Belief & Desire

The Standard Model of Intentional Action — Critique and Defence

Petersson, Björn

2000

Document Version:
Publisher's PDF, also known as Version of record

Belief & Desire

The Standard Model of Intentional Action — Critique and Defence

ISSN 1100-4290
ISBN 91-628-4018-5

Björn Petersson

STUDIES IN PHILOSOPHY 9
Editors: Bengt Hansson and Wlodek Rabinowicz

LUND
UNIVERSITY

For my daughters

Sigrid, Tuva, Yrsa

Table of Contents

Introduction

When I regard other people as agents, i.e. as beings exercising their capacity to act intentionally, I think of them as being driven by their inner states. These internal elements are, I suppose, directed towards the realisation of some ultimate goal or, at least, towards a small part of such a goal or some means for the realisation of it. In other words, I view their actions as the product of a hierarchy or network of driving forces. Intentional actions are caused by such internal powers and at the same time rationalised, at least in a thin sense, by them. "Wants" and "desires" are common terms for pro-attitudes of the kind I have in mind. Since agents' goals are set by their desires, the role of beliefs must be subordinate; Beliefs depict, passively, the world. They provide the map enabling agents to reach destinations pre-set by their desires. Beliefs, like maps, should be modified to fit the world, while desires are satisfied when the world is changed in accordance with them. Desire's direction of fit is world-to-mind instead of mind-to-world.

So, my way of thinking about agents seems to be a straightforward application of *the belief-desire model* (henceforth abbreviated: The BD model). This means that I am prepared to explain, predict and rationalise people's actions wholly and solely in terms of two distinct psychological components: Beliefs and desires. These states constitute motivating reasons. Wishes, preferences, intentions, decisions, representations, understandings, feelings, moods and other varieties of associated states we employ in order to get a hold of people's behaviour, are either reducible to beliefs and/or desires, or contingent explanatory factors on other explanatory levels. There is no third distinct essential element in motivation, on a par with beliefs and desires.

It would be a gross understatement to point out that this way of thinking about actions is common. I would be willing to place a bet on the empirical claim that, at present in our society, at least 50% of the philosophically untutored would agree roughly with the BD picture, were they forced to produce an opinion on the matter. (A small bet though, since my estimation of the odds is a bit unreliable. I base it on a somewhat dubious generalisation from my children, friends and undergraduates in philosophy.) The purpose of the present work is to expose and summarise the conceptual and metaphysical commitments inherent in taking this widespread attitude to people. It offers a theory about how people think when they explain and predict actions, based on observations of how we talk in those terms.

Although my primary ambition is to explicate the metaphysics of action inherent in common action explanations, my own intuitions are firmly in line with the scheme I present. I am convinced that common criticisms of the BD model are misdirected, and that the BD model, even taken in its most literal and realistic form, can be upheld without internal contradiction, and without clashes with everyday experiences of agency. The two enterprises are mutually supportive. The empirical fact that a certain way of thinking about behaviour is well established gives us good reasons to require strong counter-evidence

before abandoning it. On the other hand, if it is demonstrated that the BD model suffers from internal inconsistencies and evident empirical anomalies, we would not only have reasons to reconsider our theory of action, but also to question my view of the BD model as a part of common sense.

My empirical claim is that the theory presented here reflects our ordinary ways of talking about actions. That assumption does not, of course, imply or require that people actually use the jargon and terminology I have to use in order to articulate the theory. Nor is there any reason to believe that all those who think about actions in the BD model way would agree to the explicit exposition of the theory they employ. There is no ground for thinking that anyone who masters the use of a certain notion must be able to define the notion in question, or to make explicit its conceptual connections. I want to show ontological assumptions that ordinary speech *commits* us to, rather than describe what the ordinary speaker would believe about the ontology of action if we ask him to think about it.

A comparison with meta-ethics shows the difference clearly. Try to present, to a group of children or beginner students, and without begging the question in its very formulation, the problem of what we really mean when we say things like “That was bad!” You will soon find that Richard Hare, like J.L. Mackie, is wrong about that “ordinary moral thinkers /…/ would naturally, if they started to philosophize, first embrace some form of descriptivism” (Hare 1982 p.78, Mackie 1977). There is no pre-theoretical consensus in this matter. If you manage to explain the problem in a fairly unprejudiced manner, they will, after a while, form widely differing hypotheses, much like the ones established in the history of moral philosophy.

Some of them will argue that badness is just an emotional thing, others are convinced that there must be a truth in matters of good and bad. A few of the latter might conceive from that truth in terms of correspondence to natural facts, some might not, etc. They are all apparently able to communicate about good and bad within the group. So, whatever the correct semantic theory about ‘bad’ is, a majority within your group believes in some false theory about the meaning of a simple notion, which they are fully capable of using. Hare is certainly right in stressing that if “we want to find out what ordinary people mean, it is seldom safe just to ask them” (1982 p.80).

The method at hand is, instead, to examine how our terms are put to daily practice. Even if many ordinary people would suggest other philosophies of action than the one I assign to them here, their hypotheses on this matter might be ruled out by a closer look at their own linguistic behaviour. It is likely that such an examination would favour the BD model. Frank Jackson and Philip Pettit express a similar methodological point: “The patterns we described in terms of possibilities associated with beliefs and desires are not news to the folk. They use them implicitly all the time in predicting behaviour. All that is unfamiliar to them is the jargon and the theoretical articulation.” (1995 p.270).

Which assumptions about people's inner constitution does the BD model involve? Will it commit us to controversial standpoints in philosophy of mind or metaphysics in general? Does it impose any important restrictions on behaviour? Will it give any guidance in questions about how we ought to behave? Most of the things that can be said in these matters have been said before, but I believe that there is a value in summing and displaying the BD model's typical claims in a unified picture. One reason for believing that a comprehensive critique is of value is that many of the assertions typically associated with the BD model, e.g. about its explanatory and predictive force or about its tautologous character, may seem appealing on their own but paradoxical when combined.

Some philosophers would agree with me that the belief desire model is an established part of common sense or "folk psychology" while others view it as an academic prejudice, typical among philosophers fostered in a Humean, or perhaps Davidsonian, tradition. The BD model is a trivial conceptual truth, many would say. In other camps it is regarded as a descriptive theory, either based on firm evidence of essential or important elements of the human mind, or flagrantly overlooking brute facts about human psychology. Still others claim that the BD model should not be judged in terms of truth or likelihood at all, but in terms of things like predictive and problem-solving efficiency. They believe that to apply the BD model is less like accepting a theory and more like adopting a pedagogical metaphor.

In other words, philosophers disagree about how true, commonsensical, interesting, and useful the BD model is. The purported status of the model, metaphysically and methodologically, is under debate. Furthermore, the model's alleged implications for several questions about practical rationality and ethics are focussed in many philosophical controversies. But the initial problem to be solved is, simply, to characterise the distinction, which is the model's fundament. How do desires differ from beliefs? Two metaphors were used above in order to present the distinction: The "map" figure of speech, and, more popular in philosophical jargon nowadays, the functional division between motivational states on account of their opposite "directions of fit". How informative are these pictures? Could the point they express be put non-metaphorically? Let me give a hint about the difficulties.

