattern that subsumes information about the phenotypes of the organisms ancestors, the phenotypes of its descendents, and the phenotypes of the ancestors of that organism's with which it produced these descendents. To take the simplest possible example from Mendel's original field of application, the common garden pea is said to carry the recessive gene either if it has a white flower, or if it has a purple flower and its offspring, stemming from self-pollination, has a white flower with probability 0.25. Mendelian genetics fits data of this sort extremely well; hence the predicate 'has genotype X' identifies a property if it is used in accord with the theory. But this doesn't say anything conclusive about the realism

of the conjectured mechanism and attributions itself.

Similarly, mental properties are attributed to agents to account for their pattern of choices. Nowhere are the references of these predicates directly observed; and nowhere can a causal relation between a mental property and an observable stimulus or choice be ascertained before the whole of the theory has been applied to the behavioural phenomena. Davidson expresses this characteristic of mental properties as follows:

> 'if we were to ask for evidence that the explanation [of behaviour in terms of reasons] is correct, this evidence would in the end consist of more data concerning the sort of event explained, namely further behaviour which is explained by the postulated beliefs and desires. Adverting [mental properties] to explain action is therefore a way of fitting an action into a pattern of behaviour made coherent by the theory.' (Davidson 1975, 159)

To assign mental properties to an agent is therefore not just an epistemological exercise: the result of finding out about the agent's mental constellation. It is also a determination of the meaning of mental predicates: if a theoretical framework fits the choices an agent has made, then the mental predicates assigned to the agent through that theoretical framework obtain their meaning from their role in the true framework.

Lewis (1972) employs this general idea for his discussion of the meaning of mental predicates. According to him, any such theory consists of three types of statements: (1) conditionals expressing causal relations between stimuli type descriptions and mental state descriptions (mental predicate ascriptions); (2) conditionals between mental state descriptions of different kinds; and (3) conditionals between mental state descriptions and behaviour type descriptions. More formally, these three types of statements are expressed as follows, with x being the person, and M_i being a mental predicate of type i.

(1) $(x)x$ perceives [something]$\rightarrow M_i(x)$
(2) $(x)M_i(x) \wedge M_j(x) \rightarrow M_k(x)$
(3) $(x)M_i(x) \wedge M_j(x) \rightarrow x$ does [something]

where $\rightarrow$ denotes a causal relation. The mental predicates and the

interrelations between them constitute the T-theoretical part; the T-nontheoretical part is constituted by the assumed sensory perception and the observed choices between options of an assumed calibration. I quoted typical examples for the relations between theoretical concepts in section 2.3.1; here they are again for illustrative purposes:

> $(x)(p)[(x$ fears that $p) \rightarrow (x$ desires that $\neg p)]$
>
> $(x)(p)(q)[((x$ believes that $p) \wedge (x$ believes that (if p then $q))) \rightarrow$ (barring conflicting desires, distractions, etc., x believes that $q)]$
>
> (Churchland 1981, 71)

The theory T – if true – covers the causal relations of mental properties. Now instead of specifying first the meaning of all predicates included in T and then determining whether the theory is true, Lewis suggest that we can first check whether the theory is true and then determine from the theory what the predicates mean. This strategy is pursued with the help of the Ramsey sentence, in which all mental predicates M_i are replaced with variables bound in existential quantifiers:

$$\exists M_1, \exists M_2, \ldots, \exists M_n : T(M_1, \ldots, M_n)$$

The Ramsey sentence says that there is at least one realization of T, with some constellation of mental predicates $M_1 \ldots, M_n$. The meaning of any predicate M_i is then determined as M_i : $(\exists M_1)$, $\ldots, \exists M_{i-1}, M_i, \exists M_{i+1}, \ldots, \exists M_n$: $T(M_1, \ldots, M_n)$. This of course implies that M_i does not have any meaning at all if T is not true for any such constellation. But if T is true for some specification of predicates, then that specification of predicates names the properties which make the theory true. The Ramsey-Lewis strategy thus leads to the specification of mental properties as causal roles between stimulus input and behavioural output, assuming the causal structure in which they occur.

This provides the mere bare bones of an architectonic of theories of behaviour. In the following section, I will discuss some concrete examples of these theories, and explain what implication their interpretation under a Ramsey-Lewisian framework has.

2.4 Cognitive Theories in Action

In the last subsection, the overall structure of all cognitive theories of behaviour was discussed. But how is this structure filled? What sort of theoretical concepts are conjectured, and what sort of axiomatic relationships hold between them? Naively, one might think that all conjectures are random concoctions of some rich syntax, only governed by rules of consistency and non-repetitiveness. The resulting collection of possible theories might compete with Borges' *Library of Babel*; but it is certainly not the way social scientists arrive at their conjectures. Rather, they can help themselves from an affluent set of proposed theories of deliberation, that have been found plausible and convincing – without, of course, being necessaily true. In the following, I will briefly review three types of these theories: those that attribute the will an autonomous role in deliberation; those that model deliberation as formed only by passions and beliefs; and those that reduce all passions to one form of motivation: desire, want or preference. I will then discuss two theories of the latter type, Ramsey's version of Bayesian Decision Theory and Loomes' and Sugden's regret theory.

2.4.1 Will, Passions, Desire

From Plato to Kant, the will has been assigned an important role in theories of human agency. Neglecting many differences between authors, their common denominator was the claim that the will has an autonomous influence on human behaviour. The will was understood as the capacity to become motivated to act based on deliberations about what actions would be justified. Aquinas, for example, maintained that the practical intellect determined which action should be performed; the will as a 'rational appetite' thus represented a motivating force in competition with the passions. Kant, in a different theory, cast 'Wille' as the capacity for autonomous self-legislation, restricting or overriding the motivating force of the passions. In contemporary philosophy, theories of human agency that cater for a direct influence of rationality as a motivating force still play an important role – for example Sen (1977), who stresses the role of commitment as non-welfare-maximizing motivation; Nagel (1976), who argues in that morality is concerned with the will; and Bratman (1987) who claims that intention cannot be assimilated

under the concept of desire.

In contrast to this, Hume conceived of the will as 'nothing but the internal impression we feel and are conscious of, when we knowingly give rise to any new motion of our body, or new perception of our mind' (Treatise in 'Of liberty and necessity'). Will as an autonomous motivation was thus eliminated, and the passions gained importance as the only motivating forces. Consequently, the philosophical interest of the passions rose. Descartes, although still committed to the notion of will as an autonomous motivating faculty, developed a system in which six passions were conceived of as primitive (wonder, desire, love, hatred, joy and sadness), while 'all the others are either composed from some of these six or they are species of them' (On the Passions of the Soul P69).

In the second part of the Treatise, Hume fleshed out the notion of passion. A passion is a reflective impression; it arises not immediately and directly from some impression of the senses, but through the mediation of the idea of that impression.

> An impression first strikes upon the senses, and makes us perceive heat or cold, thirst or hunger, pleasure or pain of some kind or other. Of this impression there is a copy taken by the mind, which remains after the impression ceases; and this we call idea. This idea of pleasure or pain, when it returns upon the soul, produces the new impression of desire and aversion, hope and fear, which may properly be called impression of reflexion, because derived from it. (Hume 2000, 7-8)

As any impression, a passion is an original existence, a state of a person analogous to other physical states of the person.[19] The fact that it arises through the mediation of an idea only specifies the causal conditions of its existence; it does not mean that it can be reduced to other mentally represented components, like ideas or other impressions. Hume even doubted the possibility of a comprehensive but non-reductive analysis. As he said about the (indirect) passions of pride and humility:

[19] 'A passion is an original existence, or, if you will, modification of existence, and contains not any representative quality, which renders it a copy of any other existence or modification. When I am angry, I am actually possest with the passion, and in that emotion have no more a reference to any other object, than when I am thirsty, or sick, or more than five foot high.' (Hume 2000, 415)

> . . . it is impossible we can ever, by a multitude of words, give a just definition of them, or indeed any of the passions. (Hume 2000a: 277)

In this sense, passions are primitive, irreducible entities of the mind. Hume nevertheless ventured to undertake an extended analysis of passions, carefully avoiding any reductive attempt. His analysis, instead, relied upon a dual classification scheme. On the one hand, passions are distinguished by their felt intensity. Those commonly experienced at low intensity (or not experienced at all) are considered the calm passions, while those experienced at high intensity are the violent passions. Hume made it clear that experienced intensity of a passion is not at all correlated to the strength of its effect.

On the other hand, passions can be classified in terms of the causal conditions under which they come about. The *primary passions* arise from 'a natural impulse or instinct' without any intermediary role of pleasure or pain. Their violent type is manifested in hunger and lust, for example, their calm type in benevolence, resentment and the love of life. The *secondary passions* are aroused by the preceding impression of pleasure or pain or the idea of these. Of these, the *direct passions* are those that are from pleasure or pain immediately. Their violent type is manifested in desire and aversion, joy and grief, hope and fear, despair and security, while their calm type is exemplified in the basic moral and aesthetic sentiments. The *indirect* (secondary) *passions* are caused by pleasure or pain, but the object of these passions is not identical with that cause. For example, hatred is an indirect passion of the violent type: caused by the pain of a particular event or series of events, but directed at a person, not that event. Similarly, this relation holds for pride, humility and love, or in the calm variant, the approval and disapproval of persons. According to Hume, nothing further can be said about the constitution of the passions, and he consequently spends most of Book II of the *Treatise* analyzing the specific causal conditions under which indirect passions come about.[20]

[20]It is important to repeat that despite this categorization, which consists in a causal analysis and the comparison of similarities between the different passions, Hume thinks of *all* passions as simple and non-reducible. This interpretation is supported by Ardal: 'A simple perception cannot be analyzed into distinct parts. Yet Hume thinks that it can be characterised by pointing outs its similarity to other simple perceptions or its difference from them. One can also state the conditions under which it is found to arise, or, in other words, its causal conditions. Thus, for Hume, a simple perception is not just something that can only be

Little remains of Hume's non-reductive theory of agency in contemporary philosophy and social sciences.[21] Instead, a unified account of the passions became dominant, first through psychological hedonism as exemplified by economists like William Stanley Jevons or Francis Edgeworth, and later through the representation of all types of motivations by a preference order – pioneered by Vilfredo Pareto, and developed into the now orthodox economic position by Kenneth Arrow and Gerard Debreu. This unified account has found entrance into the philosophical literature through Donald Davidson's work, who speaks of *pro-attitudes* towards state of affairs, under which are included 'desires, wantings, urges, promptings, and a great variety of moral views, aesthetic principles, economic prejudices, social conventions, and public and private goals and values' (Davidson 1963, 4) and who just a few pages later claims that 'it is not unnatural, in fact, to treat wanting as a genus including all pro attitudes as species' (ibid. 6). Whether one speaks of wantings, desires or preferences then mainly is a formal question; contentwise, they all express a disposition to behave under specific conditions.

The three positions on agency sketched above are *prima facie* distinguished by the number of motivational forces they allow. As I have argued above, they therefore have to be distinguished by the whole of their theoretical architecture. In particular, in the terminology of the Ramsey-Lewis account of theories, they are to be distinguished by their definitions – the 'contact points' between theory and non-theoretical terms – and by their axioms – the interrelation between the theoretical concepts themselves. In the following, I will discuss two cognitive theories that employ a unified account of motivation.

2.4.2 Ramsey: Bayesian Decision Theory

Ramsey's account of decision starts out with three assumptions: that the definitions must only refer to observed actions or reported hypothetical choices, but not to any other introspective data (compare also his quote in section 2.2.1); that choices are to be inter-

pointed to or given a name. Many things may be predicated to it. I shall, indeed, emphasise that the bulk of the second book of the *Treatise* is concerned with stating the causal conditions for the emergence of simple impressions, and indicating various similarities between them.' (Ardal 1989, 12).

[21] For an account of the relevance of Hume's notion of the passions for his economic and political thought, see Gruene and McClennen (forthcoming).

preted as choices over gambles; and that choices are the outcome of only two psychological factors, beliefs and desires, where the concept of desire is cast as a preference ordering and the concept of belief as a probability measure.[22] The problem of such a theory is nicely summarised by Donald Davidson:

> Choices between gambles are the result of two psychological factors, the relative value the chooser places on the outcomes, and the probability he assigns to those outcomes, conditional on his choice. Given the agent's beliefs (his subjective probabilities) it's easy to compute his relative values from his choices; given his values, we can infer his beliefs. But given only his choices, how can we work out both his beliefs and his values? (Davidson 1974a, 145)

I will give a rough account of Ramsey's solution to this problem.[23] Ramsey takes as evidence hypothetical choices, in particular indifference judgments. An agent is supposed to have a preference over all available options, and she is supposed to have beliefs 'about everything'. Then, to measure her beliefs and evaluations, an agent is offered two gambles, $(a; E; b)$ and $(b; E; a)$, where the options a and b are of different value to the agent.[24] E is an event such that the agent is indifferent to whether E is realised or not. The probability of E is defined to be 0.5 iff the agent is indifferent between the two gambles. This definition is plausible insofar that if the agent judged E to be of different probability, she would – under the given interpretation – be indifferent between a smaller and a larger expectation of her preferred option. This would be incoherent.

Then, in a next step, all options are assigned numbers to reflect their position in the preference ordering. This numerical assignment only represents the order in which they stand. In a further defini-

[22] Ramsey cautions about the approximateness of the theory that he is about to use: 'In order therefore to construct a theory of quantities of belief which shall be both general and more exact, I propose to take as a basis a general psychological theory, which is now universally discarded, but nevertheless comes, I think, fairly close to the truth in the sort of cases with which we are most concerned. I mean the theory that we can act in the way we think most likely to realise the objects of our desires, so that a person's actions are completely determined by his desires and opinions. This theory cannot be made adequate to all the facts, but it seems to me a useful approximation to the truth in the case of our self-conscious or professional life, and is presupposed in a great deal of our thought' (Ramsey 1926, 69).

[23] See Bradley (2001) for a detailed account and a representation theorem for Ramsey's measures.

[24] Read $(\$x; E; \$0)$ as the gamble that if E is the case, you receive $\$x$, if E is false, you receive \$0.

tion, Ramsey gives meaning to the numerical differences between two options. He defines that the numerical differences between options a and b and between c and d are the same iff the agent is indifferent between the gambles $(a; P; d)$ and $(b; P; c)$, where again P is assumed to be evaluatively neutral and has probability 0.5. This definition is plausible insofar that if the agent judged the evaluative differences to be non-identical, she would – under the given interpretation – for example trade b for a more readily than d for c, despite the fact that she was indifferent between gambles that yielded a or d and b or c in equal parts. This would be incoherent.

With the help of this definition, as well as a few further axioms, Ramsey is able to provide a utility measure $u(\cdot)$ for all options and gambles between options such that $u(a) > u(b)$ iff a is preferred to b and $u(a) - u(b) = u(c) - u(d)$ iff the difference between a and b is the same as that between c and d.

Last, Ramsey presents a definition for belief in all kinds of propositions, including those that are not ethically neutral. Given that an agent is indifferent between an option a and a gamble $(b; P; c)$, his belief in P is determined by the utilities for a, b and c : $Pr(P) = \frac{U(a)-U(c)}{U(b)-U(c)}$. This definition is plausible insofar that if the agent judged P to have a different probability, she would – under the given interpretation – be indifferent between options of different expected utilities. This would be incoherent.

Ramsey's measurement of utilities and beliefs illustrates well the gist of the Ramsey-Lewis theory. The method takes hypothetical choices as given data, and superimposes a theoretical structure to subsume this data under a systematic pattern. Davidson summarises this as follows:

> The explanation of a particular preference [standing here for hypothetical choices] involves the assignment of a comparative ranking of values and an evaluation of probabilities. Support for the explanation doesn't come from a new kind of insight into the attitudes and beliefs of the agent, but from more observations of preferences [choices] of the very sort to be explained. In brief, to explain (i.e. interpret) a particular choice or preference, we observe other choices or preferences; these will support a theory on the basis of which the original choice or preference can be explained. Attributions of subjective values and probabili-

> ties are part of the theoretical structure, and are convenient ways of summarizing facts about the structure of basic preferences; there is no way of testing them independently. (Davidson 1974a, 146)

Ramsey's theory neatly exemplifies such a pattern. The (extratheoretical) data – real and hypothetical choices over certain and prior specified outcomes as well as choices over gambles with specified outcomes – is subsumed under the theoretical pattern of expected utility maximization. The hypothesis consists of definitions: of choices over certain outcomes in terms of preferences over those outcomes, and of choices over gambles in terms of an expected utility index. It further consists of a series of axioms that restrict the theoretical terms: the transitivity of the preference ordering over certain outcomes; the interrelation of preferences between gambles, preferences over certain outcomes and the subjective probability of neutral events; the interrelation of preferences between gambles probability of the neutral event and the numerical evaluative differences between gambles; and the interrelation between numerical differences between gambles, probabilities of gambles, preferences over outcomes and the expected utility index. These two realms of the theory are illustrated in figure 2.1.

Through its definitions, the theoretical terms are connected to the extra-theoretical data depicted in the lower box. This connection only runs from theoretical to extra-theoretical terms, not the other way around. I therefore speak of 'fitting the data' instead of defining the theory in terms of it. Now the theoretical terms 'preferences over outcomes' and 'preferences over gambles' are in turn governed by a regime of axioms that interconnect all theoretical terms. As I showed above, the interconnection is established by construction. The 'probability of the neutral event' is constructed out of the 'preferences over outcomes' and 'preferences over gambles'. The 'numerical difference between gambles' is constructed out of the 'probability of neutral events' and the 'preferences over gambles', which in turn defines the 'probability of gambles'. 'Probability of gambles' in combination with the 'preference over certain outcomes' constructs the 'expected utility of gambles', which in turn defines the extra-theoretical terms of 'choices over certain outcomes' and 'choices over gambles'. The theory now has come a full way around; because the 'expected utility of gambles' again defines an

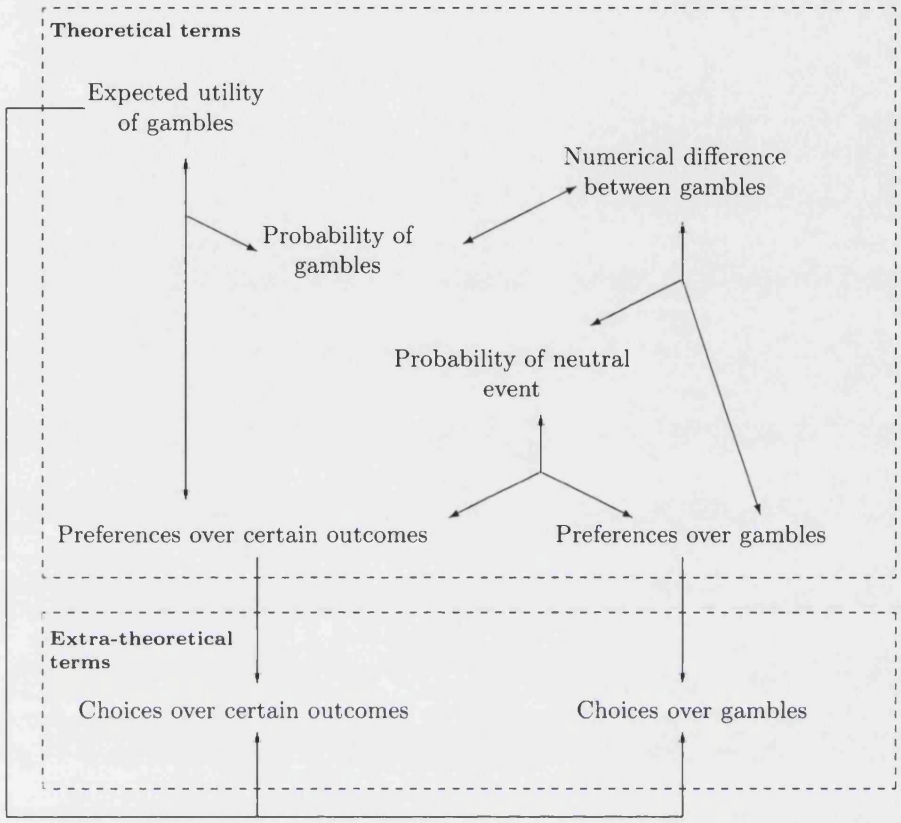

Figure 2.1: The basic architecture of Ramsey's Theory

extra-theoretical term, the theoretical terms between it and 'preference over gambles'/'preferences over outcomes' have to construct these terms in such a way that they can fit the data. For this reason the argument for the construction of each theoretical term was: 'if this term was constructed differently (e.g. if the probability of the neutral event E was not 0.5 in the case where the agent was indifferent between $(a; E; b)$ and $(b; E; a)$) then the theory would make incoherent predictions'.

Ramsey's theory is tested in two ways: (i) Existing evidence – that was not directly involved in the theory's construction – is subsumed under it, beliefs and utilities derived from it under obedience

of the axioms governing these measures. (ii) From the measures, with increased generality, choice behaviour is predicted. If any prediction does not fit with actual evidence, the theory is incoherent. In the synchronic case, the theory is thus shown to not fit the data and has to be reconstructed. In the diachronic case, where behavioural evidence from a later time is used, the question how to deal with the theory is more complex. This will be discussed in chapter 4.

If my account of the construction Ramsey's theory is correct, then the theory does not have any claim to the realism of the mechanism it uses. Available data severely underdetermines the theoretical structure; even from a good fit of the theory to the data we cannot hope to derive an argument for the reality of preferences and beliefs in human agents that would satisfy Ramsey's axioms.

As it turns out, Ramsey's theory – or rather the further developments it has enjoyed by the hands of von Neumann/Morgenstern and Savage – does not enjoy such a perfect fit even with the data available. A number of decision situations was developed for which these theories could not account; most prominently amongst them the Allais and the Ellsberg Paradoxes.[25] Two responses to the bad fit are of particular interest here. The one is to expand the theory to include some of the extra-theoretical concepts in its domain. Broome responds this way to the Allais Paradox by suggesting that the theory should take into account the individuation of outcomes. While the original theory from Ramsey to Savage takes the choices over specified outcomes as extra-theoretically given, a refined theory would determine the individuation as part of the utility maximization mechanism (Compare Broome 1991, 95-100). The second response is to change the mechanism of the theory itself, attempting to achieve a better fit. Many of these attempts exist; I will focus here on an alteration made in direct response to the Allais's Paradox – regret theory.

2.4.3 Regret Theory

Regret theory as an alternative to Bayesian Decision Theory was proposed independently by Bell (1982) and Loomes and Sugden (1982). In the following I will discuss Loomes and Sugden's version.

[25] The paradoxes are well described in many textbooks, amongst them in Hargreaves Heap et al.(1992), so that I do not need to describe them here again.

Regret theory proposes an alternative motivation for actions. Instead of choosing actions according to their expected utility, regret theory claims that agents choose an action according to the experience of regret and rejoice expected from it.

Loomes and Sugden start by defining a choiceless utility function $C(\cdot)$ over all possible consequences. $C(\cdot)$ represents the agent's evaluation of prospects she experiences without having chosen them. If an agent chooses between two actions, she evaluates the prospects she experiences as an effect of her action differently from $C(\cdot)$. Let the consequence of an action A_1, given that the state j of the world occurred, be x_{1j}. Then the evaluation of action A_1 does not only depend on x_{1j}, Loomes and Sugden claim, but also on x_{2j}, the consequence of the alternative action A_2. By having chosen A_1, the agent forgoes the outcome of A_2 if j occurs; the possible *relative* loss, for which she is partly responsible herself, will influence her decision in addition to the absolute value of the outcomes attained. Each consequence of an action A_i, conditional on a state of the world j, where A_i is chosen over action A_k, is therefore evaluated by a function M of the choiceless utility indices of the consequences of both actions, x_{ij} and x_{kj}: $m_{ij}^k = M[C(x_{ij}), C(x_{kj})]$. The thus defined *modified utility function* m_{ij}^k is proposed as a measure of action evaluation that takes regret and rejoicing in pairwise action comparisons into account. Loomes and Sugden then impose certain restrictions on M.[26]

In a decision situation between two actions and a partition of the world into states $i, j, k, \ldots$, an action A_i is thus evaluated by its *expected modified utility* $E_i^k = \sum_{j=1}^n p_j m_i^k j$. This evaluation is different from classical expected utility frameworks as long as $m_{ij}^k \neq C(x_{ij})$.[27] One interesting consequence of this difference is that the expected modified utility index defined over actions does not necessarily represent a transitive preference order over these actions. Thus if the modified expected utility would define choices over gambles directly, the theory would be incoherent. Instead, Loomes and Sugden propose an additional theoretical term, *weighted* expected modified util-

[26] In particular, they assume that (i) if the evaluations of the consequences are identical, there is no regret and $m_{ij}^k = C(x_{ij})$; (ii) Increase in the evaluation of the foregone consequence increases regret and hence decreases modified utility: $\partial m_{ij}^k / \partial C(x_{kj}) \leq 0$; (iii) Modified utility rises with the evaluation of the obtained consequence: $\partial m_{ij}^k / \partial C(x_{ij}) \geq 0$.