Maps are external navigation tools enabling travellers to reach their destinations by providing information about the exterior. That analogy appears plausible as a characterisation of the role of belief, but it offers little information about desires. Travellers choose destinations and select routes. One might go one step further and say that the role of desires is that of selecting destinations and route in response to the map, i.e. goal and strategy given beliefs.[1] In order for that characterisation to avoid describing desires as homunculi, i.e. as little agents themselves, "selecting" must be taken metaphorically. Agents choose, although their inner states explain and rationalise their choices, according to the BD model. A more truthful picture would therefore be that of a pre-set automatic

pilot with a satellite navigator, where the programmed map as well as the programmed destination commands are parts of the same steering system. Such a revision of the metaphor would, however, fail to catch the difference between two fundamental roles in the way it is supposed to. The “map” and the “command” are internal features of the steering device which both perform the same task: they make the rudder react in pre-programmed ways upon external stimuli (the satellite signals). Both of them direct the vessel towards its destination in response to signals from the world outside it. In this interpretation the metaphor would not mark out the belief desire model from a “desire” model, i.e. a model in which one type of internal state combines the motivational roles of beliefs and desires by reacting upon evidence and motivating action.

The picture of desires and beliefs as having different directions of fit is initially more appealing. But in which sense are beliefs and desires “directed” differently towards the world? Causally, they seem to operate along similar headings. Experiences of the world affect these inner states and they both determine how we react upon the world. Different directions of fit do not, then, translate into courses of causation.

The metaphor is also applicable to speech acts. Commands are supposed to make the world right, while reports are right when they fit the world, for instance. But that does not help us to unfold the metaphor either. As John Searle (the inventor of the explicit expression “direction of fit”) notes, it seems more reasonable to regard criteria of plausibility of speech acts as parasitic upon those of intentional states than vice versa. (1991, p.101)

Michael Smith proposes that the fitness direction of an internal state is explicable in terms of certain counterfactuals, which are true of the agent who has the state. A “belief that p tends to go out of existence in the presence of a perception that not p, whereas a desire that p tends to endure, disposing the subject in that state to bring it about that p.” /…/ We might say that this is what a difference in their direction of fit is.” (1994, p.115) This conditional statement (simplified, Smith admits) appears to be true of most beliefs and desires.

But as Hume notes, some desires tend to go out in the presence of evidence that their object is absent. “Absence”, he says, “is observed to have contrary effects, and in different circumstances either increases or diminishes our affections.” (1739, p.162) Other desires tend to endure as long as the agent perceives their effect. Fred Dretske characterises the object of desire, generally, as an effect reinforcing the desire’s influence on the behaviour. (1992, ch.5.) R.B. Brandt defines ‘happiness’ as a state such that it makes the agent disposed to stay in it, and this depends on the internal qualities of the state. Both these characterisations appear to fall in between Smith’s categories; They describe desires, which endure in the presence of perceptions of their objects but still make the agent disposed to keep on realising that object. In other words, it depends on the content of the desire, whether it goes out in the presence of a perception of its object, and if evidence that the

object is absent will make it endure and make the agent disposed to bring it about. Within a broadly functionalist framework, like Smith's own, where belief- and desire *contents* are also determined by functional role, this way of distinguishing between directions of fit seems especially problematic.

I will elaborate the point in section 1.3.2, but my suspicion is that what makes the metaphorical notion of fitness directions appealing to adherents of the BD model is its normative character. It suits a Humean norm of practical rationality, which says, roughly, that rational criticism of motivation should be confined to instrumental means-ends reasoning. At the bottom of all motivation there is an action-guiding element with full immunity to reason, an element without obligation to answer to any external fact. Although that restrictive norm naturally and conventionally is thought of as being intertwined with the BD model, it is not an internal part of the model's inherent picture of the motivational process.

The catchy phrase 'direction of fit' does not, then, provide us with a substantial description of how desires differ from beliefs. Considering "how much slack there is in the phrase "direction of fit", it is for certain purposes "better to talk directly in terms of patterns of dispositions" Michael Smith says in another context (1994 p.209). I am inclined to apply that reservation to the BD model. To understand the model's internal features, the most crucial task is to examine in greater detail its dispositional but realistic concept of desire.

One of the things many people appear to believe, if forced to theorise about their abilities to explain actions, is that there is a peculiar difference in kind between actions as seen from the agent's view, and as seen from an impersonal perspective. They might admit, though, that empathy and imagination allows something close to first person understanding even from a third person's viewpoint. Insofar as ordinary people embrace this sharp distinction between personal and impersonal perspective on agency, they are in good company. Many philosophers defend some versions of the Kantian assumption that there has to be a fundamental antinomy in a person's view of persons.

"We can act only from inside the world, but when we see ourselves from the outside, the autonomy we experience from inside appears as an illusion, and we who are looking from the outside cannot act at all" (Nagel 1986 p.120). It is sometimes presumed not only that the two perspectives are mutually exclusive, but that one of them is superior and should replace the other. Various forms of scientism might suppose that ordinary action explanation in terms of intentional states reflects a mild form of Cartesian superstition and that it is really no explanation at all. Philosophers such as Thomas Nagel and Jennifer Hornsby declare, on the contrary, that agency "is absent simply, from the impersonal point of view" (Hornsby 1995 p.185) and that "the intentional explanation of my action /…/ is comprehensible only through my point of view" (Nagel approvingly quoted by Hornsby p.179).

People should not be so pessimistic about reconciling their daily impressions of their own agency with an impersonal view of what they do. When we think of our actions in terms of beliefs and desires, we understand actions in causal terms. Belief-desire explanation is a subcategory of causal explanation in general. It is internalistic in that it picks out internal states as real driving forces in motivation, especially by assuming that some of these forces (non-instrumental desires) are fundamental and not based on any assertions about how external things are. Furthermore, in assigning representational content to those inner states it gives the personal viewpoint a methodological advantage in some cases. The question of how a situation is conceived from the agent's point of view will always be crucial to anyone who wants to find the inner causes that make his behaviour intelligible. On the other hand, the characterisation of those internal states can be put in purely impersonal terms. Their content is determined by their role in relation to external outputs and inputs, i.e. by facts we can describe impersonally.

A related common presumption about people's capacity to explain behaviour is that the agent has unique access to important factors enabling him to explain his actions in a more reliable way than other people could do. The second theme in my picture of the BD model, which initially might seem counterintuitive, is that the conceptual scheme we commonly employ in order to understand people gives no such priority to the agent. Each person may have unique access to some facts about his inner life, but the states that produce his behaviour are not among them. In this matter, I have always found the conventional claim strange. Let me briefly explain why.