[27] Given the above-mentioned restrictions imposed on M, the difference between the two theories can be measured as the *regret-rejoice function* R : $m_{ij}^k = C(x_{ij}) + R[C(x_{ij}) - C(x_{kj})]$.

ity E_i^S, which is the average of the modified expected utilities of A_i in comparison to all other possible actions $A_k \in S$. The maximization of this parameter allows an unambiguous choice between all actions in S. The architecture of regret theory is illustrated in figure 2.2.

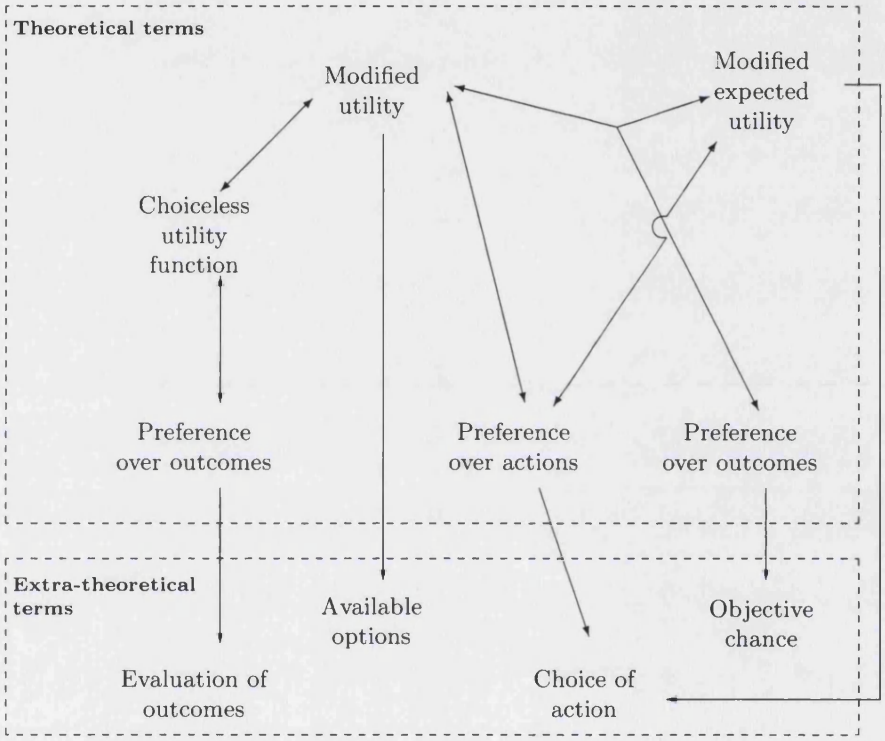

Figure 2.2: The basic architecture of Regret Theory

What is striking in comparison of the architecture of 2.1 and 2.2 is (i) that regret theory relies on a number of extra-theoretical data whose availability is strongly disputed. Loomes and Sugden express their belief 'that it is possible to introspect about utility, so defined, and that it is therefore meaningful to talk about utility being experienced in choiceless situations' (Loomes and Sugden 1982, 807). Additionally, they take the probability of states of the world as independently given. But (ii), it is remarkable that the interconnections

between the theoretical terms are weaker and less forked than in 2.1. Naturally, this allows the theory to be fitted to more versatile data; but it identifies a pattern that is not as powerful as Ramsey's, exactly because of its higher versatility.

I conclude from this that the development of cognitive theories of behaviour is evaluated by two standards. On the one hand, the constructed pattern should fit the available extra-theoretical data as well as possible; on the other hand, the more parsimonious it is, and the more the interconnections of the theoretical terms are branched, the more powerful the theory is, and the better it is able to explain and predict. Given the currently available (and justifiable) data, Ramsey's theory might be too tight, while regret theory might have some slack. Which of them – and the many other alternatives available – will be adopted, is a question at least partially dependent on further empirical research.

All these theories, however, are within the category of a unified concept of motivation – i.e. operate with only two mental properties, desires and beliefs. With the given data, there is no reason to get more complex than this – the current evidence base already underdetermines the dual theories, and even more complex theories could not be matched up by it.

Two possible future developments could change this diagnosis: first, the data base could be expanded, for example through the development of a better method of introspection than currently available. The discussion of Ericsson and Simon is instructive here, as well as Davidson's program of a unified account of thought and action. Second, arguments from other human faculties – e.g. morality – could lead to a notion of the autonomous will as their necessary condition. Nagel (1976) makes such a point, but I will not follow this argument any further here. Instead I conclude that so far belief-desire theories, without having any rock-bottom arguments for them, seem the most parsimonious and yet versatile candidates to fit the data. The question now is therefore not whether to expand on the mental properties, but which axiomatic framework is appropriate for the given ones.

2.4.4 Holism, Rationality, Intelligibility

A third response to the paradoxes mentioned above is to retreat to a position that proposes the normative validity of the theory, and to denounce any deviation from the theories predictions as irrational mistakes, that might occur locally, but are non-sustainable in the long run.[28] In a normative interpretation, the theory's axioms are interpreted as *principles of rationality.*

These axioms have sometimes been called rationality requirements. But when employing such a grand title, a few caveats are in order. First, these axioms only represent an absolute minimal form of rationality. They only regulate the interaction between mental states; no judgment is passed as to the rationality of any mental state by itself.

> No factual belief by itself, no matter how egregarious it seems to others, can be held to be irrational. It is only when beliefs [or other mental states] are inconsistent with other beliefs according to principles held by the agent himself . . . that there is a clear case of irrationality. Strictly speaking, then, the irrationality consists not in any particular belief but in inconsistency within a set of beliefs. I think we must say much the same about . . . other propositional attitudes. (Davidson 1985, 348)

Further, the axioms only restrict the interaction of mental properties in a static framework. The mental states are only judged consistent or inconsistent in relation to those other mental states which are present in the agent at the same time. Questions of dynamic consistency, that relate present to future mental properties – for example Quine's principle of conservation – are not covered at all in this simple framework. (An extension to that effect is developed in chapter 4).

Within these confines, however, it has been claimed that the axioms of the theory are rationality principles. A proponent of this view is Donald Davidson. He reminds us that the theory (and hence its axioms) are constructed such as to achieve a 'best fit'. It is supposed to subsume the observed actions under a *true* pattern – i.e. a pattern that will manifest itself in future behaviour again. For

[28] I will neglect here the arguments that the paradoxes as well exhibit the problems of the normative validity of the theory.

that to be the case, the ascribed theoretical terms must be largely *in agreement* with the observed evidence. Observing an agent's choice, we ascribe those desires and beliefs to her that would have motivated a fully rational agent to choose this way. A critic might object that we don't know whether this particular agent is rational at all, and that we hence beg the question in the ascription. Davidson's response is that the diagnosis whether an agent is rational or not presupposes that we have ascribed mental properties already. This is only possible under the assumption of *global* rationality. Thus while it might be possible to identify *local* irrationality, the condition for such a possibility is the global rationality of the agent.

> What justifies the procedure is the fact that disagreement and agreement alike are intelligible only against a background of massive agreement. (Davidson 1973, 137)

The mental properties exist only insofar as they are identified by the theory in terms of their causal roles in the production of behaviour. The theory, in turn, only exists on the basis of axioms that impose restrictions on the theoretical terms. With different axioms, the whole theory and hence the mental predicates that it ascribes would be different. But against a completely different background of mental properties it is not possible to identify a particular response as a violation of rationality. Thus Davidson thinks that a unified formulation of the axioms – variably called 'global rationality' or 'massive agreement' – is the necessary condition for the possibility of ascribing mental properties.

> . . . it is a condition of having thoughts, judgments and intentions that the basic standards of rationality have application. The reason is this. Beliefs, intentions, and desires are identified, first, by their causal relations to events and objects in the world, and, second, by their relations to one another. . . . these obvious logical relations amongst beliefs; amongst beliefs, desires and intentions; between beliefs and the world make beliefs the beliefs they are; therefore they cannot in general lose these relations and remain the same beliefs. Such relations are constitutive of the propositional attitudes. . . . Rationality . . . is a condition of having thoughts at all. The question whether a creature "subscribes" to the principle[s] is not an empir-

> ical question. For it is only by interpreting a creature as largely in accord with these principles that we can intelligibly attribute propositional attitudes to it. (Davidson 1985, 351-352)

In this view, global rationality is the condition for the assignment of mental properties, which in turn is the condition for the identification of (local) irrationality. This way, rationality is the condition for the possibility of irrationality.

> The methodological advice to interpret in a way that optimises agreement should not be conceived as resting on a charitable assumption about human intelligence that might turn out to be false. If we cannot find a way to interpret the utterances and other behaviour of a creature as revealing a set of beliefs largely consistent and true by our own standards, we have no reason to count that creature as rational, as having beliefs, or as saying anything. (Davidson 1973, 137)

This is in accord with the Ramsey-Lewis account of theories. The theory is to be constructed such that it fits the data best, and this involves the manipulation of the axioms in the appropriate way. As discussed in section 2.3.2, the application of the theory constitutes both its meaning and the knowledge we derive from it. This is what Davidson has in mind when he says that 'the methodology of interpretation is, in this respect, nothing but epistemology seen in the mirror of meaning' (Davidson 1975, 169). But what Davidson seems to claim – over and above the necessity to construct the axioms in such a way that the theory fits the data – is that this exercise of axiom-manipulation and theory-fitting leads to a unique set of axioms.

This is where I disagree. I argued in the last subsection that there are different variants of the belief-desire theories that seem equally applicable to the presently available evidence. Whether the extended (outcome-individuating) utility-maximizing theory or the regret theory is correct, or maybe another variant like Machina's utility theory without the independence axiom, is currently undecided. Behaviour can be made intelligible under each of these theories, however. From a specific theoretical perspective, it is therefore necessary to stick with the axioms – e.g. it is not intelligible to al-

low global preference intransitivity in classical utility maximization – but with the change of the theoretical perspective, the axioms also change. Davidson seems to admit this when he says:

> . . . we make sense of aberrations when they are seen against a background of rationality; but the background can be constituted in various ways to make various forms of battiness comprehensible. (352)

So if the background can be constituted in various ways, Davidson's earlier claim that everybody does subscribe to the principles of decision theory, because 'if someone does go against those principles, he goes against his own principles' is problematic. The axiom of any theory can only be seen as a rationality principle from the perspective of that theory. The transitivity of preferences is a rationality requirement only for those preferences that are part of a utility-maximizing theory. For another theory, say the Regret account, transitivity of preferences (at least over actions) is not a rationality requirement: under such a regime, an agent's choices are intelligible even if they are motivated by an intransitive preference ordering.

The theory-dependence of the rationality principles supports the conclusion of section 2.3.1. There I argued that a cognitive theory of behaviour cannot be directly derived from folk psychology. In this section I have argued that the axioms of any theory can be seen as rationality principles only in so far as they are necessary for the intelligibility of an agent's choices from the perspective of that particular theory. From a different or more global perspective, they might neither be necessary for the intelligibility nor rational. To claim that any theory models the folk intuitive notion of rationality therefore implies that this theory is identical to folk theory – a claim that so far has been difficult to prove. To claim that any theory models the global notion of rationality implies that the theory is true – a claim even more difficult to prove at present.

2.5 Conclusion

I have argued on the one hand against the possibility of ascribing mental properties through introspection, and on the other hand I gave reasons for why the initial reaction of the psychological profes-

sion, to discard mental properties altogether, leads to vicious circularities in the theory of human behaviour.

Instead, I proposed that mental properties are ascribed through a theory, with the evidence exclusively based on behavioural observations. This required that the correctness of the theory and the appropriateness of the mental states had to be taken care of at the same time. I objected that common sense notions were of no help in determining the correctness of both; and I further disputed the claim that there are *a priori* reasons for any specific structure of the theory.

The approach that I advocated in its place was to take an appropriately designed theory and employ it as a test hypothesis. The attempt should be to subsume all observed behaviour under a theory as simple as possible: with as little theoretical terms as possible and as strong restrictions over them as possible. If we succeed in achieving a satisfactory fit, we have assigned meaning to the theoretical terms and learned something about the agent. This structure, I believe, is structurally similar to any measurement procedure: A theoretical structure is developed, and then applied to a set of phenomena. If the structure yields consistent results, then it is assigned a meaning (in relation to the phenomena) and it yields information about those phenomena.

A critic might respond that under the conditions specified here, the ascription of mental properties is epistemologically indeterminate. There might well be more than one measurement procedure that satisfies the constrains, leaving us with different mental property ascriptions for the same behavioural evidence. That is correct; but I do not see it as a problem. The mental properties are real insofar as they account for an objectively present pattern of behaviour (compare Dennett 1991). This behaviour is ultimately determined by physiological or physical causes, as I argued in chapter 1. Beyond the account for this pattern, I am happy to concede that the mental properties only have an instrumental quality – but a very useful one.

Grüne tries to solve problems of individuation and preference change by formal ~~definit~~ axioms in theory.
Alternative: auxiliary assumptions — theory of rational choice is untestable as it stands, but has content in standard applications.

Chapter 3

Preference Explanation on the Basis of Causal Structure

3.1 Introduction

In this chapter I discuss the minimal structure required for preferences to function in the explanation and prediction of behaviour.

As argued in the previous chapters, the explanation and prediction of behaviour proceeds in three steps. First, mental properties, understood as behavioural dispositions, are ascribed to agents on the basis of observations of their behaviour. Secondly, a hypothetical mechanism assigns causal efficacy to some of these dispositions contingent on the presence of other dispositions and environmental conditions. The effects of the efficacious dispositions are aggregated to yield the prediction or explanation (retrodiction) of behaviour.

Preferences are a particular kind of these mental properties ascribed and employed in the course of explaining and predicting behaviour. They dispose the agent who holds them to behave in a particular way, given the right constellation of other mental states (chiefly beliefs and other preferences) and environmental conditions (availability of the options, feasibility of the behaviour). Because preferences are essentially assigned to explain, the conditions for the possibility of explanation of behaviour function as restrictions on the concept of preference. These restrictions are two: First, the ascription of a preference must be *empirically justified.* Second,

preferences must be sufficiently *abstract* to be employed in the explanation of situations that are different from those in which they were ascribed.

To justify preference ascriptions empirically, one has to refer to the observed choices of the agent in question. This empirical basis includes linguistic behaviour, but it does not treat verbal reports as privileged introspective access. From choices, however, one can only infer preferences of the most specific states of the world. After having determined that the behaviour in question was indeed a choice between various options, it must be assumed that the option was chosen *in all its specificity* over all other options available. For example, when one observes someone wearing Wellington boots, one can ascribe a preference only by taking into account the full situation: the weather, the time of the day, his destination, the social surrounding, etc. One cannot ascribe a preference for wearing Wellington boots over sneakers, e.g., *in general*; but only a preference for wearing Wellington boots over sneakers, say, in wet weather, in the countryside, in a casual or work environment, when heading across the meadow. The preferences derived from behavioural evidence are therefore highly specific.

To be useful in explanation, ascribed preferences are applied in new combinations to different and possibly completely new situations. To take up the example from section 2.2.2, a pedestrian is mugged in the street. Even though she was never in a similar situation before, she remains calm, does not look at the mugger and hands over her wallet in a very cautious manner. We can explain her behaviour by pointing out her beliefs and preferences: she believed that muggers might harm her if she did not yield or if they felt threatened; and she preferred staying unharmed over retaining her material possessions.

We might have inferred her preference for physical integrity over material possessions from choices completely different from the given situation. Maybe we observed her spending a lot of money on health insurance in the past, or something similar. If, on the other hand, the wallet contained something different than money – say the heirloom engagement ring of her late grandmother – and we found her in the past to put family tradition over her personal health, we would have predicted that she would fight instead of handing over her wallet.

The point of this example is that the ascription of mental properties only lends itself to an operational framework of explanation or prediction of behaviour if the ascribed mental properties are sufficiently versatile. If one only ascribes preferences over maximally specific states of the world – as for example the preference to deliver the wallet to the muggers over a fight with them at Boerum Street, Brooklyn, on September 4th, 9.30 pm – then preferences cannot account for anything but that particular event.

This presents a problem for the explanation and prediction of behaviour. If the outcomes over which preferences are defined are too specific, then our explanatory theory is empty. But as argued above, only the most specific preferences can be derived from an agent's observable behaviour. Thus with preferences directly derived from observed choices one cannot explain any situation that is not an exact repetition of those choices. Additionally, such an exercise wouldn't really be an explanation. So how can we ascribe preferences in an empirically justified manner that are sufficiently versatile for explanatory and predictive purposes?

What is needed is a way to construct abstract preferences on the basis of specific preferences, such that the empirical justification of the specific preferences is preserved in the derived abstract ones. The degree of abstraction of a preference is determined by its relata, which can be specified within a wide range. They can be the most specific states of the world (from here on simply: 'worlds'); or they can, in the other extreme, be highly abstract properties. Over these two extreme types of outcomes, and all levels in between, preference orderings can be defined. But all that is given so far are those preferences over worlds which can be derived from observed choice behaviour. This paper develops a *principle of equivalence* that connects the world-preferences with the more abstract ones.

Two different conceptual approaches are offered here. Either, sufficient distinctions are made to achieve good explanatory and predictive results, 'without making so many distinctions that no choice bears on any other' (Pettit 1991, 211). Or, we individuate outcomes to the finest level and then look for a principle to tell us which differences ought not rationally to matter (compare Broome 1991, 95-115).

Pettit argues that world preferences are not basic in decision

making, and that instead we make decisions from values (abstract preferences) to ends (preferences over options we choose).[1] From this argument about where we deliberate from, he argues against the methodological construction of abstract preferences from more specific ones.

> For a disposition to choose to count as a preference, it must be a disposition to choose with a reason – a disposition to choose on the basis of the properties displayed by the alternatives. . . . The equation of preferences with such brute [mere behavioural] dispositions is bound to seem inappropriate under the assumption of desiderative structure. And rightly so. After all, even if a person is disposed to choose one unconsidered prospect rather than another, he will be equally disposed, if possible, to consider the properties *before* making his choice. (Pettit 1991, 209, my italics)

This may be an argument against a kind of Skinnerian Behaviourism, but not against the *methodological* identification of prospect preferences from behavioural data. What needs to be distinguished is a metaphysical from a methodological meaning of 'basic'. While the atomism-holism debate remains undecided, it is methodologically non-controversial that the only empirical justification can be obtained from specific preferences. The principle of equivalence presented here therefore does not take a stance on the former debate, but is only constructed to clarify the role of preferences in the explanation of behaviour.

In this paper, I will start with preferences over outcomes individuated to the maximal degree and then provide a principle that allows to distill preferences of considerably higher abstraction from this world-preference ordering. I will provide a principle that is based on a model of causal beliefs, and that I think is acceptable

[1] Pettit's claim is that property preferences *determine* world-preferences. The ultimate determining preference is often called value (as does Pettit himself). Disagreement prevails between those who defend value atomism – that value has its origin in a few very abstract aspects of the world (compare Harman 1967; Quinn 1974; Carlson 1997) – and those who defend value holism – that value has its origin in the most specific states of the world (compare von Wright 1963, 29-34 and 1972; Rescher 1967; Trapp 1985; Hansson 1989 and 2001). Pettit claims that it is a folk psychological platitude that 'choosing on the basis of the properties displayed by the alternatives' captures 'choosing for a reason'. As there is considerable disagreement amongst philosophers about this claim, I am cautious granting it folk status.

on the basis of very weak plausibility considerations. The final assessment of the model presented here, however, is not philosophical, but empirical. How exactly the partitions are made is a question of predictive success, not of rational appraisal. In this sense, this paper can only conjecture a structure for empirical investigations, whose utility needs to be proven in its application.

3.2 Prospect Preferences

To formally present this principle, I will make a number of assumptions. First, I assume that there exists a level of maximally specific states of the worlds, denoted $w_1, \ldots, w_n$.[2]

Second, a weak preference pre-order (i.e. a binary relation over worlds that is reflexive and transitive) is defined over these worlds, based on the agent's choices. For simplicity reasons, it will be assumed that all choices are made over *certain* outcomes.[3] Choices are made over certain, most specific outcomes – over *worlds*. Preferences over worlds are derived from these choices as follows: An agent (weakly) prefers w_i to w_j ($w_i \geq w_j$) if she chooses w_i over the available w_j. She is said to be indifferent between w_i and w_j ($w_i \approx w_j$) iff both $w_i \geq w_j$ and $w_j \geq w_i$. She is said to (strictly) prefer w_i to w_j iff $w_i \geq w_j$ and not $w_i \approx w_j$.[4] Obviously, it would be extremely unrealistic to assume preferences over worlds to be complete on the basis of such a definition.[5]

[2]The specification of which can be dependent on an array of parameters. I will discuss the problem of 'small worlds' and choice of partitions in section 3.4.2.

[3]This in effect assumes that all choices are, in Savage's terminology, *constant acts* (Compare Savage 1972, 25). I make no attempt to justify this assumption, as I do not think that it can be empirically justified. However, as I will operate in a deterministic causal framework in the rest of this paper, I wanted to exclude all considerations of uncertainty for the sake of simplicity. To elaborate in a probabilistic framework what is discussed here deterministically will be the task of another paper.

[4]This account must not be identified with *revealed preference theory* known from neoclassical economics. Revealed preference theory *defines* preferences as consistent choices over options under a given budget. The revealed preference relation xVy is defined as the result x of the choice function h selecting from the set of those options y affordable under a given price p and a given endowment m.

$$xVy \Leftrightarrow \exists(p,m)_{(p,m)\in\mathfrak{C}} : x \in h(p,m) \,\&\, p*y \leq p*x \text{ (Richter 1966, 637-638)}$$

Instead of defining a preference, the behavioural evidence in the account presented here only *indicates* a preference. The most obvious difference to revealed preference theory is that the account here only employs a '$\Leftarrow$' instead of a '$\Leftrightarrow$'. Besides choice evidence, data from hypothetical choices (i.e. the agent is confronted with counterfactual scenarios and reports which option she would 'choose') and verbal accounts are also taken into consideration.

[5]The determination of preferences through choices as propagated here leaves open the

Third, I assume that worlds are fully analysable into conjunctions of *certain* prospects. A prospect can be the particular realization of a property, or a conjunction thereof, or the fact that a property is realised at all. Trivially, worlds are prospects as well. A further restrictive assumption I make is that of determinism. Ultimately, there is no uncertainty in any world, hence every world is fully analysable into certain prospects. Prospects are denoted p, q, r and for simplicity, I take worlds to be sets of the prospects into which they are analysable: for example $p \in w_i$.

Last, I assume a deterministic causal relation to be defined over certain prospects. This relation is irreflexive, asymmetric and acyclical, but again is not assumed to be complete. I will not discuss any empirical basis for this relation. It is interpreted as the beliefs an agent holds about the causal dependence of particular prospects.

The principle of equivalence I propose comes in the guise of a definition of the preference relation $\succeq$ over prospects $p, q, \ldots$ in terms of the preference relation $\geq$ over worlds $w_1, w_2, \ldots$. It employs a *representation function* f that picks out pairs of worlds $\langle w_i^p, w_i^q \rangle$ for each pair of propositions $\langle p, q \rangle$:

Definition 1 $p \succeq q \Leftrightarrow w_i^p \geq w_i^q$ *for all* $\langle w_i^p, w_i^q \rangle \in f(\langle p, q \rangle)$.

Definition 1 is trivial if the propositions p and q are worlds themselves. Otherwise, the definition is not trivial. It now requires a specification of f such that all the relevant worlds are picked out in such a way that through the preference defined between them they determine the preference between the two prospects.[6] I will discuss the form of f in two separate installments. In the first step, I will focus on the special case where prospect preferences are only defined over a prospect p and its negation $\neg p$. In such a preference, *mutu-*

possibility that inconsistent preferences are ascribed on the basis of behavioural observation. This problem will be discussed in section 4.2.1.

[6] It now becomes clear why the paper is restricted to certain prospects. This definition does not work if p or q are gambles over worlds. Take the following example: p and q are gambles such that

p: if dice rolls 6 you receive \$100.
q: if dice rolls 4 or 5 you receive \$100.

According to the definition, one prefers q to p only if one both prefers world w_q^1: 'dice rolls 4 and you receive \$100' to w_p as well as world w_q^2: 'dice rolls 5 and you receive \$100' to w_p. But of course it is natural to be indifferent between these worlds, even though it is very plausible to prefer q to p. I thank both Richard Bradley and an anonymous referee, who independently brought this to my attention.

ally exclusive and *conjointly exhaustive* prospects are compared.[7] In the second step, I will discuss prospect preferences defined between mutually exclusive, but conjointly *not* exhaustive prospects.[8] This distinction is important, because the latter feature in preference orderings beyond the pairwise level, while the former don't. Thus preferences over mutually exclusive, but conjointly not exhaustive prospects are subject to the transitivity property, and I will present an interesting result here.