In life and science, explanatory hypotheses are tested against predictions. We may find out whether _ might have been a condition for x to happen by staging iterated attempts to call forth x, with and without the aid of _, other initial conditions varied etc. Our confidence in their partnership will corrode if _ and x are seen on their own too often, but as long as they stick together, our belief in conditional dependence is toughened. Similarly, we will be more inclined to believe the gossip that B is married to A because he wants her money, if we observe other actions one might expect if that is his desire. E.g., that he divorces her when he finds out that she is broke or that he is prepared to marry anyone with a reasonable fortune.

No earthly creature has seen more of your actions throughout your life than you have. If you want to explain why you do the things you do, you ought to be in a privileged position, just in virtue of your access to empirical material. No sample of observations of input and output frequencies could be larger than your own. You should know the most about falsified and confirmed predictions and explanations concerning this individual. So even if you restrict the base of your investigation to public sources, i.e. to your behaviour under different external circumstances, as seen by a third person, we should expect your inferred

explanations to be more reliable than other people's. All the more so, if we suppose that there are some important explanatory factors, which in principle are knowable only to you.

Why is it then that first person predictions about behaviour completely lack that firm reliability? Not only are we often proved wrong about how we will act in many situations, spectators close to us are even better at giving forecasts about our actions under different conditions. As Frederic Schick puts it, "we sometimes choose, and /…/ we can't then know beforehand what we are going to choose, can't even know what we will do — or even properly believe we will choose it or that we will do it. Others may know this, but we ourselves can't" (1999 p.11). Schick argues that the notion of foreknowledge of one's own decisions is inconsistent. That is in line with my view (to be defended) that our attempts at predicting our own behaviour tend to be less incorrect when we try to view ourselves from the other's perspective — and that this is what we should expect. The point I want to make here is just that since first person predictions are often mistaken, first person explanations are rarely reliably confirmed. That ordinary experience should undermine our confidence in the idea that the agent is privileged when it comes to knowing why he did what he did.

In daily life, I think of myself and other people in terms of the BD model. Therefore, I feel sympathetic towards Michael Smith's view that "common-sense explanations all presuppose the availability of a standard, Humean, belief/desire explanation" (Smith 1998 p.39). He argues that the BD model is the unifying common ground that individuates philosophy of action as an academic discipline (though many contributors to this enterprise so far apparently must be unaware of this). But when I think of that model in philosophical terms, I have to admit that other possible positions in philosophy of action appear. I am, e.g., inclined to think that the view of human motivation as a purely cognitive process can be consistently worked out, provided that it is combined with the right kind of positions concerning notions of knowledge and truth. (The troublesome positions are the ones in between, exploiting the force *both* of belief-based action explanation and of conventional positions regarding epistemology and metaphysics.)

The BD model is a third person perspective screen, which is designed to facilitate explanations and predictions of behaviour. It comes with a weak means-ends view of practical reason and a tendency towards moral anti-rationalism. It is incompatible with clear-cut akrasia and it has implications for related practical issues such as autonomy.

In a sense, I regard your decision about explanatory/predictive model as an existential choice. Philosophies of action select a certain role for agents in the world. The BD model stresses the agent's inner set-up as the motor of intentional change. Purely cognitive philosophies of action depict agency from the opposite perspective. Such views stress the world around you as the initial force and depict your actions as proper or improper responses to that force. While they let the world pull your actions from you, the BD model stresses agents' pushing the world. The last metaphor is, of course, a variation on the theme

"direction of fit". And as with that phrase, its substantial point is normative. It expresses two different opinions about what external facts can reasonably require from an agent. If I leave the meta-psychological plane and apply my scheme of agency to you, I cannot help but to regard your choice of a fundamental norm as rational and genuinely subjective — as an ordinary interpreter of your actions I must beg the action-theoretical question. When you adopt or reject such an attitude, you simply make up your mind about which perspective towards the relationship between you and the world you would feel most comfortable with.

---

[1] Gunnar Björnsson unfolds the map-metaphor of desires and beliefs as a picture of "strategy-selecting states" and representational states, where the role of the latter is that of "inner maps" (1998, p.20-21). F.P. Ramsey's famous use of the metaphor pictures *agents* (rather than their inner states) as steering by beliefs: "[A] belief …is a map of neighbouring space by which we steer." For an interesting interpretation of Ramsey's metaphor in this context, see Sahlin 1991.

Chapter 1

1 The Belief Desire Model as a Philosophy of Action

1.1 A Philosophy of Action

The BD model is a conceptual framework utilised in everyday explanations and predictions of actions. A theory about the BD model must therefore involve analysis of our psychological thinking. This book is an attempt to expose common sense philosophy of action. It is not only supposed to define the notions we use to explain behaviour, but also to display the ontological suppositions we commit ourselves to by applying this conceptual scheme to people.

When Michael Smith and Philip Pettit present what I regard as an analysis of the belief-desire model in "Backgrounding Desire", they note that it is irrelevant to their thesis, whether the model in question truthfully depicts what really goes on in the mind of agents. The analysis might be correct even if the model is nothing but "a useful fiction of the sort Dennett takes it to be" (1990, p.565). In one sense, this disclaimer is applicable to my description of the BD model as well. I may be justified in claiming that my description really corresponds to general elements in our psychological thinking, without committing myself to the philosophical theory inherent in those elements. Perhaps the correct diagnosis of how we think about agency is an error theory.

On the other hand, I believe that the usefulness of the BD model requires that its users really subscribe to the ontological assumptions inherent in it. I.e., when we make use of the model, we are not just applying a handy scheme of definitionally intertwined terms; we are exploiting metaphysical implications of the notions they express. It is not possible to employ the model and at the same time consider its ontological implications to be nothing but fiction. The predictive and explanatory force of the model rests on those ontological assumptions. My confidence in beliefs and desires as indicators of how people will react, does e.g. rest upon assumptions about the ontological character of beliefs and desires. In other words, the error theory cannot be part of the philosophy of action inherent in the model.

The BD model is a conceptual scheme entailing, or embedding, a philosophy of action, I claim. What is a philosophy of action? To begin with, something should be said about the philosophy of action *discipline*. The primary job of this branch of philosophy of psychology is, loosely speaking, to detect and analyse essential components in human agency. It aims to specify those elements in our picture of human motivation the absence of which would be unthinkable, and without which the picture would be unintelligible. Let me label concepts referring to such elements as *action-theoretical concepts.*

A broader characterisation should add, as Michael Bratman suggests, that philosophy of action also explores those elements of human motivation that are of fundamental

importance to human beings, whether these elements are strictly necessary for our understanding of agency or not. Some notions concerning agency are central to the kinds of lives we want to live" (Bratman 1998 p.58) Such concepts need not, however, refer to entities essential to the lives we *have* to live as agents. 'Plans' and 'policies' are important notions of that kind, according to a common theme in Bratman's works. From this it follows that it may be a proper and respectable job for philosophy of action to analyse concepts which are not action-theoretical in my narrow sense.

To take a simple example, if it is necessary to master the notion of 'belief' to some degree in order to view a creature as acting intentionally, then it is part of the first task assigned to philosophy of action to analyse that concept. But, say, *resolutions* are perhaps not necessary and irreducible elements in agency. (The example might be questioned and I am not assigning this view of resolutions to Bratman.) We might understand every action a person performs as intentional, e.g. in terms of his beliefs and desires, without assigning resolutions to him. Nevertheless, this may be a notion of great importance to us in the sense that our ideas of the nature and possibility of resolutions could affect our practical strategies, norms of rationality, self-respect and so on. That would not make 'resolution' an action-theoretical concept, but it would qualify it as an object for analysis, in accordance with the second job Bratman reserves for philosophy of action.

It must be stressed that the two tasks hereby assigned to philosophy of action are distinctly different in character. So are, partly, their methodologies. Examinations of common sense conceptions, by way e.g. of construing examples appealing to our linguistic intuitions, are useful when any concept is analysed. But since the first task of philosophy of action is to outline the essential formal structure of agency, philosophical examples will here play a more crucial role. Their purpose is in this case to explore the limits for what we can say, and still make sense, concerning actions.

Suppose someone presents a consistent and sensible example of an action in the genesis of which the agent has no resolutions whatsoever, or shows that a convincing description of resolution-based action can be analysed purely in terms of beliefs and desires. The first would be sufficient to undermine the idea that resolutions are necessary for intentional action, while the second would show that resolutions are not irreducible elements in the formal structure of motivation. Either way, 'resolution' would thereby be excluded from the conceptual scheme inherent in the very notion of agency — from the scheme of action-theoretical concepts. But if the job were to refute or underpin less general ideas about the existential, moral or practical importance of resolutions, philosophical examples would not have that conclusive character. Psychological (empirical) evidence will e.g. almost always also be of relevance to such conclusions. Furthermore, construed examples would then have to be credible, not merely intelligible.

It would be superfluous to point out this difference, were it not for the fact that it is unclear in some cases, whether a certain philosophical claim about human action is to be taken in the first or the second sense. The result may be confusion. Some examples of the importance of being clear about this are to be discussed e.g. in ch.5 and ch.7. One such example concerns the explanatory role of perspectives, understandings, and "seeing as a reason". Another example of a similar kind in connection with the BD model, is the debate over the role of *intentions* in the production of actions. Few people would now agree with Davidson's opinion in "Actions, Reasons and Causes" 1963, that 'intention' is a concept devoid of explanatory force. Even those of us who believe in the reduction of intentions to beliefs and desires would find such a position exaggerated. The explicitly non-reductive views about intentions are ontological suppositions placing intentions firmly on the action-theoretical level together with beliefs and desires. (Bratman 1987 and Mele 1992 are examples of that type) Other suggestions appear to side with those views, but are in effect difficult to distinguish from refined versions of the reduction. Davidson's later proposed identification of "pure" intentions with all-out judgements belongs to the latter category in my view (1978).

When I claim that *a* philosophy of action is inherent in the BD model, my point is that the BD model is a scheme of action-theoretical concepts in the narrow sense. To employ these concepts is to utilise specific ontological assumptions about actions in general. The model gets its predictive and explanatory force from an implicit theory about the real nature of human actions.

Summary of 1.1: In a narrow sense, *a philosophy of action* is a set of ontological assumptions expressed in a scheme of action-theoretical concepts. A concept is action-theoretical if it is essential to our understanding of agency. The *discipline* philosophy of action also examines other action-related notions playing important roles in our lives, even if these concepts are not strictly necessary for us to regard someone as an agent. The BD model embeds a philosophy of action in the narrow sense.

1.2 Philosophy of Action and Decision Theory

Like a standard view in economical decision theory, the BD model ties desire with conceptual necessity to action and choice. On both views, acting is the ultimate key to desiring. It is sometimes thought that the BD model is nothing but a terminological variation of standard formal decision-theoretical analysis of individual utilities as determined by choice behaviour. This identification is put forward from both perspectives. Richard Brandt does it from the viewpoint of the BD model, when he assumes that the precise results of decision theory can be straightforwardly employed to

measure strength of desires, within the framework of a theory where desires are real and causally potent inner forces. [1]

The identification appears to be even more common in economical decision theory. "The crux of the question", Amartya Sen notes, "lies in the interpretation of underlying preference from observations of behaviour" (1973 p.241). Like Dan Hausman (1999), and Bengt Hansson (1988), Sen finds it important to emphasise a distinction between — to use Hansson's labels — a realistic and a formalistic interpretation of axiomatised decision theory. The distinction "has to do with the degree to which the interpretation provides a bridge between the formal apparatus and some outside, empirical phenomenon." (Hansson 1988 p.142)

On the first interpretation, it is literally true that the "individual guinea-pig", "by his market behaviour, reveals his preference pattern" (Paul Samuelson 1948 quoted by Sen 1973 p.241). I.e. patterns of choice are taken as *evidence* of inner states (thus; "revealed") that drive people towards the actions they choose. "It is my guess", Hansson says, "that the realistic interpretation is by far the most common one in the literature, especially in texts addressed to business audiences" (1988 p.144). And according to Hausman, "many economists have drawn the mistaken conclusion that the proofs in this literature [the "revealed preference"-tradition initiated by Samuelson 1938] demonstrate that choice reveals preference".

The formalistic interpretation *defines* preference purely in terms of choice behaviour. On a radically formalistic interpretation, the plausibility of the theory is merely a matter of internal coherence and consistency. As a descriptive theory of decision it is less empirical - less susceptible to falsification. The distinction formalistic/realistic cuts across the categorisation of decision theories as normative or descriptive. Normatively, the formalistic interpretation will make the theory less practical and harder to violate. Values expressed in the realistic version may be offensive or insincere, its usefulness and intuitive plausibility are criteria of relevance, while the formalistic reading makes no allegations to our prior wants and values. (Hansson 1988 p.145-6)

It should be clear by now why it is a mistake to identify the BD model with the decision-theoretical analysis of desires in terms of choice, via the notion of preference. The BD model is a philosophy of action, with ontological and empirical content. It has practical implications and our set of intuitions about practical rationality is therefore one of the external factors against which it can be measured. Decision-theory need not necessarily subscribe to any such philosophical view. The "revealed choice" concept of preference, employed by economists, can be used in a technical sense without committing the decision theorist to a hypothesis about how actions are caused, or about ordinary language meaning of intentional concepts. So, although the BD model is not at odds with the decision-theoretical approach, it is not implicit in it either.

I do not deny that the common source of inspiration for traditional decision theory is the realistic view that desires and beliefs are the two types of factors that determine our

decisions.[2] So there is a natural and generic affinity between decision theory and the BD model's philosophy of action. Furthermore, the realistic interpretation claims that formal decision theory offers a true description of decision-making on a BD model basis (or, normatively speaking, it offers a useful tool for decision-making on a BD model conception of practical rationality). The realistic interpretation constitutes a testable claim about the relation between axioms of rational decision-making and the real basis of our decisions. It is not a fallacy to tie formal decision theory to a specific philosophy of action in this way, but a substantial philosophical assertion, which gives decision theory greater explanatory and normative force.