3.2.1 Conjointly Exhaustive Prospects

In this subsection I will restrict myself to cases where definition 1 defines preferences over prospects and their negations only; preferences of the sort $p \succeq \neg p$. The way f picks out worlds is of central importance for the preference relation between prospects; definition 1 says nothing about it. There are at least three different doctrines about how to specify f.

The *absolute* preference approach stipulates that all worlds which are logically compatible with a prospect have to be taken into account. That is, any world w^p that contains a prospect p has to be preferred to any other world $w^{\neg p}$ that does not contain the prospect p. This very quickly leads to enormous numbers of world-comparisons necessary for the derivation of a prospect preference. For example, imagine worlds instantiated by only four prospects, p, q, r, s. Then there would be $2^3 = 8$ different worlds that contain p, and 8 that do not. In the absolute preference approach, all possible $8^2 = 64$ comparisons between p-worlds and $\neg p$-worlds have to show a preference for p worlds, in order to derive the prospect preference $p \succeq \neg p$ from it.

In such a universe, let p be the agent's consumption of Marmite, q and r prospects irrelevant at the moment, and s the case that the agent is allergic to Marmite. Now, whether q and r are realised or not, as long as s isn't, the agent prefers the world in which she consumes Marmite to the one where she doesn't. But, quite understandably, she does prefer the world where she is allergic to the stuff and does not consume it to worlds where she does consume

[7] This is the case that comes closest to Pettit's discussion of *property desires*.

[8] I will argue that the third case, preferences over mutually non-exclusive prospects, must be translated into preferences over mutually exclusive ones. A translation procedure will be discussed in section 4

it and suffers the allergic consequences of her actions. Should her preference between those last two worlds determine her prospect preference over Marmite consumption? I don't think so. The scenario is *counterfactual*; she does not actually suffer from the allergy. This doesn't mean that counterfactual scenarios do not have any influence on prospect preferences; I will show further down that they do. But in this case, the counterfactual scenario is *causally independent* of the prospect in question; Marmite consumption does not cause Marmite allergy. The absolute account does not allow this abstraction and thus should be discarded.

The *ceteris-paribus* preference approach stipulates that only those worlds are taken into account which are as similar as possible to each other, while realizing and not realizing the prospect in question respectively. That is, any world w^p that contains a prospect p has to be preferred to that other world $w^{\neg p}$ which is as similar to w^p in as many aspects as possible.[9]

For illustration, let's imagine that the four aspects of our four-aspect worlds are logically independent. Then, clearly, there is exactly one w^p-world that is most similar to one $w^{\neg p}$-world: namely that world that shares with w^p the realization or non-realization of all aspects but p. According to the *ceteris-paribus* approach, then, there are only eight comparisons between the four-aspect-worlds necessary to establish prospect preferences. This can be illustrated in figure 3.1, where the numerals in the columns signify the realization or non-realization of an aspect in the respective world.

w_i^p	p	q	r	s		$w_i^{\neg p}$	p	q	r	s
(1)	1	0	0	0	$\geq$	(1)	0	0	0	0
(2)	1	1	0	0	$\geq$	(2)	0	1	0	0
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
(8)	1	1	1	1	$\geq$	(8)	0	1	1	1

Figure 3.1: *Ceteris paribus* comparisons

Figure 3.1 shows the sufficient conditions for $p \succeq \neg p$ according to the *ceteris-paribus* approach. Each world in which p is realised is compared with the world in which p is not realised, but which is

[9]This approach was to my knowledge first discussed by von Wright 1972, 146. It is also defended in Hansson 2001, 67-94.

otherwise as similar as logically possible. If all aspects are logically independent – that is, no aspect is implied by any other aspect nor implies any other aspect – then the two worlds compared differ *only* in the realization of p. We are free to choose how to partition the worlds into aspects, and it is not conducive for the purpose of deriving more abstract prospect preferences to partition the aspects such that they are logically dependent. Thus the situation will always look like that illustrated in figure 3.1.

There are two fundamental problems with the *ceteris paribus* account. First, it rests on a concept of logical possibility, which is too wide for the purpose at hand. Second, it disregards the world the agent is in when making the comparison. The following example will illustrate both of these shortcomings in turn.

Diogenes Laertius, the ancient chatterbox, tells of an incident where Alexander the Great puts Diogenes of Sinope to the touch. 'Ask of me any boon you like' the Macedonian is reported to have offered; to which the reply came: 'stand out of my light'.[10] The anecdote is quite popular, and rightly so. At first sight, Diogenes seems to act contrary to a knee-jerk reaction of most of us. You are offered wealth or power for free – then take it! In this version of the story, Alexander embodies the ancient idea of Kairos, Machiavelli's Fortuna or, if you will, one of the brothers Grimm's good fairies. When Diogenes declines the seemingly irresistible offer, he must have good reasons for it.

As revealed in his choice, Diogenes prefers a world w^u undisturbed by any patron, however powerful, to a world w^o which promises all the wealth and influence Alexander has to offer. If we now think that the two worlds differed in only one relevant aspect, wealth, we could derive Diogenes preference for poverty over wealth. But even though we don't know much about them, we can suspect that Diogenes' other choices could not have been subsumed under such a simple prospect preference. Even that most hardened despiser of material wealth, we suspect, must see that wealth and power are desirable for him too: he wouldn't have to go panhandling anymore, he could have bought his freedom from Xeniades, his owner, or he could have convinced the elders of Sinope to re-

[10]Diogenes Laertius VI, 38. I use the source as an inspiration, and hasten to add that the following is not meant as a textual analysis.

move the ban and let him return to his homeland. So if Alexander – who's immediate reaction is not reported – had asked, in slight astonishment: 'but don't you want to be rich?', Diogenes answer, if for once straightforward, would have been complex. 'On the one hand', he would have retorted, 'there is a sense in which I want to be rich. But on the other hand look at the world I live in: if I took a significant boon from you, I would be obliged to show my gratitude. Further, my lifestyle would be considered implausible; and people would envy me for my easily achieved wealth. Under these conditions, I do not want to be rich.'

With this extra bit of information, we may try to apply the *ceteris-paribus* framework for an analysis of Diogenes' preferences. According to the account that I put into his mouth, Diogenes identifies four aspects of w^u and w^o to be relevant: wealth (r), independence from donors (i), personal credibility (c) and envy of others (e). But clearly, all these aspects are *logically* independent. Thus the specification of f in figure 3.1 applies. According to it, Diogenes compares $w_1^u = \{\neg r, \neg i, \neg c, \neg e\}$ with $w_1^o = \{r, \neg i, \neg c, \neg e\}$, $w_2^u = \{\neg r, i, \neg c, \neg e\}$ with $w_2^o = \{r, i, \neg c, \neg e\}$, etc. Whatever his preferences between those worlds are, and whatever the resulting prospect preferences are, this specification of f does not capture his story at all if it goes: 'on the one hand, I want to be rich. But on the other hand look at the world I live in. . . '. There, he compares $w_i^u = \{\neg r, i, c, \neg e\}$ with $w_i^o = \{r, \neg i, \neg c, e\}$. According to the *ceteris-paribus* approach and the assumed logical independence of the aspects, such a comparisons is not admissible, because the worlds are too far apart. So does Diogenes tell us an incoherent story, or is the *ceteris-paribus* approach wrong?

I propose that it is the *ceteris-paribus* approach that is flawed. Diogenes does not employ logical but *causal* possibility when assessing the independence of the worlds' aspects. He envisages a particular way in which he can achieve wealth: through his submission under a donor. As he tells us, he believes in the causal dependence of the other relevant aspects on this genesis of wealth. His wealth would cause the envy of others; his submission under a donor would cause the loss of his independence, which in turn would cause the loss of his credibility. Given the causal dependence Diogenes believes in, worlds which are most similar to w^u but for the realization

of wealth are not the ones the *ceteris-paribus* account suggests. It is *causally impossible* for Diogenes to be wealthy *without* being envied; it is equally impossible for him to be wealthy through the benefits of a donor *without* becoming dependent on him and hence losing his credibility. Even though these worlds are logically possible, I have argued that what matters for a principle of equivalence is *causal* possibility. Logical possibility only forbids what is inconsistent, while causal possibility allows only *what can be produced.* The agent takes only those worlds as possible given p which are producible according to her *causal beliefs.* This epistemic notion of causality will restrict the selection function in the following way:

Restriction 1 *f picks out only those worlds which are causally compatible with p and $\neg p$, respectively.*

But this restriction alone is not sufficient for the right choice of f. The causal structure an agent believes in restricts the worlds she will deem possible; but she will not compare all possible worlds, as some of them are too far removed from her actual situation. Thus, *facts* believed to be actual play a role too.

To stay with the above example, Diogenes might reasonably believe that secretly inheriting from a distant relative causes one to be wealthy without any strings attached. Thus, such a causal story would allow him to introduce into definition 1 the world w^o where he is wealthy, independent, credible and not envied by anybody due to the secrecy of the inheritance. So it might seem that because of the possibility that this belief opens, Diogenes does not prefer poverty over wealth *simpliciter.* It seems he only prefers it conditional on other aspects, in this case the absence of any living bequeather.

This appearance is wrong. Diogenes does not have any wealthy relatives from whom to inherit (or at least, we, as the interpreters of his behaviour, do not know of any). To define his prospect preferences, we not only take into account the causally possible worlds that realise the relevant prospects; we only take into account the *actual* causal possibilities, which can be realised, conditional upon the world the agent is in.

The above preference expression should therefore be interpreted as taking the relevant causal background conditions to be the same as in the actual world. Of course, not all background conditions can be the same: otherwise no counterfactual world could be constructed

adhering to the causal structure. For Diogenes to imagine a world in which he is wealthy – seen from his actual predicament of poverty – a counter*factual* change is necessary. But changing the facts does not mean changing the causal dependency structure. The change of facts, under stable causal dependencies, will require certain causally prior prospects to change as well: somehow, his wealth has to be caused in this possible world. But there are facts in the actual world which offer themselves as ready causes: there *are* donors offering their support, but there *aren't* any wealthy distant relatives ready to bequeath Diogenes. Those facts that do not have to be changed in order to accommodate the counterfactual – either because there is no causal link to them at all or because there are other causes closer to the actual situation – remain as they are in the actual world. Hence,

Restriction 2 *f picks out only those worlds which realise p and q but maximally comply with those background conditions pertaining to the actual world.*

Under the two restrictions on f for which I argued here, we can indeed say that Diogenes preferred poverty to wealth *simpliciter*. What definition 1 in combination with the now specified f does, is to identify the necessary preferences over worlds in order to determine a prospect preference. This prospect preference – so far discussed only in the context of mutually exclusive and conjointly exhaustive prospects – then represents a highly abstracted disposition to choose, given that an option promises the realization of that prospect. In this sense, definition 1 is a principle of equivalence.

3.2.2 Conjointly Non-Exhaustive Prospects

Prospect preferences are not only used in the sense that one prefers the realization over the non-realization of a prospect, as Diogenes prefers poverty over wealth, according to the scheme $p \succeq \neg p$. Preferences also occur in contexts where the two relata do not exhaust the possibilities. For example, over breakfast I prefer reading an English paper to a German one; and I prefer a German to a Russian newspaper. These three types of newspapers certainly do not exhaust the possibilities of breakfast reading, nor do they exhaust my ordering of breakfast readings. However, it is perfectly intelligible to hold preferences between conjointly non-exhaustive prospects; the

problem is only that such preferences cannot be represented as cases of the scheme $p \succeq \neg p$.

Conjointly non-exhaustive relata occurring in preference types $p \succeq q$ are not necessarily mutually exclusive. For example, one can meaningfully hold a preference like 'I prefer an apartment in New York to a house in Tuscany', even though it is clearly possible to own an apartment in New York and a house in Tuscany at the same time.[11] However, to express a preference $p \succeq q$ without meaning to express a preference for $p \wedge \neg q$ over $q \wedge \neg p$ violates Grice's *Cooperative Principle* (Grice 1989). In particular, if uttered in a situation of choice between either p or q, the conversational contribution made does not satisfy the pragmatic convention of *relevance*: preferences over relata involving $p \wedge q$ do not help making such a choice. If uttered in a situation where information about the speaker's evaluations is sought, it does not satisfy the pragmatic convention of *informativeness*: $p \wedge q \succeq q \wedge p$ is tautological and thus empirically empty. By conversational implicature, then, a preference between mutually non-exclusive relata is interpreted as a preference between the corresponding mutually exclusive relata.[12] This conventional translation procedure has to be amended for cases where at least one relatum logically implies the other or causally requires the presence of the other. Thus $p \succeq q$ is translated to $p \wedge \neg q \succeq q \wedge \neg p$ only if it is *possible* that $p \wedge \neg q$ and $q \wedge \neg p$. In cases where it is not, the original relatum remains untranslated (compare Hansson 2001, 68-70). Thus restriction 1 needs to be reformulated for conjointly non-exhaustive prospects in the following way.

Restriction 3 *f picks out only those worlds which are causally compatible with $p \wedge \neg q$ and $q \wedge \neg p$, respectively.*[13]

Concerning the actual causal background, the same restriction holds as for the conjointly exhaustive case. An example is given by Trapp (1985) for preferences over different diseases (of which the comparison between two of them is obviously not conjointly exhaustive). A man who prefers contracting cholera to being ill with cancer should

[11] Trapp claimed that 'no two relata of a preference relation should be considered to be true in the same possible world', at least in those worlds that are chosen by the selection function (Trapp 1985, 301). For a rejection of this view, see Hansson 1989, 6.

[12] Such an translation, albeit without the conversational implicature justification, was first presented by Halldén (1957, 28).

[13] Or, if one of the relata is not causally compatible with any world, f picks out worlds which are compatible with the untranslated relatum.

not be interpreted as preferring a situation where there is no cure for cholera (say for example he would be in a country where there are no antibiotics available). The existence of a cure has significance consequences if one has either cholera or cancer, and hence naturally plays a crucial role in the evaluation of both situations. Thus the agent prefers cholera to cancer iff he prefers a world where he has cholera and all the contemporary cures are available to a world where he has cancer and all the contemporary cures are available. The restriction is thus reformulated as follows.

Restriction 4 *f picks out only those worlds which realise $p \wedge \neg q$ and $q \wedge \neg p$ but maximally comply with those background conditions pertaining to the actual world.*

A particularly interesting feature of preferences over conjointly non-exhaustive prospects is that the pairwise comparisons *can* give rise to a preference ordering — but they don't have to. Under particular conditions, the preference pairs $p \succeq q$ and $q \succeq r$ imply the additional preference pair $p \succeq r$. This transitivity property of preferences need not be fulfilled by prospect preferences, even though it is (by assumption) satisfied by the preferences over worlds underlying it. All that needs to be established is that the world $w^{p \wedge \neg q}$ – compared in $p \succeq q$ with the world $w^{q \wedge \neg p}$ – and the world $w^{r \wedge \neg q}$ – compared in $q \succeq r$ with the world $w^{q \wedge \neg r}$ – are not the same as the worlds $w^{p \wedge \neg r}$ and $w^{r \wedge \neg p}$ compared in $p \succeq r$. Thus, if $w^{p \wedge \neg q} \geq w^{q \wedge \neg p}$ and $w^{q \wedge \neg r} \geq w^{r \wedge \neg q}$ but $w^{p \wedge \neg q} \neq w^{p \wedge \neg r}$ and $w^{r \wedge \neg q} \neq w^{r \wedge \neg p}$, it does not follow that $w^{p \wedge \neg q} \geq w^{r \wedge \neg q}$; and hence it does not necessarily follow from $p \succeq q$ and $q \succeq r$ that $p \succeq r$.

3.2.3 Actions

An action is a particular kind of prospect. It has been attempted to characterise it metaphysically as a species of events 'that can be described under an aspect that makes it intentional' (Davidson 1971, 46). What the correct account of agency is will not be discussed here; it suffices for the present purpose that there is *some* criterion of distinction.

The relevant point to note in this context is that agents evaluate their own actions, and sometimes those of others, in a different way from evaluating other prospects. While the evaluation of a prospect

takes into account all causal antecedents of that prospect, the evaluation of an action only takes into account the action itself and all its consequences, while disregarding any causal history that led to the action.

Take Diogenes' example again. The only way for him to achieve wealth would have been to submit under a donor, which in turn would have had consequences for his independence and credibility. All in all, he preferred a world without those consequences to a world with them; thus he preferred poverty to wealth. But if he took those indirect consequences of wealth and poverty into account, shouldn't he cast the net even wider? Let us imagine that Diogenes' choice to reject a donor would be caused by his contempt for authority. If Diogenes had this character trait, he would have so mortified and frustrated his father as an adolescent that (unbeknownst to him) his begetter subsequently would have turned into an unbearable tyrant, spelling doom over Diogenes' mother and siblings after Diogenes had left the Ionic coast.

Now, if these causal dependencies did indeed exist, would it make Diogenes change his preferences between being rich and being poor? Some claim it would:

> . . . to the extent that acts can realistically be identified with propositions, the present notion of preference is active as well as passive: it relates to acts as well as to news items. . . . From this viewpoint, the notion of preference is neutral, regarding the active passive distinction. If the agent is deliberating about performing act A or act B, and if AB is impossible, there is no effective difference between asking whether he prefers A to B as a news item or as an act, for he makes the news. (Jeffrey 1983, 84)

On Jeffrey's account, Diogenes takes his rejection of a donor as the news for his character trait and its consequences just as he takes the observation that he is wealthy as the news that he accepted the donor's offer. Presumably, what Jeffrey means by 'he makes the news' is that there is no further causal history to an action that carries news characteristics. But the above example shows that this assumption is not generally true. If Diogenes took his action as a news item, then his choice to reject the donor would tell him about his contempt for authority, his father's frustration and the plight of

his family. If he then found that world worse than a world where he himself became wealthy, dependent and incredible, he would indeed prefer being wealthy over being poor.[14]

I think this model of evaluation is flawed. Diogenes – nor any other responsible actor – takes into account the causes of their actions and the effects of these causes when evaluating their actions. An agent who evaluates a non-action state of the world takes a passive outlook: he takes into account what consequences this state has, and how this state came about, with the other consequences which that cause witnessed. An agent who performs an action exhibits an active outlook: she chooses between various options according to the benefit of their consequences; but she takes the world as it is, disregarding any influences that might have caused her action.

> Statements that describe acts are different in kind from other sorts of propositions simply because the actor has the power to make them true. With this power comes a kind of responsibility. An agent must, if rational, do what she can to *change* things for the better. . . . rational decision makers should choose actions on the basis of their efficacy in bringing about desirable results rather than their auspiciousness as harbingers of these results. Efficacy and auspiciousness often go together, of course, since most actions get to be good or bad news only by causally promoting good or bad things. In cases where causing and indicating come apart, however, the causal decision theorist maintains that it is the causal properties of the act, rather than its pure evidential features, that should serve as the guide to rational conduct. (Joyce 1999, 150)

Thus Diogenes would disregard the causes of his choices and their respective effects when evaluating the prospects of wealth and poverty, respectively. Acts must be considered *exogenous*. Instead, he would fill in those parameters with what he actually believes happened, irrespective of what option he chooses. The principle of equivalence is therefore amended for the case of actions.

Restriction 5 *If p is an action, f picks out all those worlds that*

[14] This situation is in many ways similar to the so called *Newcomb's Problems* in probabilistic models of decision making.

are causally compatible with $p \wedge \neg q$ and its consequences only, while disregarding any causal history of $p \wedge \neg q$.

The disagreement between the two positions sketched remains, however, in so far as prospects often cannot clearly be identified as actions or non-actions. Thus the allies of Jeffrey might be right in insisting that some apparent actions are evaluated as news items. This does not touch on the basis of the argument, and is of no further relevance here. With these amendments added to the specification of f, definition 1 is a principle of equivalence for all prospect preferences.

3.3 Constructing the Selection Function

The concepts of causal compatibility, maximality of compliance with the actual world and causal history so far have been given only intuitive meaning. This section seeks to specify their meaning more formally, by reference to a formal concept of causal models.

A causal model is defined by Pearl (2000, 203) as a triple

$$M = \langle U, V, G \rangle$$

where:

1. U is a set of *background variables*, determined by factors outside the model.
2. V is a set of *endogenous* variables, determined by variables of the model – that is, variables in $U \cup V$.
3. G is a set of functions $\{g_1, g_2, \ldots, g_n\}$ such that each g_i is a mapping from $U \cup (V \setminus V_i)$ to V_i and such that the entire set G forms a mapping from U to V. In other words, each g_i tells us the value of V_i given the values of all other variables in $U \cup V$, and the entire set G has a unique solution $V(u)$. Symbolically, the set of equations G can be represented by writing

   $$V_i = g_i(V_j, U_i), \quad i = 1, \ldots, n$$

   $U_i \subseteq U$ stands for the unique minimal set of variables in U sufficient to determine V_i on the basis of G.

The variables in Pearl's model are random variables. I take the individual realization of a random variable to be equivalent to a prospect, e.g. $p \equiv (V_i = v_i^1)$. Given a particular constellation of background variables $u^* = \{U_1 = u_1^1, \ldots, U_n = u_n^2\}$, the model has the unique solution $V(u^*)$. Prospects can be directly deduced from this solution: $V(u^*) \vdash p$, where $\vdash$ is the classical inference relation.

M can be represented as an acyclical directed graph, with the arrows representing the function g. Forked arrows show that g has more than one argument. Figure 3.2 is an example of such a representation of $M^* = \langle U^*, V^*, G^* \rangle$, with all variables in $U^* = \{U_1, \ldots, U_4\}$ and $V^* = \{V_1, \ldots, V_4\}$ having only two realisations each, and $G^* = \{V_1 = g_1(u_1), \ldots, v_4 = g_4(v_2, u_4)\}$. Each realization is then equivalent to a proposition or its negation. Let the first realization of a background variable be expressed by $a, b, c...$, i.e. $a \equiv (U_i = u_i^1)$, etc.; the realization of a endogenous variable by $p, q, r, \ldots$, i.e. $p \equiv (V_i = v_i^1)$, etc.; and the respective second realization by a negation of that proposition: $\neg p \equiv (V_i = v_i^2)$, etc.

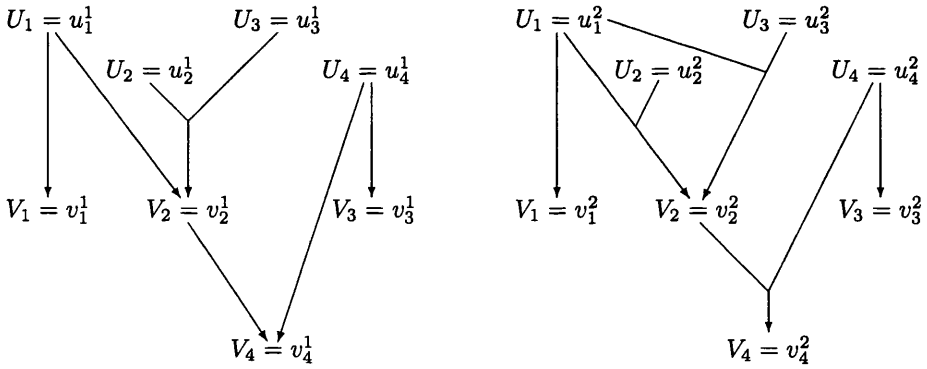

Figure 3.2: An example of a causal graph

Each world w specifies the values for all U_i and for all V_i of every M. Because the functional relationships g_i of M restricts the endogenous V's given the exogenous U's, not all worlds are consistent with a specific causal model.

Definition 2 *A world w is* consistent *with a causal model M =*

$\langle U, V, G \rangle$ *iff there is a set of realisations* $u^* = \{U_1 = u_1^1, \ldots, U_i = U_i^1\}$ *for which* $w \vdash u^*$ *and* $w \vdash V(u^*)$.

For example, the world $w_1 = \{a, \neg b, c, \neg d, p, q, r, \neg s\}$ is consistent with the model M represented in figure 3.2, while the world $w_2 = \{\neg a, b, \neg c, d, p, q, r, s\}$ is not. Having specified the relations between prospects and worlds on the one hand and the causal model and its variables on the other, we can now define causal compatibility:

Definition 3 *w is* causally compatible *with p with respect to M iff there is a causal model $M = \langle U, V, G \rangle$ such that w is consistent with M, and $w \vdash p$.*

For example, the world $w_1 = \{a, \neg b, c, \neg d, p, q, r, \neg s\}$ is compatible with p with respect to M. Worlds which are causally compatible with p represent the possible causal histories of p. In such a world there is at least one 'chain' that leads from background conditions to p in the following way.