The fallacious identification occurs only if we equivocate the formalistic definition of a technical notion of "preference" with the realistic concept of preference derivable from the BD model's notion of desire. This is what some economists do if they, as Hausman says, believe that formalistic proofs concerning patterns of choice "demonstrate that choice reveals preference". Facts about "preferences" in the technical and non-explanatory sense can perhaps be demonstratively proved from rational choice axioms, but not facts about the preferences explaining choices, since they are external to the formal apparatus. The error stems from the exaggerated verificationist standard that a "scientifically respectable" theory about economic behaviour must "explain that behaviour without reference to anything other than behaviour". (I.M.D. Little, quoted by Sen 1973) As Sen remarks, on such an interpretation, the theory about how choice reveals preference would appear to use the word "preference /…/ to represent an elaborate pun". (1973, p.243) Any theory exploiting the explanatory and normative force of concepts like 'preference', 'desire' and 'belief', will appear to depict people with *some* "pretence to see inside their heads" (which is impermissible for "the econometric theory of demand" according to Hicks, quoted by Sen 1973 p.242).

Summary of 1.2: The BD model may be an important source of inspiration for traditional decision theory, and it is the sort of ontological underpinning that gives axiomatised decision theory greater explanatory and normative force. Nevertheless, it is a fallacy to equivocate the technical "revealed choice" notion of preference with the realistic concept of preference derivable from the BD model's 'desire'. The source of the fallacious tendency to equal these views is an exaggerated verificationism, according to which a scientifically respectable theory must "explain behaviour without reference to anything other than behaviour."

1.3 Evidence for the BD model? Another Methodological Note.

1.3.1 Patterns and Predictions

In the interminable Library, "a thinker once observed that all the books, however diverse, are made up of uniform elements". The thinker "deduced that the Library is total and that its shelves contain all the possible combinations of the twenty-odd ortographic symbols". One of the much consulted books, for instance, "is a mere labyrinth of letters, but on the next-to-last page, one may read *O Time your pyramids."* (Borges 1941). Perhaps you found the book in front of you in Jorge Luis Borges' Library. It is there, for sure.

According to Borges' narrator, it would not be altogether improper, in that case, of you to regard this combination of natural symbols (ink shapes eventually imitated by the inventors of language), as altogether meaningless. Like the empiricist and theist Kleanthes in Hume's Dialogues, Borges' narrator (I am not sure about Borges, though) apparently thinks that it makes sense, at least, to distinguish the meaningfulness of the book under these circumstances, from what the pattern would convey if we assumed a certain specific known history. However, Kleanthes, like some of the Library's believers, would on account of induction be convinced that the pattern in front of you, discernible to you in virtue of your linguistic capacities, *reveals* intention. Kleanthes asks his friends to imagine that books grew wild in nature, and continues rhetorically:

Suppose, therefore, that you enter your library thus peopled by natural volumes containing the most refined reason and most exquisite beauty; could you possibly open one of them and doubt that its original cause bore the strongest analogy to mind and intelligence? (1779 part 3)

Independently of where you found this book, you have imposed your scheme of interpretation on it, and found some pattern on its pages. Other patterns may be there as well, discernible on other schemes, typographical as well as linguistic. (Should you have a considerable amount of time left, you may consult the Library. In there you will find innumerable grammars and dictionaries providing you with just as many different decipherings of these combinations, some of them even expressing refutations of what you just read.)

A brilliant and imaginative decoder might detect a pattern from just viewing a couple of pages, without recognising a language to begin with, and without having seen similar combinations to generalise from. Then she might make explicit the syntax etc. of the pattern, and start making predictions about which combinations that are to be expected. She may "explain", in terms of syntactical rules, why certain combinations are frequent and others are rare. Suppose her predictions are correct up to now. Should you place a

bet on her next forecast?

My guess is that you will not feel any comfort in her predictions unless you assume that there is *some* underlying explanation of the pattern she has seen. If you are sure that the book is just one out of zillions of randomly produced series of signs, you will hold on to your purse. It is because you take it for granted that her successful predictions exploit some more basic "natural" order, not just an order imposed by her, that you are inclined to risk your money. You also, in that case, regard deviation from the pattern as *inconsistent* with the general rules she has access to. And the notion of consistency required in this context is, I suggest, narrower than the concept of a pattern.

The relevance of my present digression about patterns lies, of course, in the fact that the BD model is one kind of scheme employed to extract patterns and make predictions. Although Borges' short story would serve the purpose, I am not interested in making any points about *meaning* (whether of life or of language) here, just about predictions. (The BD parallel to the Library would be a world where a thinker, perhaps a clever semiotician, had observed that in an interminable population of mortal agents, the number of human action-constituents is finite, and also found that every conceivable combination of such constituents is realised.)

The view I want to defend here is, firstly, that the predictive value of the BD model is dependent upon a realistic application of it. The BD model thinker must not only take seriously that patterns of behaviour are real; he must to some extent also exploit the commonsensical idea that 'belief' and 'desire' corresponds to real natural kinds, *underlying* those behavioural patterns. (What follows in chapters 2 and 3 is an attempt to spell out the weak ontological assumptions he is thereby committed to.) The explanatory and predictive powers of the model will otherwise vanish.

Secondly, I want to make clear that we assign to agents an amount of consistency on the BD model scheme, not merely behavioural regularities. 'Consistency' is normative to some extent, and when the term is applied to behaviour, the norms involved are about rationality. Furthermore, these two claims are not only compatible with each other but also mutually supportive. The normative force of consistency claims *exploits* the model's allegations about underlying causes. (I guess that the standard view is the opposite one; that the idea of an inescapably normative element in our interpretation of actions, as presented by Dennett or Davidson, goes hand in hand with the assumption that the hypotheses about complex underlying causes of behaviour are of little importance to the usefulness of our interpretative schemes.)

As Sen notes concerning the possibility of redefining the notion of 'preference' in terms of mere consistency in choice:

[It] is then legitimate to ask what does "consistency" of behaviour stand for and on what basis are the required consistency conditions chosen. The alleged inconsistency between (i) choosing *x* when *y* is available and (ii) choosing *y* when *x* is available, would seem to

have something to do with the surmise about the person's preference underlying his choices. (1973 p.243)

Dennett in "Real Patterns" suggests a useful weak notion of 'pattern'. Any series "of dots or numbers or whatever" has a pattern if it is not random. The series "is random if and only if the information required to describe (transmit) the series accurately is *incompressible:* nothing shorter than the verbatim bit map will preserve the series." (1991b p.32) No piece of information could transmit a sufficiently chaotic labyrinth of letters without being at least as complex as the labyrinth itself. But the book in front of you has patterns that could be described in terms of some general rules (typographical, grammatical, logical etc.) and notes about exceptions etc. that with some effort could be made shorter than a transcription of the text itself. A pattern in this weak sense is by (Dennett's) definition *real*, independently of its genesis. Dennett uses a clarifying example, roughly similar to this briefer and simpler version:

Make a program that lets your computer produce a clearly visible pattern, like a row of five black squares separated by four white squares. Allow "noise" to interfere with the printing increasingly, until the original pattern is barely discernible and finally becomes indiscernible. Then, for comparison, you can construct a program, the capacity of which to produce distinct patterns (of a kind visually similar to the first pattern) slowly degenerates, so that it gives you decreasingly discernible patterns without taking the route via designing a constant distinct pattern and then adding gradual contamination with noise. Suppose we compare the two stages where the pattern is *almost* indiscernible, as a result of the two different processes. Dennett's point is that

even if the evidence is substantial that the discernible pattern is produced by one process rather than another, it can be rational to ignore those differences and use the simplest pattern description as one's way of organizing the data. (1991b p.44)

So, if the pattern is there when the clear-pattern generator continuously works, though the picture has become blurred by noise into indiscernibility, then it is there just as much when the picture has become fuzzy due to a non-distinct-pattern generator. If the pattern really becomes *in*discernible, he argues, it is also obliterated — independently of how this came to happen. (A description of that picture would have to be *longer* than a piece by piece transcription, in order to depict the indiscernible pattern beside the discernible chaos.) It is not as if the first pattern is there all the time, becoming more and more covered by noise, while the second is gradually transformed into something more complicated. Patterns have unbreakable links to observers.

How does this kind of parallel relate to BD model explanations? What kind of conventional story about underlying conditions for patterns does he want to undermine? To begin with, *where* would Dennett say that the pattern is in this case? The answer to

the last question is clear. Although he stresses that the success of BD predictions "depends on there being some order or pattern in the world to exploit" (1991b p.30), behaviour is what matters. The "pattern is discernible in agent's observable behavior when we subject it to 'radical interpretation' (Davidson) from 'the intentional stance' (Dennett)" (1991b p.29). I.e. behavioural patterns exist in behaviour, due to the spectator's interpretative schemes. The appearance of definiteness and precision in BD descriptions comes from language, not from distinct counterparts in the mind.

One of Dennett's explicit ambitions in "Real Patterns" is to convince philosophers that his position "is not just the desperate dodge of ontological responsibility it has sometimes been taken to be." (p. 29) Nevertheless, one way of reading his parallel is as just another confident declaration of the verificationist methodology suggested by numerous metaphors, analogies and overt commitments elsewhere in his works:

The verificationist's general complaint about the realist is that he is insisting on differences (between, e.g., bats with private lives and bats without, dogs with intrinsic intentionality and dogs without) which make no difference: that his intuitions cannot be integrated in an explanatory scheme because they are "wheels which play no part in the mechanism" [Wittgenstein, 1953, I, #271]. This seems to me a good complaint to make, and the only one we need to make. (Richard Rorty, approvingly quoted by Dennett, 1991a p.461)

There is a vacuous ring to the verificationist's rebuke. Is not the insistence on "differences which make no difference" simply a straw man? However, I believe that the claim really is substantial and practical, as Dennett's metaphorical introduction to "Qualia Disqualified" suggests:

When your kite string gets snarled up, in principle it can be unsnarled, especially if you're patient and analytic. But there's a point beyond which principle lapses and practicality triumphs. Some snarls should just be abandoned. Go get a new kite string. It's actually cheaper in the end than the labor it would take to salvage the old one, and you will get your kite airborne again sooner. (1991a p.369)

His point is pragmatic. Applied to philosophy of action, it implies, roughly, that we should not bother with the entangled underlying "reality" of behavioural patterns, as long as we get what we need — reliable predictions — from what is out there in the light anyway. It will simply be a waste of energy.

Things should not be multiplied beyond necessity (Ockham) and I am prepared to grant that in such a well-intentioned piece of economical advice, "necessity" may well be understood as *practical* necessity. To admit just one more metaphor here: If someone wants to put more weight in your rucksack, the burden of proof is on him to show that

the extra items will actually make things easier for you. So, where does my picture of the BD model or "folk psychology" depart from Dennett and the tradition he represents?

Before answering that question, let me mark some positions I find the BD model to be compatible with:

a) The multidimensional complexities of processes underlying behaviour need not have any distinct limits and simple physical or mentally present characteristics making them correspond in a direct way to belief and desire talk. Propositional attitudes are identified in terms of dispositions for behaviour, not neurally or introspectively. The reference to underlying processes I claim that BD talk commits us to requires merely that these processes can be delimited (in principle) via their functional roles as bases of dispositions.

b) Differing types of underlying processes might produce similar observable patterns of behaviour, and similar dispositions for action. In such cases, it *may* be proper to ascribe similar beliefs and desires to the agent on account of the behaviour, in spite of the imaginable difference in genesis. If the relevant conditional predictions are similar, there is no reason to vary the BD labels.

c) More than one pattern can be discerned in a sequence of behaviour. Our pick of intentional description on the BD model is in that case not an entirely descriptive matter. Firstly, some background assumption of shared beliefs and preferences is needed to reach any kind of conclusion about the agent's propositional attitudes in a specific situation. Secondly, we may have a choice between rationalisation (attitude makes sense of behaviour) and charity (attitude is true or good) in deciding how far from our own convictions the other's attitudes could depart before we must regard him as having renounced agency altogether. (See section 1.3.2.)

d) It is *imaginable* that an entirely different conceptual scheme — a purely cognitive model of our internal set-up, for instance — could be as efficient as the BD model in helping us to extract patterns and make predictions about actions. These two models could both be structuring a sequence of behaviour and its underlying basis, although they would yield different structures. That possibility is related to the fact that mere regularities in behaviour are insufficient to fix patterns. Our ability to see patterns in behaviour is not just a generalisation from observations of similarities. This ability consists, also, in imposing certain norms of rationality on the data. (It could be compared to the economical "aim" of our perceptual apparatus to organise impressions in accordance with a limited set of *Gestalts*.)

Evaluation of these schemes has to involve things like their conceptual consistency, normative acceptability, and (linguistically) intuitive plausibility in handling delimiting

cases — these elements is what the present defence of the BD model is about. Since their predictive capacities also exploit realistic assumptions about agents' inner states, they both also in that sense have a testable empirical content. If both models are just as reliable in making predictions and accounting for observations, the empirical evidence is simply insufficient to make clear which one gives the true picture. In such cases, what actually settles the question for us may well be a matter of attitude — although the matter *is* determinable in principle.

The last possibility does *not* imply that there could be two interpretation schemes, both structuring the very same data correctly but in different ways (i.e. in accordance with their inherent norms of order, respectively), which were in genuine disagreement with each other. (In my view, this is where the BD model must leave Dennett's path.)

It is of course possible that two ways of structuring the data would yield different predictions, though both ways were just as well founded, in relation to available evidence. It is also quite possible that two ways of applying the *same* structuring device, the BD model, could fit the behavioural pattern and yield different predictions. B's behaviour might indicate that he wants to kill himself or that he cries out for help. You and I may have just as good grounds for ascribing these opposing desires to B, and subsequently we would make differing conditional predictions. But in that situation, at least one of the two ways of ascribing propositional attitudes to B would have to be mistaken *in principle*, although we may never come to find out which.