Definition 4 *A prospect p is* dependent *with respect to the background conditions in $U^* \subseteq U$ iff there is a functional chain: $V_1 = g_1(u^*), V_2 = g_2(u^*), \ldots, V_n = g_n(V_1, \ldots, V_{n-1}, u^*)$ with $g_1, \ldots, g_n \in G$ and p being equivalent to $V_n = g_n(v_1, \ldots, v_{n-1}, u^*)$.*

According to M represented in figure 3.2, for example, q is dependent on a and (b, c), while r is dependent on a, (b, c) and d.

Now if a prospect p is not realised in the actual world $w^@$, all it takes for p to be realised is that one background condition on which p is dependent is realised. Of course p is realised as well in worlds where more than one background condition on which p is dependent is realised, but in those cases the ensuing worlds are not as *similar* as possible to the actual world.

Definition 5 *A world w^* is maximally similar to the actual world $w^@$ iff for w^* out of the set of all worlds: $max(\sharp(w^* \bigcap w^@))$.*

$\sharp$ here signifies the cardinality of the intersection of the respective world with the actual world. By maximizing the cardinality of this set, those worlds are chosen which have the highest overlap with the actual world.

Restrictions 1 and 2 (or 3 and 4, respectively) are satisfied if f selects worlds w^p and w^q, which are compatible with p and q

with respect to M, respectively, such that both w^p and w^q are most similar to $w^@$ by the above similarity measure. With the concepts discussed in this section, we can therefore specify definition 1:

Definition 1* $p \succeq q \Leftrightarrow w_i^p \geq w_i^q$ *for all* $\langle w_i^p, w_i^q \rangle$ *which are compatible with* $p \wedge \neg q$ *and* $q \wedge \neg p$ *with respect to* M*, respectively, such that both* w_i^p *and* w_i^q *are most similar to* $w^@$.

Definition 1* yields a preference relation $\succeq$ over propositions with the following properties.

Theorem 1 *If the causal model is non-cyclical,* $\succeq$ *is reflexive.*

Proof For each world w_i compatible with p, there is a realization of the background variables u_i such that the proposition equivalent to $V(u_i) \cup u_i$ contains w_i. u can be distinguished into the independent and the dependent background conditions, u^*. If there is only one set u^* for p, the proof is trivial, because there is only one world that is compatible with p. If there is more than one u^*, then the similarity relation ensures that only identical u_i^*'s are paired. Hence,

$$for\ all \langle w_i, w_j \rangle \in f(\langle p, p \rangle) : w_i = w_j.$$

Given that $\geq$ is reflexive, the relation $\succeq$ defined thus is equally reflexive.□

Theorem 2 *If for all prospects* $p, q, r \ldots$*, all causally possible conjunctions* $p \wedge \neg q, p \wedge \neg r$ *are dependent on the same background variable* u^{p*} *(and similarly for* $q \wedge \neg p, r \wedge \neg q, \ldots$*), then a prospect preference ordering over* $p, q, r \ldots$ *is transitive.*

Proof Without loss of generality, we take the case where $p \succeq q$ and $q \succeq r$. If all $p \wedge \neg q, p \wedge \neg r$ are causally possible, then there are causally compatible worlds $w^{p \wedge \neg q} \vdash p \wedge \neg q$ and $w^{p \wedge \neg r} \vdash p \wedge \neg r$. If for all p, q, r, $p \wedge \neg q$ and $p \wedge \neg r$ are dependent on the same variable u^*, then there is at least one world $w^p = w^{p \wedge \neg q} = w^{p \wedge \neg r}$ which is causally compatible with both $p \wedge \neg q$ and $p \wedge \neg r$. If $p \wedge \neg q$ and $p \wedge \neg r$ depend only on u^*, then $w^p = w^{p \wedge \neg q} = w^{p \wedge \neg r}$ is the world causally compatible with $p \wedge \neg q$ and $p \wedge \neg r$ which by definition 5 is most similar to $w^@$ (for the same reasons, *mutatis mutandis*, $w^q = w^{q \wedge \neg p} = w^{q \wedge \neg r}$ is the world causally compatible with $q \wedge \neg p$ and $q \wedge \neg r$ which is most similar to $w^@$). By definition 1*, and $p \succeq q$

and $q \succeq r$, $w^{p \wedge \neg q} \geq w^{q \wedge \neg p}$ and $w^{q \wedge \neg r} \geq w^{r \wedge \neg q}$. By the argument above, $w^{p \wedge \neg q} = w^{p \wedge \neg r}$ and $w^{q \wedge \neg p} = w^{q \wedge \neg r}$, hence $w^{p \wedge \neg r} \geq w^{q \wedge \neg r}$ and $w^{q \wedge \neg r} \geq w^{r \wedge \neg p}$, and thus by transitivity of $\geq$ $w^{p \wedge \neg r} \geq w^{r \wedge \neg p}$. Then by definition 1*, $p \succeq r$.$\square$[15]

It is further noteworthy that $\succeq$ is not complete, even if $\geq$ is. This can easily be seen by the following counterexample. Take a pair $\langle p, q \rangle$ such that $\langle w_1^p, w_1^q \rangle \in f(\langle p, q \rangle)$ and $\langle w_2^p, w_2^q \rangle \in f(\langle p, q \rangle)$ such that $w_1^p > w_1^q$ and $w_2^q > w_2^p$. Then $\succeq$ is not defined over $\langle p, q \rangle$.

These results are quite weak, but they represent genuine properties of pairwise preferences. The antecedent of theorem 2 is of course often not fulfilled, which explains the manifold existence of intransitive preference comparisons. That preferences are not complete over the set of all propositions, should not be surprising at all.

The formal apparatus developed in this section can now be applied to the case of Diogenes' discussed in section 3.2.1. Diogenes lives in world where he is without donor, and therefore poor and not envied, but independent and credible in his ideology: $w^{@} = \{\neg s, \neg r, \neg e, i, c\}$. The causal model M that Diogenes believes in is represented in figure 3.3.

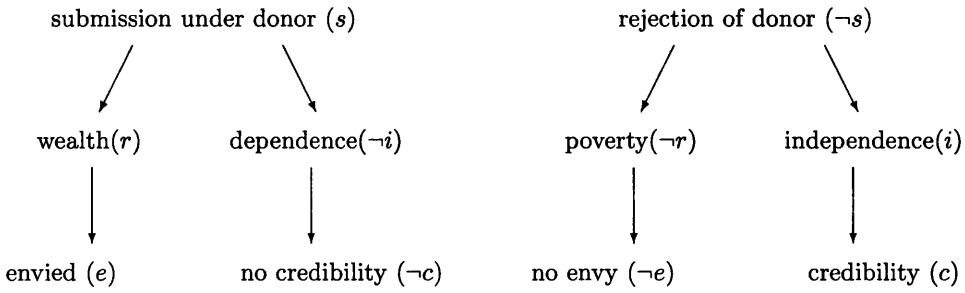

Figure 3.3: Diogenes' causal beliefs

The actual world is thus causally compatible with the prospect of poverty ($\neg r$) with respect to M, and it is obviously maximally

[15]The reverse claim does not hold: one cannot infer from the transitivity of a preference relation over $p, q, r, \ldots$ that all their causally possible conjunctions $p \wedge \neg q, p \wedge \neg r$ are dependent on the same background variable u^{p*}. For example, the evaluation of $w^{p \wedge \neg q}$ and $w^{p \wedge \neg r}$ might coincide without the two worlds being identical.

similar to itself. The world $w^o = \{s, r, e, \neg i, \neg c\}$ on the other hand, is compatible with the prospect of wealth (r) with respect to M. Because wealth is dependent on only one background variable in model M, there is no other world compatible with the prospect of wealth with respect to M. Thus even though $\sharp(w^o \bigcap w^@) = 0$, w^o is picked by f. By definition 1*, $\neg r \succeq r$ iff $w^@ \geq w^o$. Diogenes' behaviour in front of Alexander, as reported by Diogenes Laertius, does reveal his preference for $w^@$ over w^o; and hence – through his causal beliefs – his preference for poverty over wealth.

But what if the causal model gets extended to include causes of Diogenes choice between accepting and rejecting the donor? The background intuitions of such an extended model were discussed in section 3.2.3. In figure 3.4, the corresponding causal model M' is represented.

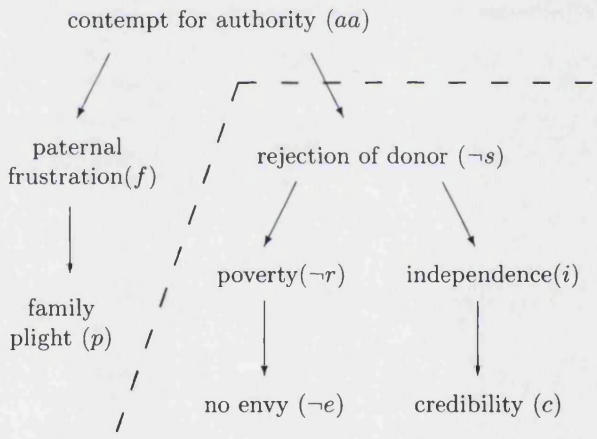

Figure 3.4: Truncating the causes of Diogenes' action

If definition 1* operated with M' instead of M, the conclusions of the above example would no longer be valid. Diogenes would prefer poverty over wealth if and only if he preferred the world $w^{o'} = \{\neg aa, \neg f, \neg p, s, r, e, \neg i, \neg c\}$ over the world $w^{aa} = \{aa, f, p, \neg s, \neg r, \neg e, i, c\}$; which is a completely different condition from preferring w^o to the actual world.

However, restriction 5 tells us to neglect all causal antecedents of

a prospect if that prospect is an action. When evaluating non-action prospects, we assumed the truth of a prospect counterfactually and investigated how the causal dependencies and effects of that counterfactual assumption would determine the worlds compatible with that prospect. When evaluating an action, we assume the truth of that action-prospect not counterfactually, but as an *intervention*. An intervention, in contrast to a counterfactual assumption, does not have a retrospective influence on the past.[16] An intervention is represented as a *truncation* of the causal graph: all direct ancestors of the model are removed from a causal model M, the model thus transformed into a truncated model M^T.

Definition 6 *A causal model M is transformed into a truncated causal model $M^T = \langle U, V, G^T \rangle$ by eliminating all $g_i \in G$ which have an action prospect in their range.*

The thick dotted line in figure 3.4 shows such a truncation. The function that connects *aa* with $\neg s$ is eliminated, thus cutting the causal connection between *aa* and $\neg s$ in M^T. By including M^T instead of M into definition 1*, restriction 5 is always satisfied.

Definition 1** $p \succeq q \Leftrightarrow w_i^p \geq w_i^q$ *for all $\langle w_i^p, w_i^q \rangle$ which are compatible with $p \wedge \neg q$ and $q \wedge \neg p$ with respect to M^T, respectively, such that both w_i^p and w_i^q are most similar to $w^@$.*

In cases where M does not include any action prospects, definition 1** is of course identical with definition 1*. In all other cases, definition 1** still satisfies theorems 1 and 2, as they were proven for *all* causal models, including truncated ones.

Thus, despite Diogenes' belief in the extended causal model M', definition 1** secures that his preference for poverty over wealth is still derived on the basis of the truncated model M^T, which in this case coincides with the original model M.

3.4 Conclusion & Remarks

I have offered a principle of equivalence between an agent's preferences over prospects and her preferences over worlds. More specifically, I represented the agent's beliefs as a causal model, and argued

[16] For a more extensive discussion of intervention, see Pearl 2000, 85-89, Spohn 2002, 23-27.

with the help of this model which of the agent's preferences over worlds serve as definiens for her preferences over propositions.

I have argued why such a principle of equivalence is necessary for the explanation and prediction of behaviour with preferences; however, the model presented here leaves open many important questions. I will finish with three remarks on how to develop the discussion further.

3.4.1 Possibility or Probability

The criterion of the causal possibility of a world might be too rough a distinction to be viable. Instead, it has been suggested that prospects should be evaluated according to a weighted average of value of those worlds in which they are realised. The weighing can be determined as a probability index which measures the likelihood of a world occurring given the actual world. Ideally, such a measure combines the criteria of causal possibility and actuality.

A first step was made in Rescher (1967). He constructed a ranking over worlds by assigning to them a numerical *index of merit.* From this ranking he derived an index over states: The index number of a state $\sharp(a)$ is the arithmetic mean over the index numbers of all possible worlds in which a is true. These index numbers over states give rise to a semantic definition of preferences over states: a is preferred to b iff $\sharp(a) > \sharp(b)$.

Trapp (1985) picked up Rescher's idea; but unlike him, Trapp suggested a probabilistic weighing of the index of possible worlds. Such a weighing can be interpreted as a continuous similarity metric: An agent assigns higher probability to those worlds that he thinks are closer to actuality. A very similar account was given by Jeffrey, who derived the desirability index over propositions from the desirability index over worlds:

> . . . the desirabilities of a proposition is a weighted average of the desirabilities of the cases [worlds] in which it is true, where the weights are proportional to the probabilities of the cases. (Jeffrey 1983, 78)

The most pressing problem of these accounts is their uniform treatment of actions and non-action prospects, as discussed in section 3.4. More generally, the probabilistic weighing of the worlds does not

necessary coincide with the concept of causal compatibility presented here. Causal decision theory has tried to remedy this problem by recasting the probability measure as a specification of objective chances or a measure of counterfactual dependency. Instead of trying to import all relevant information into the probability measure, the natural expansion of the account presented here suggests to employ a subjective probability measure *conditional* on other relevant causal factors held fixed. The notion of relevant causal factors, of course, needs to be provided independently and prior to the probability measure; a task fulfilled by the causal graph discussed in this paper. The structure needed for a probabilistic weighing of worlds to determine the preferences (expressed as a utility index) over prospects then requires a *Bayesian Network* which consists of a causal model and a probability function defined over it, satisfying certain conditional dependencies. To construct a utility function on the basis of Bayesian Networks will be the task of future work.

3.4.2 Small Worlds

What are the objects of world preferences? They are the most specific items we assign as the content of mental properties. But how is 'most specific content' defined? This question has not been answered in the account presented here. I have only assumed that every world can be partitioned into some collection of prospects.

However, this question is of central importance: if one interprets Diogenes' choice as based on a less or more fine-grained situation than I did in w^o and w^u, the prospect preferences derived from those worlds might be different from the ones that I got. This partition-dependence of the evaluation of (action-) prospects was first discussed by Savage as the *small world* problem (1972, 82-91). A solution in the form of a partition-invariant utility function was proposed first by Jeffrey for evidential decision theory and then by Joyce (1999) for causal decision theory. The problem of Jeffrey's theory were briefly mentioned in the first remark. The problem of Joyce's utility measure is that it rests on a measure of counterfactual dependence, which ultimately depends on a similarity measure between possible worlds – a very unwieldy and mysterious notion.

Instead of attempting to solve the problem of small worlds in all its generality, some decision theorists lately have argued that

one needs to justify the particular partition in which the situation is modelled.[17] Two kinds of arguments are of interest here. First, partitions have been justified as *rational*. A partition is rational if it serves an advantageous purpose, as expressed for example in Broome's 'principle of individuation by justifiers':

> Outcomes should be distinguished as different if and only if they differ in a way that makes it rational to have a preference between them. (Broome 1991, 102)

One particular way to cash out this idea is by specifying the costs involved in refining a partition and comparing them with the expected gains from such a refinement.[18]

Secondly, partitions have been justified by pointing to allegedly corresponding representations on the nervous system level of individuals. Monitoring neuron activity in the paretial cortex of animals or humans exposed to different stimuli, it is claimed, provides evidence to what extent the tested agents differentiate environmental stimuli. Interestingly enough, recent research in this field has found evidence for cost-benefit considerations influencing these neuron activities:

> Current sensory data would reflect the observer's best estimate of the current state of the salient elements of the environment. As such, it would be influenced by stored information that could improve the efficiency of sensory processing through selective attention. (Platt and Glimcher 1999, 233)

Both approaches to partitions are in the midst of current research; conclusive arguments are not available for either. Any fuller discussion of these important arguments would therefore be beyond this chapter; I restrict myself to pointing out these open questions for the framework provided here.

3.4.3 Prospect Preference Aggregation

The model presented here provides a definition of more abstract prospect preferences in terms of world preferences. Once the prospect preferences are specified for a particular agent, the question arises:

[17] Compare Lewis 1981, 11, Sobel 1994, 161.
[18] Compare Halldin 1986.

how are they employed in the prediction or explanation of the agent's behaviour in novel situations? In the simplest case, the new situation is analyzed into its aspects, and the prospect preferences of the agent may provide us with clues of what the agent will do or why she did what she did. An example of such an application was the above mentioned case of the mugging. The victim's behaviour is explained with reference to her prospect preferences, in this case her preference for physical integrity over material possessions. Given the absence of other applicable prospect preferences, she will choose an action that she believes will leave her unharmed but without her money. Such an application is easy in cases where the available prospect preferences are unanimous – that is, where all aspects of one situation are either preferred or non-comparable to the aspects of another situation. But what can we say if conflicts arise? For example, the agent prefers physical integrity to material possession and from this perspective would prefer a world in which she complied with the muggers; but she prefers her honour untouched to feeling cowardly and from this perspective would prefer a world in which she attacked the muggers.

One suggestion for such a case is to employ the framework of ordinal preference aggregation as found in the social choice literature. But instead of using it for questions of how the rational preferences of a group of individuals can be aggregated into a coherent ranking of the group, the strategy proposes 'to apply interpersonal economic theory to intrapersonal problems' (Elster 1985, 232) – i.e. the different prospect preferences are aggregated back into one world preference.

The general results of such an application are that there is no aggregation rule for prospect preferences that satisfies certain minimal constrains and results in a coherent, transitive world-preference order (Compare Steedman and Krause 1986, Rizvi 2001). It does however not preclude that in many situations, coherent world-preferences can be aggregated from prospect preferences, or that with the help of external information, the decisiveness of some prospect preferences can be justified.

Again, a further discussion of these questions goes beyond the scope of this chapter. With the current state of research in this area, however, there are *prima faciae* reasons to believe that prospect preferences can, in many cases, be aggregated to coherent world

preferences and can thus function in the prediction and explanation of action.

Chapter 4

A Model of Preference Change

4.1 Introduction

Preferences are theoretical terms that represented a pattern in an agent's behaviour. They are identified through the causal role they play – in conjunction with other mental properties – in bringing about behaviour. Methodologically speaking, as preferences are not directly observable, and introspective evidence is not reliable, preferences are assigned exclusively on the evidential basis of observed behaviour.

> It [intentional action explanation] explains what is relatively apparent – an arm-raising – by appeal to factors that are more problematical: desires and beliefs. But if we were to ask for evidence that the explanation is correct, this evidence would in the end consist of more data concerning the sort of event being explained, namely further behaviour which is explained by the postulated beliefs and desires. Adverting to beliefs and desires to explain action is therefore a way of fitting an action into a pattern made coherent by the theory. (Davidson 1975, 159)

For all that I have discussed so far, there is still a problem with this view. The preferences inferred are preferences at a particular point in time. They represent (part of) a pattern from which the agent would act at that particular moment; a pattern that restricts

and regulates the interconnections between specific preferences and other mental properties, as it is spelled out in preference and expected utility theory. But this pattern is *synchronic*; it only incorporates mental properties at one time, not at different times. As Davidson says: 'The theory merely puts restrictions on a temporal cross-section of an agent's dispositions to choose' (Davidson 1971, 235). That the theory ascribes particular preferences to an agent at present does not imply that this agent will have these preferences at some future point in time.

The static nature of preference theory spells problems both for the ascription of preferences on the basis of behavioural evidence, as well as for the application of preferences in explanations and predictions.

First, we are supposed to ascribe and confirm a particular synchronic preference configuration on behavioural evidence that is necessarily diachronic. Behavioural evidence is diachronic because agents do not exhibit many different forms of behaviour at one time. People might smoke while listening to the radio, telephone in the bathroom, dream while sleeping, etc., but none of these simultaneous behaviours are sufficient to ascribe any meaningful preference pattern. What is needed is a considerable number of observations, and these necessarily have to be made over a period of time. This is pretty much the experience in everyday intercourse with other people. When meeting a stranger, we are cautious as to what character traits to ascribe to her; the more of her behaviour we get to see, the more complete our image of her character gets. Once we feel confident in ascribing certain preferences to her, we then test our ascriptions against the further observations we make of her behaviour. Hardly ever do we ascribe anything to a stranger upon observing her behaviour once; and if we do we are prone to error and subsequent disadvantage.

So if the evidence consists of diachronic data, which of it is allowed in the construction of a synchronic preference ordering? How 'thick' is the temporal cross-section allowed to be? A compromise has to be found such that the cross-section is not 'too thick' but there are sufficient observations as evidence for the ascription of the preference ordering. A static preference theory does not give an answer to these questions.

Second, a static theory ascribing synchronic preferences at present does not justify their application at a future time. It is the nature of synchronic preferences that they are causally efficacious only at one point in time. In economics, this deficit is remedied with the additional assumption that tastes are stable over time. This assumption could be termed the simplest possible type of diachronic preferences. It establishes a diachronic preference framework, but it fails to establish a dynamic preference theory: nothing explains *why* preferences should be stable over time.[1] The remedy, therefore, is *ad hoc*: economists generally admit that tastes change *in the long run*, but insist that they do not change in the period relevant for their explanation/prediction. But the length of such an assumed stable period is completely unwarranted; it is held only on the basis to remedy the troubling defect.

For both of these reasons, it is therefore necessary to develop the concept of diachronic preferences within a dynamic framework. In this paper, I will approximate such a dynamic framework by modelling the transformations an agent has to perform on her preferences in order to maintain certain rationality requirements.

In the next section I discuss the various possibilities of how inconsistency of an observed behaviour with an ascribed preference order can be interpreted. I argue that one important interpretation is that the preferences have changed. Given this interpretation, the ascribed preferences need to be transformed to reflect the preference change. This is done with the help of a theory of preference change, whose guiding principles I will present at the end of the next section. I illustrate the different types of preference order transformation in the third section. In the fourth section I develop a model of these types of preference transformation. The first part discusses the basic construction of such a structure; the second part actually constructs the preference representation, and the three change operators of expansion, revision and contraction. For each of these operators, relevant properties are proven. In the fifth and last section I will compare the model developed here with the only other model of preference change known to me, by Sven Ove Hansson.

[1] Attempts to show empirically that diachronic preferences are stable suffer from this lack of a dynamic theory. As an example, see Landsburg (1981).

4.2 A Theory of Preference Change

A theory of preference change starts on the same basis as a static preference theory. Observations of present and – within a limited horizon – past observations of behaviour at time t are fitted under a pattern of preferences and other mental properties as the causes of the observed behaviour. Now the preferences ascribed on this basis are employed as diachronic preferences: under the *principle of unwarranted divergence*, it is assumed that the ascribed preference ordering remains stable in time unless there are discernible causes that change it. Methodologically, this principle is a bit of a *carte blanche*, as the theory of preference is not sufficiently far developed to comprehensively cover the causal mechanisms that change preferences. So even if one subscribes to the principle of unwarranted divergence, it will not help much as we do not know what sort of causes to look for. Instead, one is thrown back upon behavioural observations as the only evidence available for preference change.

Evidence of this sort comes in the guise of future behaviour which does not fit the preference ordering assigned. An agent's behaviour that contradicts the preferences assigned to her can be interpreted in five ways.

4.2.1 Five Kinds of Interpreting Contradictory Behaviour

First, the agent can simply have made a mistake. Due to a misperception of her own preferences, due to beliefs not concurrent with available evidence, or due to lack of computational effort, she acted in a way contrary to her own preferences. The correctness of this interpretation can be tested in cases where the ascriber has recourse to post-observational interviews. If it can be communicated to the agent that she acted contrary to the preferences assigned to her, and she agrees with this verdict, this interpretation is confirmed.

If the agent however does not agree that she made a mistake, maybe we can identify her behaviour as irrational in the sense that – due to external causes – it was not determined by her preferences and other mental properties in the way the theory predicts. Momentary influences like drug abuse, physical exhaustion or illnesses, as well as permanent defects like amnesia[2] or the loss of sensory-

[2] Particularly interesting here is the case of losing one's memory while retaining all reasoning faculties, called *Korsakov's Syndrome*

conceptual capacities.[3] If any of these or similar causes can be observed, the irrationality-interpretation will explain why the observed actions were inconsistent with the ascribed preferences. If the disturbance is momentary, one can expect to explain or predict the behaviour after the disturbing cause has ceased (for example, after sobriety is restored). In permanent disturbances, obviously, the applicability of a preference theory is annihilated, unless the disturbance had only a local effect, such that a new pattern could be established.

If the agent rejects the possibility of error, and no external disturbing causes can be identified, a third interpretation of the inconsistency between observed behaviour and assigned preferences is that the description of the deliberative situation the agent supposedly faces is incorrect. In particular, what has to be re-assessed is the description of the relevant alternatives that the agent faces, and over which she has preferences that determine her behaviour. Such an interpretation has been offered, for example, by Broome in defense of Savage's subjective expected utility theory against the arguments derived from Allais' Paradox. Broome argues that the specification of the situation by Allais is incomplete, and requires a further individuation of outcomes. Given that there is indeed a criterion for reasonable individuation (as Broome 1991, 103 provides), the full specification shows that agents who choose the seemingly erratic combination in the Allais situation do indeed comply with the sure-thing principle.

Attempts to explain seemingly erratic behaviour (i.e. behaviour inconsistent with preferences ascribed to an agent) by further individuating the alternatives can be found for example in Freud's theory of repression. It postulates that a desire – under appropriate conditions – not only causes behaviour that leads to the desire's satisfaction; rather, an unfulfilled desire as well produces anxiety about its satisfaction. Behaviour that seemingly contradicts the existence of a particular constellation of desires then is explicable as the attempt to thwart the anxiety produced by these unfulfilled desires.

Similarly, agents might have preferences on which they act but which they consciously withhold from publication, thus giving rise

[3] O. Sacks describes the case of a patient who lost his capacities to see due to a localised stroke. He had lost not only his eyesight, but 'he had lost the very idea of seeing – and was not only unable to describe anything visually, but was bewildered when I used words such as "seeing" and "light" ' (Sacks 1985, 39).

to the impression that their behaviour is inconsistent with prior ascribed preferences. Such a concealment tactic is often exhibited in legal cases, as illustrated tongue-in-cheek in the court case of the former butler of the Princess of Wales:

> The police report listed the many explanations Mr. Burrell gave for not having returned objects to the family. They include not getting around to it, being too traumatised by Diana's memory to confront certain objects, keeping things for sentimental reasons, guarding things that she expressly meant him to have, being unaware that a particular object was in his possession, thinking something was not appropriate for the charity to which it was destined and taking off her hands gifts she had never liked in the first place. Mr. Boyce [the lead prosecutor] commented dryly, "You may observe some inherent contradiction between these various explanations, all of which, according to the statement, were existing in the same man during the same period". ('Diana's Faithful Butler: In the End, Was He False?', New York Times 18.10.2002)

The multiplicity of motives at the same time indicates, or so the prosecutor at least implies, that some crucial motive is missing, or is indeed concealed by the agent. Instead, the agent offers different motives for similar actions, as if to contradict the impression that there was a common pattern displayed in his behaviour.

The above interpretation already proves the preference theory partly wrong: it attributed the wrong preferences, by lacking insight into the correct causal relations and by neglecting the degree of individuation in the preferences' objectives. If no reasons can be found for this interpretation either, one might want to go further and denounce the whole framework of the preference theory as flawed: its assumption of a particular causal pattern which underlies behaviour, that this pattern is ascribable on the basis of behavioural evidence alone, and that the ascribed mental properties are governed by logical consistency. Examples of this interpretation are the development of regret theory in reaction to the preference-reversal phenomena, or the 'toolbox' approach to bounded rationality in reaction to experimentally observed systematic deviations from probability calculus. However, this interpretation is reserved to deviations that are *sys-*

tematic, and offers itself only if an alternative theoretical framework is available.

If none of the above interpretations apply, another possibility is that the preferences the agent had at a particular time have *changed* in such a way that they produced the observed behavioural effect inconsistent with the original preferences.

According to the principle of unwarranted divergence, this interpretation depends on the presence of a potential preference-changing cause. Thus, a preference change is diagnosed only if there is behavioural evidence for it *and* a cause for it can be identified. Notice that now the reasoning does not go from the presence of a cause of preference change to the preference change itself. Rather, the argument starts from the observed inconsistency of behaviour with the prior preference ordering. It then proceeds through the various potential explanations for such an inconsistency. It concludes that preference change is the correct explanation if some potential cause for such preference change can be identified.

The role of the causes of preference change is so weak in this explanation because no proper theory of them has been successful as of yet. The literature on potential causes of preference change is extensive; but their unifying moment is that the causes they identify are neither necessary nor sufficient for a change in preferences. Even if the list of causes offered in these investigations were comprehensive, one could not infer from the observation of any of them that a preference change in a particular direction has occurred. Given that one knows that the agent has been exposed, for example, to so many billboards praising a low-carbohydrate diet, we still cannot infer that the agent now prefers a diet rich in protein and fat to one rich in carbohydrates. The conclusion is only that some agents are influenced and some are not – and however detailed the investigation becomes, none has so far managed to provide robust predictions of preference change from the presence of certain causes.[4]

Instead, the existence of such causes functions as *secondary* evidence for the preference change interpretation in the following way. *If* observed behaviour is inconsistent with the agent's preferences, *if* external interfering causes do not exist and *if* mistakes are ruled out, *then* the existence of potentially preference-changing causes deter-

[4]It seems that all that has been accomplished in this field so far are categorisations of preference change. Two good examples of these are Elster (1982) and Bowles (1998).

mines the interpretation of the inconsistency as a preference change. Accepting this interpretation, the – seeming – inconsistency becomes explicable, and further preference changes predictable.

A theory that takes the inconsistency of behavioural evidence and attributed preferences as principal evidence for preference change – supported by the existence of some potential cause for preference change – is not *ad hoc*. On the basis of the evidence, the theory first of all modifies the prior preference ordering in such a way that it is consistent with the observed behaviour. But this initial modification will require further modifications. These modifications, in turn, have potential consequences on the agent's behaviour, which can be observed under the right conditions. Thus, the preference change theory proposed here is not a mere integration of observed evidence into the theory for reasons of immunization, but it proposes a more general modification of the preference ordering that has empirical significance beyond the observations made.

Given the observation of an appropriate cause of preference change, the theory takes the observed behaviour as the evidential *input* and models the *collateral change* of preferences on this basis. By focusing on the input and collateral consequences, the theory proposed here disassociates causes from inputs. Causes function only as secondary evidence, while preferences revealed in observed behaviour function as input. It is important not to confuse causes and preference inputs in this context.

4.2.2 The Principles of a Theory of Preference Change

The collateral preference change is governed by four principles: consistency, success, conservatism and entrenchment.

Consistency is the principle that regulates synchronic as well as diachronic preferences. Consistency requirements – typically transitivity, completeness, asymmetry, plus requirements on preferences over lotteries – stipulate or prohibit the existence of certain binary preferences conditional on the existence of others. When new binary preferences are introduced into the ordering as the preference change input, these requirements will stipulate the introduction or elimination of further preference pairs. To function as an ordering at all, the preference ordering has to satisfy all consistency requirements after the initial input was introduced.

Success requires that the preference ordering regain consistency without eliminating the input. The latter might sometimes be the easiest way to regain consistency, but clearly defies the purpose. The input *must* be accommodated. The possibility that the preference change was only momentary, and not maintainable in the light of the ensuing computational costs, is a question concerning the causes of preference change and is consequently not debated here.

According to the principle of *conservatism*, no preferences should be given up if not necessitated by a good reason. To understand this principle, two types of reason need to be clearly distinguished: reasons for *holding* a preference and reasons for *discarding* a preference. The arguments for the conservatism principle are related to the computation costs of these respective reasons; and as I shall show, the costs of reasons for holding preferences is much higher than the costs of reasons for discarding them. Thus conservatism can be split into a negative and a positive claim: that one should not in general keep track of the justification of one's preferences, and that one should retain all preferences that one has no reason to discard.

As I discussed in chapter 3, I believe that reasons for holding a preference for a world w_1 over a world w_2 are the preferences over the aspects $p, q, r \ldots$ that w_1, w_2 respectively realise. For example, the reasons that Mrs. Juniper prefers Paris for her next weekend-vacation destination over Barcelona and Barcelona over Moscow are as follows: she thinks that Paris has the most interesting art collections, followed by Moscow and then Barcelona; further, she thinks Paris has a cosmopolitan air to it, while Barcelona strikes her as a little touristy (she doesn't know what to make of Moscow in this category); the climate strikes her as best in Barcelona, followed by Paris and then Moscow; she knows Moscow and Paris are expensive, while Barcelona is moderate in its cost; the best food with the greatest choice she will get in Paris, followed by great food, albeit a limited variety, in Barcelona, while Moscow's culinary promises are dubious at best. Mrs. Juniper has preferences over these aspects, and the result of aggregating these preferences lead her to prefer Paris to Barcelona to Moscow for her next trip.

According to this view, the reason for holding a preference $a \succ b$ consists of three elements. First, the set of possible worlds is partitioned in such a way that one part represents a feature realised

by a while the other part represents a feature realised by b. Infinitely many partitions are possible, of course, but only some will be relevant to Mrs. Juniper. She might for example find irrelevant the characterization of Moscow as the city with the highest rate of luxury cars, followed by Paris, with Barcelona trailing last; maybe because she does has no evaluation associated with it, or because she is not even aware of this fact. How many partitions are made to characterise the different options, as well as what these partitions are, is therefore an essential part of the reasons for preferences.[5] Second, a preference relation is defined over these partitions. The aspects associated with each option then are preferred or dispreferred to the aspects associated with another option. Third, if it is not the case that all aspects associated with one option are preferred to the aspects associated with another option, there must be a process of weighing the preferences. As noted in section 3.4.3, there are no good solutions available to this process of aggregating prospect preferences yet, but that should not concern us here. If it is correct that reasons for holding a preference over options are the preferences over the prospects, aspects or properties the relata of the target preference respectively realise, then there must be a process that involves forming partitions, specifying preferences over these partitions and weighing those preferences.

A reason to discard a preference, on the other hand, is based on the principles of consistency and success. For example, an agent has a preference $a \succ b$. She then comes to prefer $b \succ a$. By consistency, this new preference provides a reason to discard the preference $a \succ b$.

Comparing reasons for holding preferences with reasons for discarding preferences thus reveals two differences. First, a reason for a preference is always a reason only for *that* preference. The combination of specific partitions, aspectual preferences and their weighing justifies only this one specific preference. The reasons to discard a preference are far more universal: they are the general principles of consistency and success. Second, a reason for a preference might go back into the past, and involve a large number of elements. Reasons for discarding preferences, on the other hand, are always

[5] Just how important this aspect is for evaluation is betrayed by the effort the advertising puts into manipulating people's partitioning. Successful examples of newly introduced partitions include: 'cool – uncool', 'diet/light – non-diet',' '80s – '70s', 'individualistic – conformist'.

determinable in the present: consistency is violated now by such and such a preference, or success needs to be satisfied now. So in both comparisons the reasons to discarding preferences come out more parsimonious, more accessible and thus less expensive than reasons for holding preferences.

The negative claim of conservatism states that agents should not keep track of the justifications for all the preferences they hold, because the costs are high and the benefits dubious. Costs for keeping track of all justifications are high, because human resources of thought are scarce, and filling up memory space with 'clutter' (Harman 1986, p. 41) is wasteful of these scarce resources. Given what I said about the elements of reasons for holding preferences, the partitions, aspect preferences and weighing can indeed amount to quite some clutter. Take for example one's preferences over different kinds of baby food. Once this was an extremely important issue, and one had clear preferences for some kinds over others, based on specific aspects of the different foods – say the delightful mushiness of the carrots versus the displeasing tartness of the spinach. But one is very unlikely to ever have to choose between eating any of those again – hence why should one maintain all one's reasons for those preferences?

Further, the benefit might easily not match up to the high costs. Because of the instability and manipulability of the partitions, secure justification are very hard to come by. There is always the chance that one's senses did not work correctly when one experienced something, or that one deceived oneself. Hence a reason one had for holding a preference might have always been a bad reason. The mere possibility of such an error – as small as it may be – puts even more emphasis on the cost-benefit argument. If an agent does not have rock-bottom justified preferences in the first place, then the high costs involved in tracing everything back to some presumed foundations are even less likely to result in a net gain.

The positive claim of conservatism states that people do and should retain all preferences which they have no reason to discard. First, this is again a question of cost efficiency. To discard preferences one holds is a difficult business, be it to stop desiring a cigarette or to distance oneself from a person one was close to – how-

ever good one's reasons are. Preferences, unlike beliefs, are strongly habitual, and to abandon a habit one's reasons deem bad is costly.[6] In order to minimise costs, no one will or should discard preferences if she has no good reasons to do so. Second, the agent might assume that she never embraced a preference without a reason. The reason might now be forgotten, but as long as it is not actively contradicted by a reason to eliminate or revise the preference, nothing should be done about it.

But there are situations where there are equally good reasons to discard one preference or the other. For example, an agent who holds preferences $\{a \succ b, b \succ c, a \succ c\}$ and then comes to prefer $c \succ a$ by success and consistency has a reason to discard $a \succ c$ and either $a \succ b$ or $b \succ c$. In those situations, the principle of conservatism is of no help to the agent. Instead, the agent has a choice between three actions. She can randomly choose which preferences to eliminate, she can exploit further reasons she might have for retaining one preference over another, or she might violate conservatism and discard all preferences in question. I will discuss these three options in turn.

First, the agent might randomly decide which preference to eliminate. This option satisfies conservatism, but introduces capriciousness, and hence should be rejected.

Second, in special cases like these, the agent might revert to reasons for holding preferences and eliminate that preference which she has less good reasons to hold. In such a case we say that the retained preference is more deeply *entrenched* than the discarded preference. Entrenchment orders preferences in a ranking of 'having better reasons for this preference than...'. It is important to note that entrenchment is a property of a binary preference relation, not of the relata of that preference. Following my discussion of reasons for holding preferences, I think that a preference is the more deeply entrenched (i) the more unanimous the aspectual preferences are in its favour, (ii) the larger the number of aspectual preferences that support it and (iii) the more salient the partitions that underlie all the relevant aspectual preferences. Because entrenchment represents reasons for holding a preference, it faces the same cost-benefit con-

[6] Contrast this with the discussion and rejection of the notion of beliefs as habits of thought in Harman 1986, 38-41.

siderations discussed above. However, entrenchment is called upon only in particular situations, where the other three principles will not yield a unique result. Given the cost considerations, one can therefore expect that entrenchment information is not available for most preferences.

The third option the agent has – given she has no entrenchment information available and randomization is rejected – is to eliminate *both* preferences. This case has some intuitive plausibility: if things grow too complicated, if there is evidence that our values are inconsistent with a preference we have strong reasons to hold but cannot determine which of the values is the 'culprit', we do have a reason to suspend *all* the value judgements inconsistent with that preference. But if we agree with that, then conservatism is violated – the agent removes both his preferences, even though he could have remedied the inconsistency by removing only one. To be more exact, conservatism is violated in relation to those preferences which are potential inconsistency makers. In the broader picture, for all other of the agent's preferences, conservatism remains intact. I will call this the principle of *weak conservatism*.

In the following section, I will discuss an example of preference change and illustrate the applications of the four principles discussed here.

4.3 Three Illustrations of Input-Driven Collateral Preference Change

Hubert, a culinary greenhorn, moves to Cuisineville with its three restaurants (Thai, Italian and Chinese). At his arrival, he prefers Thai to Italian, but he hasn't made up his mind about the other cuisines. Thus Hubert holds the preference $t \succ i$, which is depicted as the directed arrow in figure 4.1.

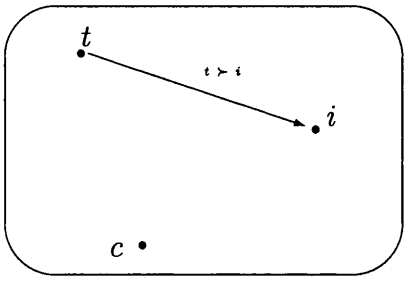

Figure 4.1

Then, after repeated visits to both the Italian and the Chinese restaurant, he comes to prefer Italian over Chinese. Thus Hubert now holds the preferences $t \succ i$ and $i \succ c$, and the preference between Italian and Chinese is depicted in figure 4.2.

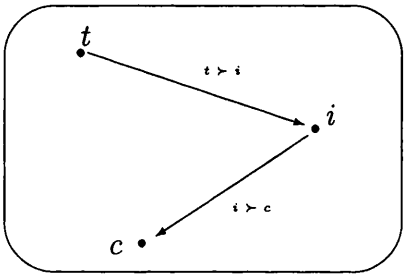

Figure 4.2

But given that he holds both of these preferences, transitivity now requires him to hold a third – he is forced upon the threat of inconsistency to prefer Thai over Chinese. Thus the moment he obtains his second preferences, the principles of success and consistency require that he has to accommodate a third preference such that he holds $t \succ i, i \succ c$ and $t \succ c$, as illustrated in figure 4.3.

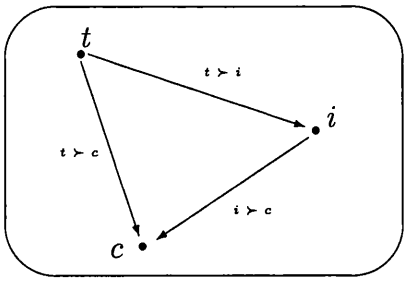

Figure 4.3

Hubert has now acquired a complete preference ordering over t, i, c. Some *Lo Mein* later, Hubert realises that he now certainly prefers Chinese over Thai. If this new input is simply integrated into the existing ordering, it results in a preference cycle as illustrated in figure 4.4.

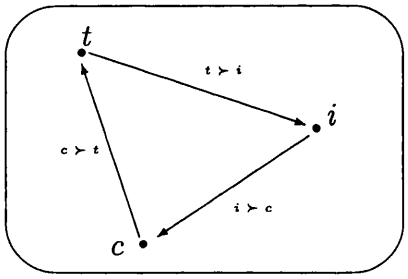

Figure 4.4

The ensuing preferences $t \succ i, i \succ c$ and $c \succ t$ violate transitivity. To honour consistency and success, the whole preference ordering has to be transformed. Three different options are open: (i) the other preference pairs are removed. This cannot be done individually, as removing either $t \succ i$ or $i \succ c$ would leave the resulting ordering still violating transitivity. Thus both $t \succ i$ and $i \succ c$ need to be removed to accommodate $c \succ t$. This in turn violates conservatism. Instead (i) $t \succ i$ is changed to $i \succ t$ or (iii) $i \succ c$ is changed to $c \succ i$. Both (ii) and (iii) satisfy conservatism, as the ensuing ordering retains one preference pair of the order depicted in figure 4.3 each. But conservatism provides equally good reasons for (ii) or (iii). If one must not randomise, as I have argued, additional entrenchment information is needed to choose between (ii) and (iii).

If such information is not forthcoming, then Hubert will revert to option (i) and remove both preferences that *together* contradict his new preference $c \succ t$. Thus either Hubert holds only $c \succ t$, or he holds $c \succ t, i \succ c$ and $i \succ t$, or he holds $t \succ i, c \succ i$ and $c \succ t$. To continue with Hubert's story, let's assume he changes his preferences to $t \succ i, c \succ i$ and $c \succ t$. That is, Hubert's preference for t over i is more entrenched than his preference for i over c; he held the former for longer and has made many more comparisons between t and i than between i and c. His preference ordering is then illustrated in figure 4.5

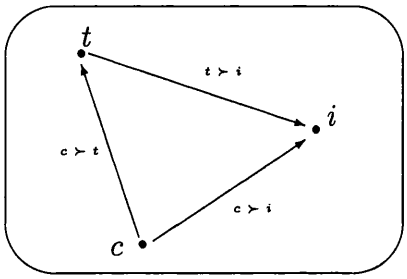

Figure 4.5

After further explorations of the local culinary scene, and after the exposure to some rather exotic Chinese fish eye and jellyfish dishes, Hubert loses confidence in his preference for Chinese over Italian cuisine. This preference retraction leads to the problematic preference constellation depicted in figure 4.6.

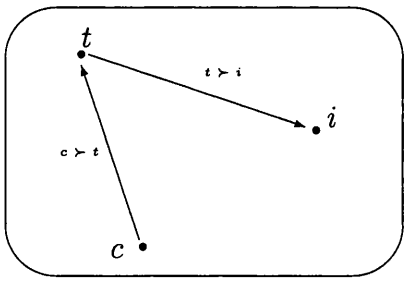

Figure 4.6

But this ordering again violates transitivity. To honour success and consistency, the whole preference order needs to be transformed.

Hubert has three options to do so: he can remove either of the preference pairs or both. Again, this choice will depend on the available entrenchment information. As we know from before, Hubert's preference over t and i are most deeply entrenched, hence the collateral change results in the removal of $c \succ t$ and Hubert's final order consists only of $t \succ i$, depicted in figure 4.7.

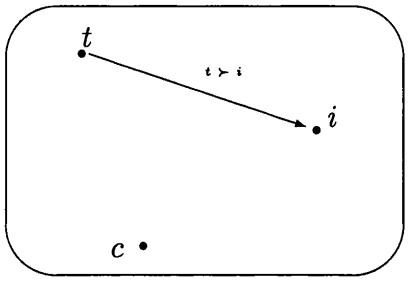

Figure 4.7

4.4 The Model

4.4.1 General Considerations

The model presented here consists of a representation of an agent's diachronic preferences, and three operations of preference change defined over this representation, which are contingent on a preference input. Before I will present the formal model itself in section 4.4.2, I will discuss some general features of the preference representation, as well as some postulates for the operators that connect it to the four principle of preference change theory.

An agent's preferences are represented as a set of binary relations over a set of mutual exclusive alternatives. The assumption of mutual exclusivity is made here for reasons of simplicity. In an environment where some alternatives imply other alternatives or where alternatives are probabilistically correlated, instrumental dependencies arise which have to be reflected in the preference structure. These dependencies complicate the model and prohibit a clear focus on the mechanism of collateral change. I have discussed the different kinds of preferences and their relation to each other in section 3.2. With the principle of equivalence provided there, the framework

used here could be expanded; but now I just want to provide the basics of any such model.

The preferences of the agent are assumed to be reflexive and transitive, but not complete. Completeness means that any two alternatives are connected. The assumption that an agent's preferences are complete is very unrealistic. First, to establish and to store preference information of this magnitude is very costly, and is not always justified by the benefits derived from a complete preference ordering. For example, a preference for a over b and for a over c in a three-alternatives environment gives the agent a full recipe what to choose; the preference comparison between b and c is irrelevant as long as a is available. The agent can now consider the probability with which a might become unavailable – contrary to her earlier assessment of the situation – and then compare the expected loss of having to choose between b and c without preference guidance with the costs of establishing a preference between b and c. Second, it might be categorically impossible to form a reasonable preference between two alternatives – e.g. whether one prefers the destruction of Mars to that of Venus – because not enough information about the alternatives is available to form any judgement, or because one has good reasons to refuse a judgment.[7] Third, completeness restricts the scope of preference change. As I will argue, there are three important types of preference changes: expansion, contraction and revision. If preference orderings were complete, non preference could be added to it, and none could be removed. As the examples of section 4.3 showed, a model that only deals with revision does not capture the full width of preference change.

On the other hand, a complete ordering is technically easier to handle. In particular, it can be represented by a utility function, and allows – with the addition of a probability measure – to construct an expected utility index. To combine the technical advantages of completeness with the argument against completeness presented above, an agent's preferences will be modelled by the intersection of a set of total orders. The intuitive idea behind this is that the

[7]Bernhard Williams gives an example of a moral dilemma, and proposes that in a situation like this, a reasonable reaction is to withhold judgment. A traveler in remote and dangerous territory stumbles across a guerrilla group, whose leader promptly offers the unexpected guest a delicate choice: if he chooses one of their ten hostages to be executed in the traveler's honour, the other nine will go free. If he prefers not to choose, the guerrilla will murder all ten. Compare Williams(1981).

incomplete order is represented by all those total orders which are possible ways to 'fill in' the incomplete order.[8]

I will think of preference change as induced by an input command of the following sort: (i) 'Add $a \succ b, \ldots, y \succ z$ to the preference order of the agent'; (ii) 'Eliminate $a \succ b, \ldots, y \succ z$ from the preference order of the agent'; (iii) 'Revise the agent's preference order by $a \succ b, \ldots, y \succ z$'. An input command thus consists of a change command and a string of preference sentences. There are only three commands: expansion, contraction and revision. They correspond to the three cases illustrated in Hubert's example.

The input to each of the three operators is a set of total orders itself. The intersection of this set represents the preference by which the preference ordering is to be expanded, contracted or revised. This way, the model easily deals with multiple changes, and singleton changes as a special case of them.

The change operators are generally restricted by the principles of consistency, success, conservatism and entrenchment. The first requirement is that the result of applying a change operator to a preference ordering must be a preference ordering itself – it must satisfy reflexivity and transitivity. This *closure* property is a specification of the consistency requirement for theories of preference change. In the model presented here, its satisfaction is greatly facilitated through the mode of representation. The preference ordering is modelled as the intersection of total orders. As each total order satisfies the consistency axioms – otherwise it would not be an order – their intersection similarly has to satisfy these axioms.