That has to do with the BD model's inherent assumption that the behavioural consistency detected enables us to make predictions only due to some internal warrant. Such an internal warrant is what I assign to B when I say that he has a certain desire. The predictive success of the BD model exploits that sort of realism about beliefs and desires. It would be *inconsistent* of B to want to kill himself, *and* to cry out for help. He cannot both be in a state such that he is disposed to kill himself intentionally, and in a state such that he predominantly wants to live (other things equal). And as Sen remarks, this kind of inconsistency requires some intentional state *underlying* his behaviour. If we should let behaviour *define* beliefs and desires, they would "not enable one to predict, explain or recommend choices" (Hausman 1999 p.17)[3]

None of the two barely discernible patterns generated by your computer were *coincidences* (in the example mentioned before). They were there because the program designed them, they appeared on the screen, and your perceptual apparatus aimed at pregnancy in sorting your visual impressions. Whether you know anything about the details of the program or not, you are clearly justified in making hypothetical inferences about how the pattern will develop, from watching a part of it. Knowing that the pattern *has* a non-coincidental cause is enough. Compare that to this book, as found in Borges' Library. A mass of signs is visible, and you can make use of interpretative capacities on various levels to extract patterns from it. The non-random patterns are clearly there, as

real as the patterns on the computer's screen. But suppose nothing distinguishes the history of this book from the history of the other books in the Library. It just so happens that the order of signs forms a pattern that *could* have had a non-random cause. When you know this, the next intelligible sentence should be just as unexpected to you as if what you had seen on the previous pages were, in Borges' words, "leagues of insensate cacaphony" (1941 p.74).

The BD model does imply that there are inner states of agents, demarcated by their function as causal bases of behavioural dispositions. The predictive success of the model indicates that these kinds of states *exist*. One of Dennett's formulations could (out of its context) be taken to express a similar idea: Beliefs are real as long as belief talk "measures these complex behavior-disposing organs as predictively as it does." (1991b p.46)

Even if overt behaviour allows conflicting interpretations, the question of what is really in the head of the agent concerns a difference that *does* make a difference. To *predict* behaviour without reference to anything other than behaviour is just as difficult as explaining it without such references. For prediction to work, it may suffice to know that this is one of the kite strings which in principle *can* be unsnarled, i.e. that there *is* an internal state of the kind needed to warrant certain behavioural dispositions. Proponents of the BD model do not have to subscribe to any specific account about the details of these inner states, and they have certainly no reason to think of them in simplistic terms.

Summary of 1.3.1: When we view people's behaviour as intentional acting, we use our interpretative capacities on what we observe, and extract patterns. This activity is not entirely descriptive. In order to see a pattern in something, we have to impose norms of consistency, and when it comes to behaviour, we identify rationalising elements via assumptions about shared attitudes. This point is epistemic, and it does not imply that two incompatible interpretative schemes in principle could yield different predictions from observations of behavioural patterns, and both be true. That has to do with the fact that predictions exploit realistic assumptions about the underlying causes of the patterns we observe. Predictions of behaviour, just like explanations, cannot be supposed to work without reference to other things than behaviour.

1.3.2 Direction of fit: A Priori Evidence for the BD Model?

Fit is symmetrical. When the shoe fits, the foot fits. Aims, goals, desires, values and the like often make one or the other of the relata prior to the other, though. Normally, the shoe should be adjusted if it does not fit, but when one of Cinderella's stepsisters carves off her heel, and the other one her toes, as in the Grimms brothers' tale, the shoe is the

objective for their feet to follow. (John Searle's nicer example concerning the same point contrasts Cinderella looking for a shoe with the Prince seeking a foot. — 1983 p. 8) As G.F. Schueler points out, this is a way of clarifying 'direction of fit' in terms of goals. "If we drop those goals, then we just have the fit between the shoe and the foot, but no "direction"." (1991 p.278) Unless the metaphor can be unfolded in non-normative terms, then it will be of little use in helping us understand what it is to have a goal. It may still be a useful metaphor, of course, in catching a certain normative point. But as an explication of what it *means* for an entity to aim at something (like the truth, or the realisation of some proposition) it moves in too small a circle.

John Searle, like Mark Aulisio (1995 p. 341) and most others in this debate, plausibly attributes the original idea of dividing psychological entities on account of their direction of fit to Elisabeth Anscombe. The explicit expression "direction of fit," was coined by Searle, as far as I can find out.[4] Some of Hume's characterisations of the distinction between 'reason' and 'passion' in *Treatise,* e.g. in 3:1:1, of reason as aiming at truth, and passion as setting ultimate goals, seem to express a similar idea. David O. Brink interprets Aristotle as saying the same thing in *De Anima* 3.3 – 9 (Brink 1997).

The notion is crucial to Searle's taxonomy of speech acts, as well as to his theory of action. "It would be very elegant if we could build our taxonomy entirely around this distinction in direction of fit, but though it will figure largely in our taxonomy, I am unable to make it the entire basis of the distinctions." (Searle in Gunderson, ed., 1975) W.P. Alston apparently shares many views on speech acts with Searle. However, he objects to an insufficient explanation of the notion of 'direction of fit'. In reply, Searle writes "Well, any philosophically difficult concept is always subject to further elucidation, but I have attempted to explain it in *Intentionality* to such an extent that I do not feel seriously concerned with this objection." (LePore/Van Gulick 1991, p. 101)

Anscombe outlines the distinction in *Intention* as an attempt to elucidate the difference between intending and predicting. Though Searle's application of the distinction is wider, he endorses her description. Like Anscombe, Searle appeals to intuitions about when speech acts or mental states are mistaken or unreasonable in order to illustrate direction. Suppose, e.g., that I find, in my pocket, a list of groceries which answers to the content of my shopping van. How can I tell which direction of fit this list might have? Should it fit the world or should the world fit with it? That depends, Anscombe and Searle would say, on how we would diagnose a discrepancy between the list and the content of the van. Would such a discrepancy indicate a mistake in my shopping, or a mistake in the list? In the first case, I am supposed to make the world fit the list (as a list of commands or expressions of my intentions), otherwise I am supposed to make the list fit the world (as a list noting what I have actually bought). (Anscombe 1957 p.56, Searle 1975 p.346)

Say that I am in a hurry, and must hand over the list and the shopping van to my daughter and then leave. When she is about to pay, she finds peppers in the shopping

van, but no “peppers” on my list. Did peppers by mistake, i.e. contrary to the intention revealed by my list, or contrary to the commands expressed on it, come to get into the van? In that case, the fault lies in the world, and she should adjust the world to the list by leaving the line and put the peppers back: World-to-speech act or world-to-mind direction of fit. Or did I not note on the list that we put peppers among the items in our van? Then, she has detected a fault in the list’s description, and that may reveal a flaw in my beliefs — perhaps I did not know that we were buying peppers. She should then adjust the list after the world, and eventually my beliefs as well: Speech act-to-world or mind-to-world direction of fit.