When changing a preference ordering, the new preference ordering has to incorporate the commanded change. If expanded or revised by a number of preferences, then each of these preference must be represented in the new set; if contracted by a number of preferences, then none of these preferences must be represented in the new set. This way, the change operators satisfy the *Success* postulate.

Further, change operators should change preference orderings only to a minimal degree. I have discussed this principle of conservatism in section 4.2.2. Applied to change operators, four postulates

[8] Beyond the technical convenience exploited here, this representation has the further advantage that an incomplete ordering can be represented by a set of utility functions. This idea has been developed in Aumann (1962) and Seidenfeld, Schervish, Kadane (1995); existence and uniqueness proofs for such a representation are given in Dubra, Maccheroni, Ok (2004).

can be distinguished, although some of them apply only to specific operators. First, an expanded preference ordering should include the prior non-expanded ordering; while a contracted preference ordering should be the subset of the prior non-contracted ordering. This is captured in the postulate of *Inclusion* for expansion and contraction respectively. Second, under specific conditions, a change command should be vacuous. To expand or revise an ordering by a preference that is already in it should not change the ordering at all; neither should the contraction of an ordering by a preference that is not in it. These commands should leave the prior ordering completely unaltered. The *Vacuity* postulate determines that under these conditions, the new preference ordering is the same as the old one. Third, if a preference ordering and its subset are expanded by the same preferences, and none of these input preferences make any of the two sets inconsistent, then the expanded preference ordering must still include its prior subset. This *Monotonicity* postulate only applies to expansion. It will be shown with an example that monotonicity does not hold for contraction. Last, changes induced by identical inputs should result in the same changed preference orderings. This *Extensionality* postulate applies to all three change operators.

To prove the satisfaction of these postulates for the constructed model is a first step on the way to a representation result. The proof that the model satisfies the postulates is a form of existence result: given that the postulates are well chosen, it is shown that the structure constructed here represents and models preference change. However, this is possibly not the only structure available. To prove that, we have to show that if an operator satisfies the postulates, then it must have the structure presented here. Last, it is desirable to show that all possible forms of preference change – all possible forms of preference model permutations – can be obtained through the change operators constructed here. This chapter will only prove the existence results.

Very mechanical? Looking for a formal solution to a rationalistic problem?

4.4.2 The Formal Structure

Let A be the finite set of mutually exclusive alternatives $a, b, c \ldots$. Further, let $\mathbb{S}(A \times A)$ be the superset of the Cartesian product of A with A. Then the set of all total orders, $\mathbb{P}_{A \times A}$, is defined as:

Definition 7 $\mathbb{P}_{A\times A}$ *is the set of all* $X \in \mathbb{S}(A \times A) \setminus \emptyset$ *such that*

1. *for all* $a \in A, \langle a, a\rangle \in X$.
2. *for all* $a, b, c \in A$ *if* $\langle a, b\rangle \in X$ *and* $\langle b, c\rangle \in X$ *then* $\langle a, c\rangle \in X$.
3. *for all* $a, b \in A$ *either* $\langle a, b\rangle \in X$ *or* $\langle b, a\rangle \in X$

$\mathbb{P}_{A\times A}$ is thus defined as the set of all those members of $\mathbb{S}(A \times A)$ which are reflexive, transitive and complete. Each of these $R \in \mathbb{P}_{A\times A}$ is complete in the sense that *all* elements of the alternative set A appear in a binary tuple $\langle a, b\rangle \in R$ at least once. However, the number of tuples differs between different R because some tuples might be symmetric – for every pair a, b, R might contain $\langle a, b\rangle$ but not $\langle b, a\rangle$; or it might contain $\langle a, b\rangle$ *and* $\langle b, a\rangle$.

An agent's (incomplete) preferences are represented by a set $\mathbb{R} \in \mathbb{P}_{A\times A}$, whose intersection $\bigcap\mathbb{R}$ is the set of all the binary comparisons endorsed by the agent. An agent *strictly* prefers a to b iff $\langle a, b\rangle \in \bigcap\mathbb{R}$ but not $\langle b, a\rangle \in \bigcap\mathbb{R}$; she is indifferent between A and B iff $\langle a, b\rangle \in \bigcap\mathbb{R}$ and $\langle b, a\rangle \in \bigcap\mathbb{R}$.

The relation between the Rs in $\mathbb{R}$ and $\bigcap\mathbb{R}$ is as follows. The agent holds an incomplete preference ordering $\bigcap\mathbb{R}$, which is represented as the intersection of a set of total orders, $\mathbb{R}$. One might think of these total orders as the orders compatible with the agent's preferences, as the agent's *potential* complete preference orderings. The more potential preference orderings the agent has, the less specific are his preferences, and vice versa. Thus the more total orders are members of $\mathbb{R}$, the less specific the preference ordering $\bigcap\mathbb{R}$, as shown in the following corollary.

Corollary 1 $\mathbb{Q} \subseteq \mathbb{R} \Rightarrow \bigcap\mathbb{R} \subseteq \bigcap\mathbb{Q}$

Proof *Case 1:* $\mathbb{Q} = \mathbb{R} \Rightarrow \bigcap\mathbb{Q} = \bigcap\mathbb{R}$. *Case 2:* $\mathbb{Q} \subset \mathbb{R}$. Then every member of $\mathbb{Q}$ is in $\mathbb{R}$ but there is at least one member of $\mathbb{R}$ which is not in $\mathbb{Q}$: $\mathbb{Q} = \{Q_1, \ldots, Q_m\}$, $\mathbb{R} = \{Q_1, \ldots, Q_m, R_1 \ldots, R_n\}$, with $Q_i \neq R_j$ for all i, j. Thus $\bigcap\mathbb{R} = \bigcap\{Q_1 \ldots, Q_m, R_1 \ldots, R_n\} = Q_1 \cap \ldots \cap Q_m \cap R_1 \cap \ldots \cap R_n \subseteq \bigcap\mathbb{Q} = \bigcap\{Q_1, \ldots, Q_m\} = Q_1 \cap \ldots \cap Q_m$.□

However, one and the same $\bigcap\mathbb{R}$ can be obtained from different $\mathbb{R}$'s.[9] In order to avoid ambiguities here, it is stipulated that $\mathbb{R}$ is

[9]This is because a total order consisting of some indifference relations always contains at least one total order consisting of strict preferences only. Either of these orders combined

maximal, i.e. *all* total orders that can represent $\bigcap \mathbb{R}$ are actually in $\mathbb{R}$.

Definition 8 $\mathbb{R} \subseteq \mathbb{P}_{A\times A}$ *is maximal iff for all* $R \in \mathbb{P}_{A\times A}$: *if* $R \notin \mathbb{R}$ *then there is a tuple* $\langle a, b\rangle$ *such that:* $\langle a, b\rangle \in \bigcap \mathbb{R} \wedge \langle a, b\rangle \notin \bigcap(\mathbb{R} \cup \{R\})$

Maximality requires that the addition of any further total order to $\mathbb{R}$ will exclude a tuple from $\bigcap \mathbb{R}$. With maximality, the following corollary can be shown.

Corollary 2 $\bigcap \mathbb{R} \subseteq \bigcap \mathbb{Q}$ *and* $\mathbb{R}$, $\mathbb{Q}$ *are maximal* $\Rightarrow \mathbb{Q} \subseteq \mathbb{R}$

Proof Let's assume that $\bigcap \mathbb{R} \subseteq \bigcap \mathbb{Q}$, $\mathbb{R}$ and $\mathbb{Q}$ are maximal and $\mathbb{Q} \nsubseteq \mathbb{R}$. Without loss of generality let the difference between $\mathbb{Q}$ and $\mathbb{R}$ be the total order X. By maximality of $\mathbb{R}$, there is a tuple $\langle a, b\rangle$ which is a member of $\bigcap \mathbb{R}$ but not of $\bigcap(\mathbb{R} \cap \{X\})$. But by assumption $\mathbb{R} \cap \{X\} = \mathbb{Q}$, and thus $\bigcap(\mathbb{R} \cap \{X\}) = \bigcap \mathbb{Q}$. Hence $\bigcap \mathbb{R} \nsubseteq \bigcap \mathbb{Q}$, which contradicts the antecedent. □

Corollary 2 clarifies the significance of the maximality requirement. If $\mathbb{R}$ and $\mathbb{Q}$ are maximal, and $\bigcap \mathbb{R} = \bigcap \mathbb{Q}$, then it follows from corollary 2 that $\mathbb{R} = \mathbb{Q}$. To obtain a unique representation of incomplete preference orderings, we represent the agent's preferences by *maximal* subsets $\mathbb{R} \in \mathbb{P}_{A\times A}$.

The following corollary shows that $\bigcap \mathbb{R}$ indeed represents an (incomplete) preference ordering.

Corollary 3 $\bigcap \mathbb{R}$ *is reflexive and transitive.*

Proof *Part 1, reflexivity:* Let's assume $\bigcap \mathbb{R}$ is not reflexive. Then there is a tuple $\langle a, a\rangle \notin \bigcap \mathbb{R} = \{R_1 \cap \ldots \cap R_n\}$. By definition of $\cap$ follows that for some i, $\langle a, a\rangle \notin R_i$, which contradicts Definition 7.1.

Part 2, transitivity: Let's assume $\bigcap \mathbb{R}$ is not transitive. Then there are some alternatives $a, b, c \in A$ such that $\langle a, b\rangle, \langle b, c\rangle \in \bigcap \mathbb{R}$ and $\langle a, c\rangle \notin \bigcap \mathbb{R}$. Then there must be some member R_i of $\bigcap \mathbb{R}$ for which $\langle a, b\rangle, \langle b, c\rangle \in R_i$ and $\langle a, c\rangle \notin R_i$, which contradicts Definition 7.2. □

---

with a third might represent the same preference ordering, but obviously they constitute different $\mathbb{R}$. Take for example following. $\mathbb{R}_1 = \{\{\langle a, b\rangle, \langle a, c\rangle, \langle b, c\rangle\}, \{\langle a, b\rangle, \langle a, c\rangle, \langle c, b\rangle\}\}$ and $\mathbb{R}_2 = \{\{\langle a, b\rangle, \langle a, c\rangle, \langle b, c\rangle\}, \{\langle a, b\rangle, \langle a, c\rangle, \langle c, b\rangle\}, \{\langle a, b\rangle, \langle a, c\rangle, \langle b, c\rangle\ \langle c, b\rangle\}\}$. Then $R_1 = \{\langle a, b\rangle, \langle a, c\rangle, \langle b, c\rangle\}$ is both in $\mathbb{R}_1$ and $\mathbb{R}_2$, as is $R_2 = \{\langle a, b\rangle, \langle a, c\rangle, \langle c, b\rangle\}$. But $R_1, R_2 \subset R_3 = \{\langle a, b\rangle, \langle a, c\rangle, \langle b, c\rangle\ \langle c, b\rangle\}$ which is a member only of $\mathbb{R}_2$. Hence $\mathbb{R}_1 \neq \mathbb{R}_2$, but $\bigcap \mathbb{R}_1 = \bigcap \mathbb{R}_2$.

Hence $\bigcap \mathbb{R}$ is a representation of an incomplete preference ordering. In exactly the same way, the input set $\mathbb{I}$ is a set of total orders whose intersection represents an incomplete preference ordering. In the following subsections, three kinds of collateral change operators are constructed, and relevant postulates of these change operators are proven.

Expansion introduces one or more new preferences represented in $\bigcap \mathbb{I}$ to the preference ordering $\bigcap \mathbb{R}$, thus changing it to $\bigcap \mathbb{R}_{\cap\mathbb{I}}$. The expansion operator $\cdot_{\cap}$ removes certain sets from $\mathbb{R}$ and hence makes $\mathbb{R}$ more specific such that $\bigcap \mathbb{R}_{\cap\mathbb{I}}$ now also represents the preferences in $\bigcap \mathbb{I}$. Four cases can be distinguished. The trivial case is where $\mathbb{R} \subseteq \mathbb{I}$. Here $\mathbb{R}$ is already more specific than $\mathbb{I}$ and no actual expansion takes place. Second, $\mathbb{I} \subseteq \mathbb{R}$, where $\mathbb{I}$ is more or equally specific than $\mathbb{R}$ concerning every possible preference sentence. Hence $\mathbb{R}$ is replaced by $\mathbb{I}$ in the expansion. Third, $\mathbb{R}$ and $\mathbb{I}$ shares some members in common, without one including the other. Then $\mathbb{I}$ is more specific about some preferences than $\mathbb{R}$, but less so for others. Hence those members which make $\mathbb{R}$ less specific than $\mathbb{I}$ are excluded in the expansion. Fourth, if $\mathbb{R}$ and $\mathbb{I}$ do not share any members at all, no expansion can take place. The expansion operator $\cdot_{\cap}$ is thus defined as follows:

Definition 9 $\mathbb{R}_{\cap\mathbb{I}} = \begin{cases} \mathbb{R} \cap \mathbb{I} & iff \quad \mathbb{R} \cap \mathbb{I} \neq \emptyset \\ \mathbb{R} & iff \quad \mathbb{R} \cap \mathbb{I} = \emptyset \end{cases}$

The expanded set is maximal just as the prior set was.

Corollary 4 $\mathbb{R}_{\cap\mathbb{I}}$ *is maximal.*

Proof If $\mathbb{R} \cap \mathbb{I} = \emptyset$, $\mathbb{R}_{\cap\mathbb{I}}$ is trivially maximal. Otherwise, $\mathbb{R}$ and $\mathbb{I}$ are maximal, hence for all $R^* \in \mathbb{P}_{A\times A}$ if $R^* \notin \mathbb{R}$ then there is a tuple $\langle a, b\rangle$ such that $\langle a, b\rangle \in \bigcap \mathbb{R} \wedge \langle a, b\rangle \notin \bigcap(\mathbb{R} \cup \{R^*\})$. Thus there is no $R \in \mathbb{P}_{A\times A}$ which is a subset of a member of $\mathbb{R}$ but not in $\mathbb{R}$ itself. And similarly for $\mathbb{I}$: if $\mathbb{I}$ is maximal, there must be no total order I which is a subset of any I_i in $\mathbb{I}$ but not in $\mathbb{I}$ itself. Now $\mathbb{R}_{\cap\mathbb{I}}$ is the intersection of $\mathbb{R}$ and $\mathbb{I}$. Thus $\mathbb{R}_{\cap\mathbb{I}}$ contains only elements that are members both of $\mathbb{R}$ and $\mathbb{I}$. By maximality, if $R_i \in \mathbb{R}$, then all $R_i' \subseteq R_i$ are members of $\mathbb{R}$, too (and similarly for $\mathbb{I}$). Hence if $R_i \in \mathbb{R}_{\cap\mathbb{I}}$, all total orders which are subsets of R_i are in $\mathbb{R}_{\cap\mathbb{I}}$.

But then it is impossible to add another total order to $\mathbb{R}_{\cap\mathbb{I}}$ without removing at least one tuple $\langle a, b\rangle$ from $\bigcap\mathbb{R}_{\cap\mathbb{I}}$.□

The thus defined operator $\cdot_{\cap}$ fulfills the following postulates.

Theorem 3 *1.* $\bigcap\mathbb{R}_{\cap\mathbb{I}}$ *is a preference ordering (Closure)*

2. If $\mathbb{R}\cap\mathbb{I}\neq\emptyset$, $\bigcap\mathbb{I}\subseteq\bigcap\mathbb{R}_{\cap\mathbb{I}}$ *(Success)*

3. $\bigcap\mathbb{R}\subseteq\bigcap\mathbb{R}_{\cap\mathbb{I}}$ *(Inclusion)*

4. If $\bigcap\mathbb{I}\subseteq\bigcap\mathbb{R}$ *then* $\bigcap\mathbb{R}=\bigcap\mathbb{R}_{\cap\mathbb{I}}$ *(Vacuity)*

5. If $\bigcap\mathbb{R}\subseteq\bigcap\mathbb{H}$ *and* $\mathbb{H}\cap\mathbb{I}\neq\emptyset$, *then* $\bigcap\mathbb{R}_{\cap\mathbb{I}}\subseteq\bigcap\mathbb{H}_{\cap\mathbb{I}}$ *(Monotonicity)*

6. If for two input sets $\bigcap\mathbb{I}=\bigcap\mathbb{J}$, *then* $\bigcap\mathbb{R}_{\cap\mathbb{I}}=\bigcap\mathbb{R}_{\cap\mathbb{J}}$ *(Extensionality)*

Proof *Part 1, Closure:* It needs to be shown that $\bigcap\mathbb{R}_{\cap\mathbb{I}}$ is reflexive and transitive. Proof proceeds analogous to Corollary 3.

Part 2, Success: $\mathbb{R}\cap\mathbb{I}\subseteq\mathbb{I}$. Thus, if $\mathbb{R}\cap\mathbb{I}\neq\emptyset$, $\mathbb{R}_{\cap\mathbb{I}}\subseteq\mathbb{I}$. Then, by corollary 1, $\bigcap\mathbb{I}\subseteq\bigcap\mathbb{R}_{\cap\mathbb{I}}$.

Part 3, Inclusion: $\mathbb{R}\cap\mathbb{I}\subseteq\mathbb{R}$. Thus, if (i) $\mathbb{R}\cap\mathbb{I}\neq\emptyset$, $\mathbb{R}_{\cap\mathbb{I}}\subseteq\mathbb{R}$. Then by corollary 1, $\bigcap\mathbb{R}\subseteq\bigcap\mathbb{R}_{\cap\mathbb{I}}$. If (ii) $\mathbb{R}\cap\mathbb{I}=\emptyset$, by definition 9 $\mathbb{R}_{\cap\mathbb{I}}=\mathbb{R}$.

Part 4, Vacuity: From $\bigcap\mathbb{I}\subseteq\bigcap\mathbb{R}$ follows by corollary 2 $\mathbb{R}\subseteq\mathbb{I}$. Then as well $(\mathbb{R}\cap\mathbb{I})=\mathbb{R}$ and from that $\bigcap(\mathbb{R}\cap\mathbb{I})=\bigcap\mathbb{R}$. As by assumption $\mathbb{R}\cap\mathbb{I}\neq\emptyset$, $\bigcap\mathbb{R}_{\cap\mathbb{I}}=\bigcap\mathbb{R}$.

Part 5, Monotonicity: From $\bigcap\mathbb{R}\subseteq\bigcap\mathbb{H}$ by corollary 2 $\mathbb{H}\subseteq\mathbb{R}$, and from that and $\mathbb{H}\cap\mathbb{I}\neq\emptyset$ it follows that $\mathbb{H}\cap\mathbb{I}\subseteq\mathbb{R}\cap\mathbb{I}$ and $\mathbb{R}\cap\mathbb{I}\neq\emptyset$. Thus $\mathbb{H}_{\cap\mathbb{I}}\subseteq\mathbb{R}_{\cap\mathbb{I}}$, and by corollary 1 $\bigcap(\mathbb{R}_{\cap\mathbb{I}})\subseteq\bigcap(\mathbb{H}_{\cap\mathbb{I}})$.

Part 6, Extensionality: From $\bigcap\mathbb{I}=\bigcap\mathbb{J}$ by corollary 2 $\mathbb{I}=\mathbb{J}$. If $\mathbb{R}\cap\mathbb{I}=\emptyset$, then $\mathbb{R}_{\cap\mathbb{I}}=\mathbb{R}_{\cap\mathbb{J}}=\mathbb{R}$. If $\mathbb{R}\cap\mathbb{I}\neq\emptyset$, $\mathbb{R}_{\cap\mathbb{I}}=\mathbb{R}_{\cap\mathbb{J}}$, and thus by corollary 1 $\bigcap\mathbb{R}_{\cap\mathbb{I}}=\bigcap\mathbb{R}_{\cap\mathbb{J}}$.

Contraction *Contraction* removes all the preferences in $\bigcap\mathbb{I}$ from the preference ordering $\bigcap\mathbb{R}$. Such a move has any effect only if $\bigcap\mathbb{R}\cap\bigcap\mathbb{I}\neq\emptyset$. The contraction operator $\cdot_{\cup}$ adds certain sets to $\mathbb{R}$ and hence makes $\bigcap\mathbb{R}$ less specific. The addition should take stock from those total orders which do not represent any preference in $\bigcap\mathbb{I}$; the inverse of $\mathbb{I}$, $\mathbb{P}_{A\times A}\setminus\mathbb{I}$ offers itself here. But only some members

of $\mathbb{P}_{A\times A} \setminus \mathbb{I}$ should be added: we only want to remove from $\bigcap \mathbb{R}$ those preferences which are in $\bigcap \mathbb{I}$. If all members of $\mathbb{I}$'s inverse were added, $\bigcap \mathbb{R}_{\cup \mathbb{I}}$ would be empty as long as $\bigcap \mathbb{R} \cap \bigcap \mathbb{I} \neq \emptyset$. Take the following example to illustrate this point.

Let the agent have preferences over $A = \{a, b, c\}$ of the following sort: $\{a \succ b, b \succ c, a \succ c, etc\}$. Now she wants to contract $b \succ c$ from those preferences. The respective sets $\mathbb{R}$ and $\mathbb{I}$ and the inverse of $\mathbb{I}$ then look like follows:

$\mathbb{R} = \{\{\langle a,b\rangle, \langle b,c\rangle, \langle a,c\rangle, \ldots\} \ldots\}$

$\mathbb{I} = \{\{\langle a,b\rangle, \langle b,c\rangle, \langle a,c\rangle, \ldots\}, \{\langle b,a\rangle, \langle a,c\rangle, \langle b,c\rangle, \ldots\}, \{\langle b,c\rangle, \langle c,a\rangle, \langle b,a\rangle, \ldots\}, \ldots\}$

$\mathbb{P}_{A\times A} \setminus \mathbb{I} = \{\{\langle a,c\rangle, \langle c,b\rangle, \langle a,b\rangle, \ldots\}, \{\langle c,b\rangle, \langle b,a\rangle, \langle c,a\rangle, \ldots\}, \{\langle c,a\rangle, \langle a,b\rangle, \langle c,b\rangle, \ldots\}, \ldots\}$

If all members of $\mathbb{P}_{A\times A} \setminus \mathbb{I}$ were included in the contracted set $\mathbb{R}_{\cup \mathbb{I}}$, the resulting $\bigcap \mathbb{R}_{\cup \mathbb{I}}$ would be empty. This is undesirable, as the contraction in no way concerns the relation between a and b. Such a change is too radical: just because an agent contracts some preferences, she does not necessarily have to give up her whole ordering. Finding the right scope of change is subject to the principles of conservatism and entrenchment, as discussed in section 4.2. The principle of conservatism tells us to retain the relation $a \succ b$. Both the first and the third member of $\mathbb{P}_{A\times A} \setminus \mathbb{I}$ share the tuple $\langle a, b\rangle$ with $\bigcap \mathbb{R}$, while the second member does not share any tuple with $\bigcap \mathbb{R}$. On its basis, the selection function $C(\cdot)$ picks out only the first and the third member of $\mathbb{P}_{A\times A} \setminus \mathbb{I}$.

More generally, a *selection function* $C(\cdot)$ over the inverse of $\mathbb{I}$ is constructed to comply with weak conservatism. $C(\cdot)$ determines those elements of $\mathbb{P}_{A\times A} \setminus \mathbb{I}$ which are most similar to $\bigcap \mathbb{R}$. The similarity comparison makes sure that as many preferences of the old system as possible survive the contraction. Because multiple contraction is allowed (i. e. with more than one tuple in $\bigcap \mathbb{I}$), we have to make sure that not only those members of $\mathbb{P}_{A\times A} \setminus \mathbb{I}$ which are different from $\bigcap \mathbb{R}$ in one instance are picked out. To avoid such a corruption of the similarity measure, members of $\mathbb{P}_{A\times A} \setminus \mathbb{I}$ are compared with the set $\bigcap \mathbb{R}$ of which all the tuples of $\bigcap \mathbb{I}$ are removed. As both the members of $\mathbb{P}_{A\times A} \setminus \mathbb{I}$ and $\bigcap \mathbb{R}$ are sets of binary tuples a similarity metric is defined as the cardinality of the intersection between the two.

Definition 10 $C(\mathbb{P}_{A\times A} \setminus \mathbb{I}) = \{X | X \in \mathbb{P}_{A\times A} \setminus \mathbb{I}$ *and for all* $Y \in$

$\mathbb{P}_{A\times A} \setminus \mathbb{I} : \sharp(X \cap [\bigcap \mathbb{R} \setminus \bigcap \mathbb{I}]) \geq \sharp(Y \cap [\bigcap \mathbb{R} \setminus \bigcap \mathbb{I}])\}$

As I discussed above, weak conservatism has to be supplemented by entrenchment information. This is because of the following problem. Due to the transitivity requirements, an alternative a_1 that is connected to n-1 other alternatives through a preference chain $a_1 \succeq a_2 \succeq \ldots a_n$ is not just (weakly) preferred to its neighbour but to all other alternatives in the chain. This is reflected in $\bigcap \mathbb{R}$, which represents the ordering over $a_1, \ldots, a_n$: $\bigcap \mathbb{R}$ not only includes tuples $\langle a_1, a_2\rangle$,$\langle a_3, a_4\rangle$, etc., but also $\langle a_1, a_3\rangle$,$\langle a_1, a_4\rangle$, etc. Contraction of a preference ordering by $a_1 \succeq a_n$ then necessitates the removal of at least one further 'link' of the chain. That is, if $\langle a_1, a_n\rangle$, is removed from $\bigcap \mathbb{R}$ then so at least one $\langle a_i, a_{i+1}\rangle, i < n$, has to be excluded as well. This extra exclusion is automatically done by $C(\cdot)$, for there is no total ordering over $a_1, \ldots, a_n$ for which $a_1 \not\succ a_n$ if not at least for one $i < n : a_i \not\succ a_{i+1}$. The problem is that all total orders which represent $a_1 \not\succ a_n$ and $a_i \not\succ a_{i+1}$ for *exactly* one $i < n$ are equally similar to $\bigcap \mathbb{R}$ and hence are all included in $C(\cdot)$. If $\mathbb{R}_{\cup\mathbb{I}}$ was constructed on the basis of $C(\cdot)$, it would eliminate *all* members of the chain between a_1 and a_n, when contracting by $a_1 \succeq a_n$. This again is too radical a step to built it as a necessary mechanism into contraction.

Instead, the contraction in such a case can be weakened if extralogical entrenchment information is available. An entrenchment relation is a partial order of preference orderings. I have discussed the basis of entrenchment in section 4.2.2. The entrenchment relation orders preferences according to how good a reason an agent has to hold them. Eventually I hope to develop a more precise measure of entrenchment on the basis of the criteria given in section 4.2.2: unanimity of the relevant aspectual preferences, number of aspectual preferences that support it and saliency of the underlying partitions. At this point, however, I cannot provide a formalised account of it. Instead I will simply assume that there is such a partial entrenchment order available.

An entrenchment relation $\succ_{\mathbb{E}}$ is defined over subsets $\mathbb{E}_i$ of $\mathbb{P}_{A\times A}$ such that each $\bigcap \mathbb{E}_i$ represents a single preference comparison and $\mathbb{E}_i$ is maximal. $\succ_{\mathbb{E}}$ is irreflexive and transitive, but not necessarily complete. The entrenchment-regarding selection function $C_{\succ_{\mathbb{E}}}(\cdot)$ is then defined as:

Definition 11 $C_{\succ_{\mathbb{E}}}(\mathbb{P}_{A\times A} \setminus \mathbb{I}) = C(\mathbb{P}_{A\times A} \setminus \mathbb{I}) \cap \mathbb{E}_i$ *for those* $\mathbb{E}_i$ *for which* $\bigcap \mathbb{E}_i \subseteq \bigcap C(\mathbb{P}_{A\times A} \setminus \mathbb{I})$ *and not* $\mathbb{E}_i \succ_{\mathbb{E}} \mathbb{E}_j$ *for any* j

The selection function is thus refined by the entrenchment relation: those total preference orders that satisfy the similarity criterion are chosen by $C_{\succ_{\mathbb{E}}}$ which either do not appear in the entrenchment ranking at all, or which rank lowest in that ranking.

With the help of the entrenchment-regarding selection function, the contraction operator $\cdot_{\cup}$ is defined as:

Definition 12 $\mathbb{R}_{\cup\mathbb{I}} = C_{\succ_{\mathbb{E}}}(\mathbb{P}_{A\times A} \setminus \mathbb{I}) \cup \mathbb{R}$

The definition of contraction fulfills the following postulates.

Theorem 4 *1.* $\bigcap \mathbb{R}_{\cup\mathbb{I}}$ *is a preference ordering (Closure)*

2. $\bigcap \mathbb{I} \not\subseteq \bigcap \mathbb{R}_{\cup\mathbb{I}}$ *(Success)*

3. $\bigcap \mathbb{R}_{\cup\mathbb{I}} \subseteq \bigcap \mathbb{R}$ *(Inclusion)*

4. If $\bigcap \mathbb{I} \cap \bigcap \mathbb{R} = \emptyset$ *then* $\bigcap \mathbb{R}_{\cup\mathbb{I}} = \bigcap \mathbb{R}$ *(Vacuity)*

5. If $\mathbb{I} = \mathbb{J}$*, then* $\bigcap \mathbb{R}_{\cup\mathbb{I}} = \bigcap \mathbb{R}_{cup\mathbb{I}}$ *(Extensionality)*

Proof *Part 1, Closure:* It needs to be shown that $\bigcap \mathbb{R}_{\cup\mathbb{I}}$ is reflexive and transitive. Proof proceeds analogous to Corollary 3.

Part 2, Success: For all $X : X \in \mathbb{P}_{A\times A} \setminus \mathbb{I} \Leftrightarrow X \notin \mathbb{I}$. As $C_{\succ_{\mathbb{E}}}$ by definitions 10 and 11 is not empty, it follows that for all $X : X \in C_{\succ_{\mathbb{E}}}(\mathbb{P}_{A\times A} \setminus \mathbb{I}) \Rightarrow X \notin \mathbb{I}$. Hence for some $X : X \in (C_{\succ_{\mathbb{E}}}(\mathbb{P}_{A\times A} \setminus \mathbb{I}) \cup \mathbb{R}) \Rightarrow X \notin \mathbb{I}$, and thus $(C_{\succ_{\mathbb{E}}}(\mathbb{P}_{A\times A} \setminus \mathbb{I}) \cup \mathbb{R}) \not\subseteq \mathbb{I}$. Then by corollary 2 and maximality $\bigcap \mathbb{I} \not\subseteq \bigcap (C_{\succ_{\mathbb{E}}}(\mathbb{P}_{A\times A} \setminus \mathbb{I}) \cup \mathbb{R})$ and thus by definition 12 $\bigcap \mathbb{I} \not\subseteq \bigcap \mathbb{R}_{\cup\mathbb{I}}$.

Part 3, Inclusion: For all $\mathbb{X} : \mathbb{R} \subseteq \mathbb{R} \cup \mathbb{X}$. Then by corollary 1 for all $\mathbb{X} : \bigcap(\mathbb{R} \cup \mathbb{X}) \subseteq \bigcap \mathbb{R}$, and in particular $\bigcap(\mathbb{R} \cup C_{\succ_{\mathbb{E}}}(\mathbb{P}_{A\times A} \setminus \mathbb{I})) \subseteq \bigcap \mathbb{R}$. From this and definition 12 follows that $\bigcap \mathbb{R}_{\cap\mathbb{I}} \subseteq \bigcap \mathbb{R}$.

Part 4, Vacuity: If $\bigcap \mathbb{I} \not\subseteq \bigcap \mathbb{R}$ then by corollary 1 $\mathbb{R} \not\subseteq \mathbb{I}$. This is equivalent to the claim that there is a non-empty set $\mathbb{U} : \mathbb{U} \subseteq \mathbb{R} \quad \& \quad \mathbb{U} \subseteq \mathbb{P}_{A\times A} \setminus \mathbb{I}$. Then there is one such $\mathbb{U}$ that is the largest subset which $\mathbb{R}$ and $\mathbb{P}_{A\times A} \setminus \mathbb{I}$ have in common. The selection function $C_{\succ_{\mathbb{E}}}$ will pick out only those orders which are in that $\mathbb{U}$, because they all maximise the similarity with $\mathbb{R}$: $\mathbb{U} \subseteq \mathbb{R} \quad \& \quad \mathbb{U} \subseteq \mathbb{P}_{A\times A} \setminus \mathbb{I} \quad \& \quad (\mathbb{P}_{A\times A} \setminus \mathbb{I} \setminus \mathbb{U}) \cap \mathbb{R} = \emptyset \Leftrightarrow C_{\succ_{\mathbb{E}}} \subseteq \mathbb{U}$. But if $C_{\succ_{\mathbb{E}}} \subseteq \mathbb{U} \subseteq$

$\mathbb{R}$, then $\bigcap(C_{\succ_{\mathbb{E}}}(\mathbb{P}_{A\times A} \setminus \mathbb{I}) \cup \mathbb{R}) = \bigcap \mathbb{R}$, and hence by definition 12 $\bigcap \mathbb{R}_{\cup \mathbb{I}} = \bigcap \mathbb{R}$.

Part 5, Extensionality: Analogous to proof of theorem 1.6.□[10]

Revision *Revision* changes a preference existing in the order into a new preference. Such a move is different from expansion only if the input preference cannot be accommodated without changing the original preference ordering. The revision operator $\cdot_\star$ replaces the members of $\mathbb{R}$ with the members of $\mathbb{I}$ which are closest to $\bigcap \mathbb{R}$. For this similarity measure, the entrenchment-regarding selection function is employed that was defined in definition 11 Unlike standard accounts in belief revision, the revision operator here is not defined in terms of contraction and expansion.

Definition 13 $\mathbb{R}_{\star\mathbb{I}} = \begin{cases} C_{\succ_{\mathbb{E}}}(\mathbb{I}) & iff \quad \mathbb{R} \cap \mathbb{I} = \emptyset \\ \mathbb{R}_{\cap \mathbb{I}} & iff \quad \mathbb{R} \cap \mathbb{I} \neq \emptyset \end{cases}$

The revision operators satisfies the following postulates.

Theorem 5 *1.* $\bigcap \mathbb{R}_{\star\mathbb{I}}$ *is a preference ordering (Closure)*

2. $\bigcap \mathbb{I} \subseteq \bigcap \mathbb{R}_{\star\mathbb{I}}$ *(Success)*

3. If $\mathbb{R} \cap \mathbb{I} \neq \emptyset$ *then* $\bigcap \mathbb{R}_{\star\mathbb{I}} = \bigcap \mathbb{R}_{\cap\mathbb{I}}$ *(Vacuity)*

4. If $\bigcap \mathbb{I} = \bigcap \mathbb{J}$ *then* $\bigcap \mathbb{R}_{\star\mathbb{I}} = \bigcap \mathbb{R}_{\star\mathbb{J}}$ *(Extensionality)*

Proof *Part 1, Closure:* As in corollary 3.

Part 2, Success: $C_{\succ_{\mathbb{E}}}$ selects at least one member of $\mathbb{I}$ according to definition 13. Thus $\mathbb{R}_{\star\mathbb{I}} \subseteq \mathbb{I}$ and then by corollary 1 $\bigcap \mathbb{I} \subseteq \bigcap \mathbb{R}_{\star\mathbb{I}}$.

Part 3, Vacuity: By definition 13.

Part 5, Extensionality: Analogous to 1.6.

4.5 Comparison

While a broad discussion of belief change exists since the works of Quine (1970) and Levi (1974), the related topic of preference change

[10] Contraction does not satisfy monotonicity. Counterexample: An agent holds preferences $a \succ b, a \succ c$ and hypothetically also holds $a \succ b, b \succ c, a \succ c$. Then a when he contracts his actual preferences by $b \succ c$, according to 2.4 (vacuity) he retains his original preferences. But if he contracts his hypothetical preferences by $b \succ c$, he will either be left with $a \succ b$ or $a \succ c$, depending on the entrenchment of his preferences. Hence in both cases the contracted hypothetical preference ordering, which originally contained his actual preferences, is now smaller than his contracted actual preferences.

remains largely unexplored. The only exception from this are the writings of Sven Ove Hansson. In this section, I will compare his model (1995, 2001) with the one I have proposed in this chapter.[11]

Hansson's model, I will argue, differs in five central aspects from the model proposed here. It represents an agent's motivation as a sentential *preference set*, which is closed under propositional logic; while my model represents them as a preference base. Hansson further models change operators on the basis of a *priority index*, a device that does not capture as many aspects of preference change as the entrenchment ordering does, but which is computationally far more cumbersome. Thirdly, it will turn out that the sentential model of contraction has a problem to handle *multiple contraction* that the model proposed here does not have. Fourth, due to the closure of the preference set, Hansson's contraction operator satisfies the *recovery postulate*, which it should not, as I will argue. Last, Hansson's model has an advantage over mine, in that it can represent preference sentence disjuncts without having to validate one of the disjuncts. This allows to represent e.g. 'A is at least as preferred as B', without necessarily specifying this either to 'A is strictly preferred to B' or 'A is indifferent to B'.

Hansson represents an agent's motivation as a mix of relational and sentential set. On the basic level, he defines a set R, whose members are reflexive preference relations over a common domain of alternatives $\mathcal{U}$. For example,

$$\mathbf{R} = \{R_1, R_2\} = \{\{\langle A, A\rangle, \langle B, B\rangle, \langle A, B\rangle\}, \{\langle A, A\rangle, \langle B, B\rangle, \langle A, B\rangle, \langle B, A\rangle\}\}$$

with $\mathcal{U} = \{A, B\}$. In contrast to my model, the underlying preference relations $R \in \mathbf{R}$ are not complete; they are therefore not as compatible with a utility representation as my account.

Each preference relation R_i in $\mathbf{R}$ then defines a set of preference sentences $[R_i]$. That set consists of sentences corresponding to each preference comparison in R_i, and sentences from a theory T. T is formulated in the preference language itself, and consists of sentences restricting any preference sentence set, e.g. $T = \{(X \succ$

[11] It could be argued that Nayak et al. (1996)'s discussion of a model of change of belief entrenchment is relevant here too. The relation, however, is largely formal, and the model not sufficiently developed to merit a separate discussion here.

$Y) \wedge (Y \succ Z) \rightarrow (X \wedge Z)\}$. On top of that, $[R_i]$ is truth-functionally closed under propositional logic. For example, the preference relation $R_1 = \{\langle A, A\rangle, \langle B, B\rangle, \langle A, B\rangle\}$ has as its sentential representation the set $[R_i] = \{A \succeq A, A \succeq B, B \succeq B, \neg(B \succeq A), (A \succeq B) \vee (B \succeq A), (A \succeq B) \wedge \neg(B \succeq A), (B \succeq A) \rightarrow (A \succeq B) \ldots\}$.

The agent's motivations are then represented as the intersections of all of those sentential representations of the preference relations in $\mathbf{R}$: $[\mathbf{R}] = \bigcap\{[R] | R \in \mathbf{R}\}$. While this framework looks like a sentential representation, Hansson uses $[\mathbf{R}]$ only as a semantic output device; the real work – in particular the change operators – is done with the relational preference model $\mathbf{R}$.[12]

Hansson starts the construction of the change operators with the revision operator and develops the contraction operator out of that. To construct the revision operator, he develops the notion of a priority index assigned to the input sentence of the change operator. When a preference set is to be revised, say, by a sentence $A \succeq B$, the index identifies whether the set $\mathbf{R}$ should be revised by changing the position of A or by changing the position of B. I think the intuitive idea behind this priority index is that all preference changes are determined by a change in the underlying reasons for *desiring* an option. Take for example Mr. Myers' ordering of four newspapers, A to D, in the following way:

$$A \succ B \succ C \succ D$$

an order expressed by preference sentence set ♯1 (see below). Now Mr. Myers finds that newspaper B hired a right-wing commentator, who annoys him greatly, and thus he now prefers paper D to paper B. His new ordering can take two different forms. According to Hansson, he either changes it by moving D up in the ranking, obtaining:

$$A \succ D \succ B \succ C$$

an order expressed by preference sentence set ♯2. Or he changes it by moving B down:

$$A \succ C \succ D \succ B$$

[12]To increase the confusion even more, he then uses sentential inputs – such that the operators are defined as accommodating a sentential input to a relational set. Hansson briefly discusses this problem and concludes: '[a purely relational treatment] seems to be less directly related to intuitive notions of preference change. Therefore, single sentence input, and the notation $\mathbf{R}^*_B\alpha$ [for the revision operator] will be used here' (Hansson 2001, 48).

an order expressed by preference sentence set ♯3. For reasons of clarity, I have put all three sentential representation (or rather their salient parts) in one place:

$$\sharp 1 : \{A \succ B, A \succ C, A \succ D, B \succ C, B \succ D, C \succ D, \ldots\}$$
$$\sharp 2 : \{A \succ B, A \succ C, A \succ D, B \succ C, D \succ B, D \succ C, \ldots\}$$
$$\sharp 3 : \{A \succ B, A \succ C, A \succ D, C \succ B, D \succ B, C \succ D, \ldots\}$$

Note that the difference between ♯1 and ♯2 and between ♯1 and ♯3 is two binary preferences each: $D \succ B$, of course, and then one further preference each. What differs between Hansson's priority index and the entrenchment relation is the justification of determining this second binary preference.

Hansson claims that the information that determines which of the binary preferences to change comes attached to the input sentence:

> . . . the primary input that we wish to mirror in the formal model provides us with information about which of these to choose: You get tired of brand A and start to like it less than brand B, which was your previous second choice. You learn that the political party X has changed its policies on unemployment insurance and start to like it more than party Y, and so on. (Hansson 2001, 47)

I interpret Hansson to say here that the reason that makes one change the preference also determines *how*, in the sense clarified above, one should change it. If Mr. Myers changes his preference between B and D, because of a change in the properties of B, then that reason also makes him drop B in the ranking, and keep D where it is. Thus the priority index, according to Hansson, should determine B as the one whose position is changed, resulting in the ordering ♯3.

I think this framework is too narrow. Remember the notion of entrenchment I developed in section 4.2.2. Preferences over exclusionary states have their reasons in the preferences over the aspects that these states realise. For example, Mr. Myers prefers newspaper A to B because it has better book reviews; B to C because it reports more international news and C to D because it is more amusing to read. The aspects identified for each newspaper *depend on the comparison*: when compared to C, B's salient feature is its

amusement value; when compared to A, B's salient feature is its lack of good book reviews. When a new aspect is discovered or a assumed aspect found wanting, and therefore a particular exclusionary preference changed, this does *not* mean that this aspectual change is *decisive* for all other preference comparisons. But this is exactly what Hansson claims: *because* B has the negative aspect of a right-wing commentator, which makes Mr. Myers prefer D to it now, it *also* has to drop below C. In contrast to this, my model asks: which of the preferences are more deeply entrenched? And it makes the answer dependent on the number of aspectual preferences, on their unanimity, and on the importance of the partition according to which the two aspect-realizing states are compared. Under that framework, Mr. Myers might well decide that it is not so important for the newspaper he reads to be amusing as it is for it to report international news; and consequently change the preference between D and C rather than B and C – resulting in the ordering $\sharp 2$.

Given the priority index and a similarity measure (which is complicated through the incorporation of the index), Hansson constructs the revision operator $\mathbf{R}^{*}_{\mathfrak{B}}\alpha$ as changing *all* elements R^{*} of $\mathbf{R}$ in such a way that (i) they all include α; (ii) they all are close to some $R \in \mathbf{R}$; and (iii) they all are consistent. Comparing it with the one of my model, Hansson's operator is computationally more extensive: every member of $\mathbf{R}$ has to be changed to include α, and they all have to be re-shaped as closely as possible to some member of $\mathbf{R}$. In contrast to this, my model only requires that those $R \in \mathbb{P}_{A \times A}$ which validate α and are closest to $\bigcap \mathbb{R}$ are included in $\mathbb{R}$; and consecutively all those $R \in \mathbb{R}$ which do not validate α are excluded from $\mathbb{R}$.

Hansson next constructs the contraction operator on the basis of revision. To contract α from $\mathbf{R}$, Hansson forms the union of the set $\mathbf{R}$ and the set $\mathbf{R}^{*}_{\mathfrak{B}}\neg\alpha$. Thus, Hansson's contraction operator is narrowed down by the priority index, against which I have argued above. Further, the change operators take their input to be a single sentence, not a set of sentences. While in the case of revision, Hansson showed that a revision by a set of sentences can be modelled equivalently as a revision by a conjunction of all the set's members, this cannot be done for contraction.

> There is no truth-functional combination f of two sentences such that it holds in general that $\{\alpha_1, \alpha_2\} \cap [\mathbf{R} \div_{\mathfrak{B}}$

$\{\alpha_1, \alpha_2\}] = \emptyset$ if and only if $f(\alpha_1, \alpha_2) \notin [\mathbf{R} \div_{\mathfrak{B}} \{\alpha_1, \alpha_2\}$. (2001, p. 52)

In contrast, my model does not define the input as single sentences and hence is not in need for a truth-functional operator that would allow one to convert a set of sentences into a single sentence. It can therefore handle multiple contraction, while Hansson's model cannot.

Hansson's contraction operator further satisfies the recovery postulate. This is a postulate that it should *not* satisfy, however. The recovery postulate states that any preference order contracted by some preferences and then expanded by the same preferences always restores the original set. The controversial character of the recovery postulate is revealed in the following example. An agent prefers A over B and B over C. Hence by transitivity, she prefers A over C. Now she drops her preference $A > C$. In order to comply with transitivity, at least one of the other two preferences has to be removed from her overall evaluation (and it might well be possible, for lack of a specifying criterion, that she removes both). In any of the three resulting versions, a subsequent revision by $A > C$ will not restore the original preference model.

Original preference model:	$\{A > B, B > C, A > C\}$		
Contraction by $A > C$:	(i) $\{A > B\}$	(ii) $\{B > C\}$	(iii) $\emptyset$
Expansion by $A > C$:	(i) $\{A > B, A > C\}$	(ii) $\{A > C, B > C\}$	(iii) $\{A > C\}$

Models of preference change should allow for such cases, as they play an important role in preference dynamics. The recovery postulate is therefore overly restrictive. My model does not satisfy it, and I think this is right.

These advantages, I think, weigh heavily enough against the capacity of Hansson's model – due to the closure property of the preference sentence sets – to represent preference sentence disjuncts without having to validate one of the disjuncts. Overall, I therefore think the model presented here is a more adequate model of preference change than Hansson's.

but neither is much good

Chapter 5

Conclusion

In this thesis, I have argued for a particular understanding of the notion of preference used in the social sciences. The scientific context in which I have investigated this notion dictated two premises: first, that the concept of preferences must help in the scientific practices of explanation and prediction, and second, that it must adhere to certain standards of empirical adequancy.

Because the social sciences predict and explain, any notion of preference employed by them must live up to the standards of these practices. The most fundamental of these standards is that the *explanans* be a cause of the *explanandum*. I therefore argued in chapter 1 that preferences are necessary components of a sufficient cause of a particular behaviour. Beyond that, they are particular kinds of causes. They are properties, and thus their instantiations are facts, not events. As such, they can be ascribed to a behaving system, and particularly to human agents.

Further, by causing an agent to behave, preferences cause physical properties to change. I have argued that this relation is unproblematic once one agrees that preferences are programming properties: that their specific realisations are physical property realisations. However, because preferences can be realised in multiple ways, they cannot be reduced to their realisers. Instead, preferences group together under their 'realisation span' those physical properties which are necessary components of a sufficient cause of a particular behaviour; this crucial role of theirs cannot be performed by any lower-level properties. The social sciences that employ the preference concept are thus not just a transitory science that awaits

its end by successful concept reduction; rather, they are autonomous disciplines with a separate conceptual basis.

This conceptual basis, however, is ontologically compatible with those of the natural sciences. I have argued for this claim by refuting the anti-naturalist argument that preferences – or other motivational properties – are different from properties employed in the natural sciences *qua* their intentional nature. To the contrary, I have argued that preferences are not intentional properties in the sense that they essentially incorporate a semantic component. The semantic component is only a construct: a construct of the process of measuring mental properties. It is not essential for preferences, nor for any other mental states. With this argument, I have rejected two popular but incompatible positions. First, I have rejected those who claim that the semantic component is essential for mental properties, and that mental properties are compatible with natural science properties because wide semantic content still supervenes on physical properties. Second, I have rejected those who claim that the semantic component is essential for mental properties, and that therefore the social sciences dealing with these properties have to employ practices fundamentally distinct from explanation or prediction. Instead, I have argued that the social sciences do explain on the basis of preferences (and other mental properties), that these concepts are compatible with natural science properties, and are not threatened by eliminativists ambitions.

Preferences ontologically characterized in such a way can perform as *explanans*, if their attribution is sufficiently empirically justified. In chapter 2, I have demonstrated that introspection understood as sensing one's own motivation is not a reliable method to assign preferences: it is incomplete and fallible. But the failure of this kind of introspection does not lead to the conclusion that preferences and other mental properties should be given up altogether in the social sciences. To the contrary, a methodology that attempts to eschew mental properties and establish lawlike generalizations between observable stimuli and behaviour is prone to incompleteness and infinite regress.

Instead, I have argued that preference assignment rests exclusively on behavioural data (ideally construed in a wide way to include verbal *behaviour*), and that in order to explain this data, social

scientists need to employ cognitive concepts. These cognitive concepts, however, are not the same as those of introspective psychology; rather, they are non-observable, theoretical terms, determined by a theory that sufficiently fits the data. As I showed in examples from chemistry and biology, the social sciences are in good company when using theoretical concepts whose employment cannot be justified from singular observations alone. The employment of mental properties, therefore, does not make social sciences less well founded than many natural sciences.

The nature of a theory ascribing mental properties, I have further set forth, is conjectural: one cannot hope to derive and justify it from our common sense intuitions. The 'method *a priori*', as it has been defended in economics by Mill, Robbins and v. Mises, therefore has to be rejected where it claims a solid intuitive justification of its first principles. Further, the construction of decision theory as a 'collection of platitudes' cannot be more than a metaphor; it does not suffice for an operational methodology for the decision sciences.

Instead, theorists must choose from a large set of plausible theories available. Their choice, I stated, should be ultimately determined by simplicity considerations on the one hand and best-fit criteria on the other. As I pointed out, this methodological consideration leads one to admit the ultimate indeterminateness of the true theory. However, given the data available, the use of theories more complex than those that rely on the simple belief-desire distinction is unwarranted. The choice is therefore between the number of theories that are based on the dual belief-desire distinction. Beyond simplicity and versatility considerations, their correctness is an empirical question.

Without relying on any specific theory, in the following two chapters I have pointed out two central problems for all preferences ascribed in this fashion. The first problem is the modularisation of preferences. Only the most specific preferences can be empirically justified, but only sufficiently abstract preferences are useful for explaining or predicting new situations. In chapter 3 I narrowed my focus to the simplest form of preferences: most specific descriptions of certain events. I showed how these 'simplest' preferences could be derived from choices, and investigated how a more abstract kind of preference could be derived from the specific ones while maintaining

the latter's empirical justification. This framework is based on the notion of a causal structure that represents an agent's beliefs. It presents a way to ascribe preferences on the basis of observed behaviour, and then to individuate and recombine them, explaining or predicting different behaviour under different circumstances.

The second problem is the change of preferences over time. Most theories in the social sciences have so far avoided tackling this problem and have instead relied on *ad hoc* justifications. The attempts that have been made often rely on introspection and aim at identifying the causes of preference change. But, as I argued, the employment of introspection is dubious and none of the causes of preference change are either necessary or sufficient.

I therefore proposed a different approach that starts from those cases where observed behaviour was inconsistent with the ascribed preferences. Upon such a phenomenon, the scientist should choose from a menu of possible interpretations, basing her choice on secondary evidence like potential causes of preference change. If the evidence speaks for the preference change interpretation, the change itself is modelled as the transformation the agent has to perform on her preferences given that she is committed to a small number of rationality principles. By modelling this transformation, the social scientist is able to explain and predict further changes in preferences that have not been manifested in her behaviour yet.

This model is the first step to 'dynamise' cognitive theories of behaviour. With its help, behaviour observed at different times can be incorporated into the preference ordering without making strong assumptions about the stability of preferences; and the resulting preference ordering can be employed at different times for explaining or predicting behaviour without question-begging assumptions.

In this thesis I have argued that preferences are precise and accountable concepts, even though they are attributed in the process of interpreting an agent's behaviour. Explanation and prediction of behaviour on the basis of preferences is empirically justifiable and ontologically sound; by developing the conception further in the direction of modularisation and intertemporality, the social sciences will find in preferences a rich concept that is able to do a lot of work.

5.0.1 Outlook

In the course of this discussion and the ensuing development of new models that expand the applicability of preferences, a few further problems became apparent. These problems require further research work, and I will give a brief outlook to further work I intend to do on this basis.

First, the discussion in section 2.2.1 made clear that a new method was needed to incorporate verbal reports into preference ascriptions without committing the fallacies associated with introspection. As I have indicated in that chapter, I think that we can use verbal reports as data once we realise that they are governed by similar deliberation mechanisms as other behaviour is. Not only do people decide with full awareness what to tell, what to omit or how to deceive; the experiments in section 2.2.1 show that they also decide what and how to tell on the basis of parameters they are not aware of. To incorporate verbal data into decision theory thus requires the development of a model that takes into account this deliberative mechanism; which in the end amounts to developing a unified theory of thought and action.

Second, the principle of equivalence needs to be extended to preferences over uncertain and risky prospects. For this, a probability measure needs to be introduced, and the preferences are to be represented as an expected utility index. I intend to develop such a theory of expected utility on the basis of a *Bayesian Network*; and hope to show how such an index is similar, or differs, to other accounts of causal decision theory.

Third, in section 3.4.2 I mentioned the problem surrounding the specifications of world and prospect partitions. This connects directly to one element of reasons for holding preferences, as discussed in section 4.2.2. Here I intend to discuss Broome's suggestion to include the so far extratheoretically determined relata specification into the theory more seriously, possibly backed up by the recent findings in neurophysiology.

Fourth, the concept of reasons for holding preferences was crucially determined by a weighing process of the prospect preferences involved. This was also mentioned in section 3.4.3, where the references to the social choice literature showed that on the basis of preference information only, no general aggregation procedure can be found. The question that needs to be answered, then, is how

much more information is needed for such an aggregation procedure to work, and whether one can reasonably assume that this information is available in intrapersonal deliberation processes.

Last, the discussion of the last two points leads to a better understanding of preference entrenchment. With a good account of what it means to have a reason for holding a preference, it will be possible to quantify the criteria for such a reason and construct an entrenchment pre-order over preferences on that basis.

References

Antony, L. (1989) 'Anomalous Monism and the Problem of Explanatory Force', *The Philosophical Review* 98(2): 153-187.

Ardal, P. (1989) *Passion and Value in Hume's Treatise*, Edinburgh: Edinburgh University Press, 2nd edition.

Armstrong, D. M. (1968) *A Materialist Theory of the Mind*, London: Routledge.

— (1980) 'The nature of mind'' in Block (1980), 191-199.

Aumann, R. (1962) 'Utility theory without the completeness axiom', *Econometrica* 30(3): 445-462.

Balzer, W., C.U. Moulines and J.D. Sneed (1987) *An Architectonic for Science*, Reidel: Dordrecht.

Baron-Cohen, S. (1995) *Mindblindness*, Cambridge, MA.: MIT Press.

Beckermann, A. (1996) 'Is there a problem about intentionality?', *Erkenntnis* 51: 1-23.

Bell, D.E. (1982) 'Regret in decision making under uncertainty', *Operations Research* 30: 961-81.

Bem, D.J. and H. K. McConnell (1970) 'Testing the self-perception explanation of dissonance phenomena: On the salience of premanipulation attitudes', *Journal of Personality and Social Psychology*, 14: 23-31.

Bentley, M. (1926) 'The major categories of psychology', *Psychological Review*, 33:71-105.

Block, N. (1980) (ed.) *Readings in the Philosophy of Psychology*, Cambridge, MA. and London: Harvard University Press: 37-47.

Bowles, S. (1998) 'Endogenous preferences: the cultural consequences

of markets and other economic institutions', *Journal of Economic Literature* 36: 75-111.

Bradley, R. (2001) 'Ramsey and the measurement of belief', in D. Corfiled and J. Williamson (eds) *Foundations of Bayesianism*, Kluwer Academic Publishers: 263-290.

Bratman, M. (1987) *Intention, Plans and Practical Reason*, Cambridge, MA. and London: Harvard University Press.

Brentano, F. (1973) *Psychology from an Empirical Standpoint*, edited by L.L. McAlister, translated by A.C. Rancurello et al., London: Routledge and Kegan Paul.

Broome, J. (1991) *Weighing Goods*, Oxford: Blackwell, 1991.

Carlson, E. (1997) 'The intrinsic value of non-basic states of affairs', *Philosophical Studies* 85: 95-107.

Chomsky, N. (1988) *Language and Problems of Knowledge*, Cambridge, MA.: MIT Press.

Churchland, P.M. (1979) *Scientific Realism and the Plasticity of Mind*, Cambridge: Cambridge University Press.

— (1981) 'Eliminative materialism and the propositional attitudes', *Journal of Philosophy* 78: 67-90.

Crane, T. (1990) 'An Alleged Analogy Between Numbers and Propositions', *Analysis* 50: 224-230.

Davidson, D. (1963) 'Actions, Reasons and Causes', in Davidson (1980): 3-20.

— (1967) 'Causal Relations', in Davidson (1980): 149-162.

— (1971) 'Agency', in Davidson (1980): 43-62.

— (1973) 'Radical interpretation', in Davidson (1984): 125-140.

— (1974a), 'Belief and the basis of meaning', in Davidson (1984): 141-154.

— (1974b) 'Psychology as philosophy', in Davidson (1980): 229-244

— (1975) 'Thought and talk', in Davidson (1984): 155-170.

— (1980) *Actions and Events*, Oxford and New York: Oxford University Press.

— (1984) *Truth and Interpretation*, Oxford and New York: Oxford University Press.

— (1985) 'Incoherence and irrationality', *Dialectica* 39(4): 345-354.

— (1989) 'What is present to the mind', in *Subjective, Intersubjective, Objective*, Oxford and New York: Oxford University Press.

Dennett, D. (1978) 'Skinner skinned', in *Brainstorms*, Cambridge, MA.: MIT Press.

— (1991) 'Real patterns', *Journal of Philosophy* 87: 27-51.

Descartes, R. (1985, 1984, 1991) *The Philosophical Writings of Descartes*, translated by J. Cottingham, R. Stoothoff, D. Murdoch and A. Kenny, Cambridge: Cambridge University Press, 3 volumes.

Diez, J. A. (2002) 'A program for the individuation of scientific concepts', *Synthese* 130: 13-48.

Diogenes Laertius. *Lives, Teachings, and Sayings of Famous Philosophers*, Loeb Classical Library, 1931.

Dretske, F. I. (1988). *Explaining Behaviour: Reasons in a World of Causes*, Cambridge, MA.: MIT Press

— (1993) 'Conscious experience', *Mind* 102: 263-83.

Dubra, J., Maccheroni, F., Ok, E. A. (2004) 'Expected Utility Theory without the Completeness Axiom', *Journal of Economic Theory* 115: 118-133.

Eckhardt, B. v. (1994) 'Folk psychology (1)', in Guttenplan, S. (ed.) *A Companion to the Philosophy of Mind*, Oxford: Blackwell.

— (1995) 'Philosophical conceptions of folk psychology', In *Mindscapes: Philosophy, Science and the Mind*, edited by Martin Carrier and Peter K. Machamer, Universitätsverlag Konstanz/University of Pittsburgh Press: 23-51.

Elster, J. (1982) 'Sour grapes: utilitarianism and the genesis of wants', in: A. Sen and B. Williams: *Utilitarianism and Beyond*, Oxford University Press.

— (1985) 'Weakness of will and the free-rider problem.' *Economics and Philosophy* 1: 231-265.

Emlen, S.T. (1972) 'The ontogenetic development of orientation capabilities', in: *Animal Orientation and Navigation* edited by S.R.

Galler et al., NASA: 191-210.

Ericsson, K. and Simon, H. (1993) *Protocol Analysis*, Cambridge, MA.: MIT Press.

Field, H. (1980) 'Mental representation', in N. Block (ed.) *Readings in the Philosophy of Psychology* Vol. 2, Cambridge, MA. and London: Harvard University Press.

Fodor, J. A. (1974) 'Special sciences, or the disunity of science as a working hypothesis' in Block (1980): 120-133.

— (1987) *Psychosemantics*, Cambridge, MA.: MIT Press.

Frankfurt, H. (1971) 'Freedom of the will and the concept of a person', *Journal of Philosophy* 68: 5-20.

Gärdendors, P. (1988) *Knowledge in Flux*, Cambridge, MA.: MIT Press.

Gibson, J. J. (1950) *The Perception of the Visual World*, Boston: Houghton Mifflin.

Goethals, G. R. and R. F. Reckman (1973) 'The perception of consistency in attitudes', *Journal of Experimental Social Psychology* 9: 491-501.

Goldman, A. (1992) 'In defense of the simulation theory', *Mind and Language* 7: 104-119.

Gordon, R. (1986) 'Folk psychology as simulation, *Mind and Language* 1: 158-171.

Grice, H. P. (1989) 'The William James lectures', in *Studies in the Way of Words*, Cambridge, MA. and London: Harvard University Press.

Gruene, T. and E. F. McClennen (forthcoming) 'Hume's concept of the passions as the basis of his economic thought', in: M. Schabas and C. Wennerlind (eds) *Hume's Political Economy*, London and New York: Routledge.

Halldén, S. (1957) *On the Logic of Better*, Lund: Library of Theoria.

Halldin, C. (1986) 'Preference and the cost of preferential choice.' *Theory and Decision* 21: 35-63.

Hansson, S. O. (1989) 'A new semantical approach to the logic of preference.' *Erkenntnis* 31: 1-42.

— (1995) 'Changes in preference', *Theory and Decision* 38: 1-28.

— (1999) *A Textbook of Belief Dynamics*, Kluwer, Dordrecht.

— (2001) *The Structure of Values and Norms*, Cambridge: Cambridge University Press.

Hargreaves Heap, S., M. Hollis, B. Lyons, R. Sugden and A. Weale (1992) *The Theory of Choice*, Oxford and Cambridge, MA: Blackwell.

Harman, G. (1967) 'Towards a theory of intrinsic value.' *Journal of Philosophy* 64: 792-804.

— (1986) *Change in View: Principles of Reasoning*, Cambridge, MA.: MIT Press.

Hornsby, J. (1993) 'Agency and causal explanation' in: *Mental Causation*, Heil and Mele (eds), Oxford and New York: Oxford University Press.

Hume, D. (2000) *A Treatise of Human Nature* Edited by D. Fate Norton and M. J. Norton, Oxford and New York: Oxford University Press.

Jackson, F. and P. Pettit (1990a) 'Program explanation: a general perspective, *Analysis* 50: 107-117.

— (1990b) 'In defense of folk psychology', *Philosophical Studies* 59: 31-54.

James, W. (1890) *The Principles of Psychology*, New York: Dover, 1950.

Jeffrey, R. (1974) 'Preferences among preferences', *Journal of Philosophy* 71: 377-91.

— (1983) *The Logic of Decision*, Chicago: Chicago University Press.

Joyce, J. M. (1999) *The Foundations of Causal Decision Theory.* Cambridge: Cambridge University Press.

Kant, I. (1929): *The Critique of Pure Reason*, translated by Norman Kemp Smith, London : Macmillan.

Kim, J. (1973) 'Causation, Nomic Subsumption, and the Concept of Event', *Journal of Philosophy* 70: 217-36.

— (1993) 'Events as Property Exemplifications', in *Supervenience*

and Mind, Cambridge: Cambridge University Press.

— (1993): 'Multiple Realization and the Metaphysics of Reduction', in in *Supervenience and Mind*, Cambridge: Cambridge University Press.

— (1998) *Philosophy of Mind*, Bolder, CO: Westview Press.

Kitcher, P. (1984) '1953 and all that: a tale of two sciences', *The Philosophical Review* 93: 335-73.

Landsburg, S.E. (1981) 'Changes in taste: UK 1900-1955', *The Journal of Political Economy*, 89(1): 92-104.

Levi, I. (1974) 'On indeterminate probabilities', *Journal of Philosophy* 71: 391-418.

— (1985) *Hard Choices*, Cambridge University Press.

Lewis, D. (1970) 'How to define theoretical terms', *Journal of Philosophy* 81: 89-106.

— (1972) 'Psychophysicalism and Theoretical Identifications', in Block (1980): 207-215.

— (1981) 'Causal Decision Theory.' *Australasian Journal of Philosophy* 59: 5-30.

Liebermann, D. A. (1979) 'Behaviourism and the mind: A (limited) call for a return to introspection', *American Psychologist* 34: 319-333.

Locke, J. (1975) *An Essay Concerning Human Understanding*, ed. P.H. Nidditch, Oxford and New York: Oxford University Press.

Loomes G. and Sugden R. (1982) 'Regret theory: an alternative theory of rational choice under uncertainty', *Economic Journal* 92: 805-824.

Mackie, J. L. (1980) *The Cement of the Universe. A Study of Causation*, Oxford and New York: Oxford University Press

Mas-Colell, A., Winston and Green (1995) *Microeconomic Principles*, Oxford and New York: Oxford University Press.

Matthews, R. (1994) 'The Measure of Mind', Mind 103: 131-146.

McGinn, C. (1982) *The Character of Mind*, Oxford and New York: Oxford University Press.

Mellor D.H. (1987) 'The singularly affecting facts of causation', in P. Pettit et al. (eds) *Metaphysics and Morality*, Oxford: Blackwell.

Mill, J. S. (1949) *A System of Logic*, London: Longman, Green and Co.

Moser, P. (1996) 'Physicalism and mental causes: contra Papineau'. *Analysis* 56/4: 263-67.

Nagel, T. (1976) 'Moral Luck', in *Mortal Questions*, Cambridge: Cambridge University Press.

Nayak, A. C., P. Nelson and H. Polansky (1996) 'Belief Change as Change in Epistemic Entrenchment', *Synthese* 109: 143-174.

Nelson, Douglas A. and Peter Marler. 1994. 'Selection-based learning in bird song development.' *Proceedings of the National Academy of Science* 91: 10498-10501.

Nisbett, R. E. and S. Schachter (1966) 'Cognitive manipulation of pain' *Journal of Experimental Social Psychology* 2: 227-236.

Nisbett, R. and Wilson, T. (1977) 'Telling more than we can know: verbal reports on mental processes', *Psychological Review* 84 (3): 231-59.

Otis, A. S. (1920) 'Do we think in words? Behaviourist vs. introspective conceptions', *Psychological Review* 27: 399-419.

Papineau, D. (1993) *Philosophical Naturalism*, Oxford: Blackwell.

Pearl, J. *Causality*, Cambridge: Cambridge University Press, 2000.

Pennington, L.A. and J. L. Finan (1940) 'Operational usage in psychology', *Psychological Review*, 47: 254-266.

Pettit, P. (1991) 'Decision theory and folk psychology', reprinted in *Rules, Reasons and Norms*, Oxford and New York: Oxford University Press: 192-221.

— (2002) 'Three aspects of rational explanation', in *Rules, Reasons and Norms*, Oxford and New York: Oxford University Press: 177-191.

Platt, M.L. and P.W. Glimcher (1999) 'Neural correlates of decision variables in parietal cortex.' *Nature* 400: 233-238.

Putnam, H. (1967) 'The Nature of Mental States', in *Mind, Language and Reality. Philosophical Papers*, Volume 2, Cambridge:

Cambridge University Press.

Quine, W.V. and J.S. Ullian (1970) *The Web of Belief*, New York: Random House.

Quinn, W. S. 'Theories of intrinsic value.' *American Philosophical Quarterly* 11(1974): 123-132.

Rachlin, H. (1977) 'A review of M. J. Mahoney's *Cognition and Behaviour Modification*', *Journal of Applied Behavior Analysis* 10: 369-374.

Ramsey, F. P. (1926) 'Truth and Probability', in (1990) *Philosophical Papers* edited by D.H. Mellor, Cambridge: Cambridge University Press: 52-109.

— (1927): 'Facts and Propositions', in (1990) *Philosophical Papers* edited by D.H. Mellor, Cambridge: Cambridge University Press: 34-51.

— (1929) 'Theories', in (1990) *Philosophical Papers* edited by Mellor, Cambridge: Cambridge University Press: 112-136.

Rescher, N. (1967) 'Semantic Foundations for the Logic of Preference.' In *The Logic of Decision and Action*, edited by Nicholas Rescher, 37-62. Pittsburgh: University of Pittsburgh Press.

Richter, M. K. (1966) 'Revealed Preference Theory', *Econometrica* 34: 635-645.

Rizvi, S.A.T. (2001) 'Preference Formation and the Axioms of Choice' *Review of Political Economy*, 13: 141-159.

Robbins, L. (1935) *An Essay on the Nature and Significance of Economic Science*, 2nd ed. London: Macmillan.

Rosenberg S. and A. Sedlak (1972) 'Structural representatives of perceived personality trait relationships,' in A. Romney, R. Shepard and S. Nerlove (eds) *Multidimensional Scaling: Theory and Applications in the Behavioural Sciences*, New York: Seminar Press, 134-162.

Savage, L. J (1972) *The Foundations of Statistics.* New York: Dover.

Searle, J. (1995) *The Construction of Social Reality*, Penguin Press.

Seidenfeld, T., Schervish, M.J., Kadane, J.B. (1995) 'A Representation of Partially Ordered Preferences', *The Annals of Statistics*

23(6): 2168-2217.

Sellars, W. (1956) 'Empiricism and the Philosophy of Mind', in H. Feigl and M . Scriven (eds) *Minnesota Studies in the Philosophy of Science* Vol 1, Minneapolis: University of Minnesota Press: 253-329.

Sen, A. (1982) 'Rational Fools.' In *Choice, Welfare and Measurement*, Oxford: Blackwell: 84-108.

Shoemaker, S. (1984) 'Some Varieties of Functionalism' in *Identity, Cause and Mind*, Oxford and New York: Oxford University Press.

Skinner, B. F. (1953) Selections from *Science and Human Behaviour* in Block (1980): 37-47.

— (1971) *Beyond Freedom and Dignity*, New York: Knopf.

— (1990) 'Can Psychology be a Science of Mind?', *American Psychologist* 45: 1206-10.

Sobel, Howard (1994) *Taking Chances: Essays on Rational Choice.* Cambridge: Cambridge University Press.

— (1997) 'Cyclical Preferences and World Bayesianism.' *Philosophy of Science* 64: 42-73.

Spohn, W. (2002) 'Dependency Equilibria and the Causal Structure of Decision and Game Situations.' Mimeo, University of Konstanz.

Steedman, I. and U. Krause (1986) 'Goethe's Faust, Arrow's Possibility Theorem and the Individual Decision Maker.' In *The Multiple Self: Studies in Rationality and Social Change*, edited by Jon Elster (ed.), 197-231. Cambridge: Cambridge University Press.

Steward, H. (1996) 'Papineau's Physicalism' *Philosophy and Phenomenological Research* 56(3):667-672.

— (1997) *The Ontology of Mind. Events, Processes, and States*, Clarendon Press, Oxford.

Stich, S. and S. Nichols (1992) 'Second thoughts on simulation', in M. Davies and T. Stone (eds) *Mental Stimulation: Philosophical and Psychological* Essays, Oxford: Basil Blackwell.

Suppes, P. and J. L. Zinnes (1967) 'Basic measurement theory' in Luce et al. *Handbook of Mathematical Psychology*, New York and London: John Wiley and Sons.

Trapp, R.W. (1985) 'Utility Theory and Preference Logic.' *Erken-*

ntnis 22: 301-339.

Ullmann, L. P. (1980) 'Against direct perception', *The Behavioural and Brain Sciences* 3: 373-415.

Varian, H. (1992) *Microeconomic Analysis*, New York and London: Norton.

Williams, B.A.O. (1981) *Moral Luck*, Cambridge: Cambridge University Press.

Worcester, Robert M. (1986) *Consumer Market Research Handbook*, Amsterdam: North-Holland, 3rd edition.

Wooldridge (1963) *The Machinery of the Brain*, New York: McGraw Hill.

v. Wright, G.H. (1963) *The Logic of Preference*, Edinburgh: Edinburgh University Press.

— (1972) 'The Logic of Preference Reconsidered' *Theory and Decision* 3: 140-169.

Zuriff, G. E. (1976) 'Stimulus equivalence, grammar, and internal structure', *Behaviourism*, 4:43-52.

— (1985) *Behaviorism: A Conceptual Reconstruction*, New York: Columbia University Press.

PLSM.

```
 B20819158              Last updated: 09-03-05 Created: 09-03-05 Revision: 1
LANG: eng          CAT DATE:    -  -    MAT TYPE: y         COUNTRY: enk
SKIP: 0            BIB LVL: m          BCODE3: -           MARCTYPE:
LOCATION: ull
008           s2004    enk||||f ma    00| 0|eng|dntm a
040     UkLU|beng|cUkLU
100  1  Gruene, Till.
245  10 Rational causes :|bthe concept of preference in the social sciences.
300     173 leaves :|bill.
502     Thesis(PhD)--University of London, 2004.
691   1 Philosophical Studies (Subject Panel)
907     ag
981     |bT
```

```
 I21473092              Last updated: 09-03-05 Created: 09-03-05 Revision: 1
COPY #: 0          PATRON#: 0          RECAL DATE:    -  - INTL USE : 0
ICODE1: 0          LPATRON: 0          TOT CHKOUT: 0       COPY USE: 0
ICODE2: -          LCHKIN:    -  -     TOT RENEW: 0        IMESSAGE:
I TYPE: 5          # RENEWALS: 0       LOCATION: usthe     OPACMSG:
PRICE: #0.00       # OVERDUE: 0        LOANRULE: 0         YTDCIRC: 0
OUT DATE:    -  -  ODUE DATE:    -  -  STATUS: o           LYCIRC: 0
DUE DATE:    -  -  IUSE3: 0
098     PhD 2004 LSE
        1916861725
```