In the light of the central role played by “direction of fit” in Searle’s taxonomy of speech acts, as well as in his theory of action, his dismissal of Alston’s objection might seem careless. What makes Searle’s explication appear insufficient is probably that the normative element in the notion of ‘being supposed to fit’ is never eliminated from his analysis. Alston’s criticism is based on the view that Searle attempts to explain the fitness-direction of intentional states in terms of the direction of speech acts. (Alston 1991) In a reply to Alston, Searle makes clear, as I noted in the introduction, that criteria of validity of speech acts in his view must be parasitic upon those of intentional states, and not vice versa (1991, p.101).

I am not certain that it must be a flaw in Searle’s taxonomy if this central concept is essentially normative. Although if that is the case, it could perhaps have been made more explicit. When Mark P. Aulisio makes use of Anscombe’s division in a restored defence of the intention/foresight distinction, he stresses explicitly that “direction of fit” “is not about how the world comes to be in a certain way, but rather, it is about *what is normative with respect to what*, so to speak /…/ *or, what must form part of the agent’s motivational goal in acting”* (1995 p.349). A normative use of the distinction is acceptable, as long as it is not presented under the pretence of descriptivity. One may e.g. appeal to normative intuitions about directions of fit in the taxonomical enterprise of showing that valuing, intending etc., have something in common, as opposed to things like believing and predicting.

To return to the initial question of circularity: In order to expose the nature of this common element in goal-directed states or to *analyse* it, it will not do to appeal to examples of the above mentioned kind. Then we need a concept ‘direction of fit’ that does not depend on goals, purposes or values. Two questions might be distinguished in this context. Firstly, will ‘direction of fit’ provide us with an argument for a dualistic taxonomy of motivational states? Secondly, can the metaphor be unfolded to give us a substantial understanding of the difference between beliefs and desires?

An alternative formulation the central theme of the BD-model could be stated like this:

P1 R at t constitutes a motivating reason of agent A to ϕ iff there is some Ψ such that R

at t consists of an appropriately related desire of A to Ψ and a belief that were she to ϕ she would Ψ.

Michael Smith claims that a "really quite powerful argument" for P1 is that "P1 is entailed by the following three premises" none of which can plausibly be denied (1994 p.116):

(a) Having a motivating reason is, *inter alia,* having a goal
(b) Having a goal is being in a state with which the world must fit
and
(c) Being in a state with which the world must fit is desiring.

Smith cannot really mean that P1 is supposed to be deduced from (a), (b) and (c). None of the premises refer to belief, or to states or desires being "appropriately related".[5] But the three assumptions do entail P1's assertion that desires are necessary constituents in motivating reasons. The argumentative force of these three simple assumptions is therefore great, nonetheless. However, the argument works only provided that the idea of characterising beliefs and desires in terms of their different directions of fit really catches some undeniable element in our thinking about motivation. Otherwise it is hard to see why (b), "having a goal is being in a state with which the world must fit" is supposed to be self-evident. If that popular metaphor shall be considered as a premise in an *a priori* entailment of P1, it must be possible to unfold it in non-metaphorical terms. It must be made clear what it is for *x* to be such that it must fit with *y*, and how the opposite direction of fit is distinguished from this. Furthermore, such a literal interpretation of the metaphor ought to illuminate some fact besides the ones explicitly stated in P1. I.e. to be treated as a premise, and not only as a catchier way of presenting the distinction explicitly asserted in P1, the unfolded metaphor cannot lay too much weight on appeal to goals, values or other entities it is supposed to enhance our understanding of.

Searle indicates briefly that he regards the direction of causality as the oppositeof the direction of fit (1984, p.96) when it comes to motivation. If that is correct, this might provide us with a descriptive equivalent of the metaphor (although that is no explicit ambition of Searle's, here). However, Searle's explanation of this asymmetry is formulated in terms of intuitions about when beliefs/desires are successful. The world should cause my beliefs and my desires should make causal impact upon the world. (To be contrasted with 'my beliefs should fit the world' and 'the world should fit my desires'.) So, it is really not *causality* taking an opposite direction, but our ranking of adjustments, as compared to our ranking of existent states. If there is divergence between desire and world, desire is primary to world, and adjustment of world is primary to (i.e. ranked above) adjustment of desire — and vice versa when it comes to

belief. In other words, this is just another way of putting the normative point about what should be fitted to what. It does therefore not affect the point I made in the introduction, about how beliefs and desires, neutrally observed, work along the same causal paths. They are causally explained by external input, and they causally explain our responses.

Michael Smith's own suggestion is that the characterisation of desires as states, which are such that the world must fit with them, meshes with the dispositional conception of desires. The difference between a belief-state and a desire-state is that "a belief that p tends to go out of existence in the presence of a perception that not p, whereas a desire that p tends to endure, disposing the subject in that state to bring it about that p. /…/ We might say that this is what a difference in their direction of fit is." (1994, p.115) This characterisation of the dispositional account (which Smith admits as being rough and simplified) is put simply in terms of intentional states' counterfactual dependence on evidence concerning whatever the belief or desire is about. At least in that formulation, it leaves beliefs and desires insufficiently described.

As I mentioned in the introduction, Hume, Brandt, and Dretske all describe intentional states that appear to fall in between the categories, as outlined above. G.F. Schueler argues that *hope* is another example of this kind. Hope that p will be abandoned in the presence of a perception that not p, although hope in many cases, at least where p is thought of as *possible* to realise, makes the agent disposed to realise p. As Smith notes, various conditions may be necessary for us to have a specific desire or a specific belief — most notably other desires and beliefs. In this sense, some desires "may 'involve' elements of belief" (1994 p.114-15) 'Hope' could be a complex desire of this sort, characterised by its relation to various modal beliefs. In this sense, as a psychological phenomenon, hope displays both directions of fit. Though the belief and the desire components in this state might be psychologically inseparable, we can still abstract them from each other, in terms of different counterfactual assumptions. In other words, Smith's rough counterfactual account could probably be refined to accommodate counterexamples of this kind.

However, even in some simple idealised cases, it might be difficult to pick out beliefs and desires via actions and the kinds of perceptions mentioned by Smith. Imagine an almost fanatically engaged person who is prone to form strong beliefs and desires in any moral matter. Concerning, say, the lawful permission of patenting of genes, we may know that she either desires p or desires not p, but we do not know which — just that she is the kind of person who would not rest without taking a stand. Suppose she does not tend to realise p, nor to realise not p. Her passivity might be a manifestation of her desire for p combined with a belief that p, or – alternatively – of her desire for not p combined with a belief that not p. In terms of dispositions for behaviour, both pairs of belief & desire have a similar direction in the sense that they would make her disposed to stay put. Without independent insights into the content of

her beliefs or desires, the counterfactual condition above will not enable us to distinguish the first pair from the second.

It is quite plausible to assume that we can only identify a specific desire or belief from external behaviour given assumptions about other beliefs/desires. But one of the appealing features of the dispositional analysis is, as Smith stresses, "that it is precisely an account that explains how it can be that desires have propositional content; for the propositional content of a desire may then simply be determined by its functional role" (1994 p.114). I agree that content, as well as form of attitude, must be determined by functional role within the BD model. However, in a case like the one above, it appears problematic how to make clear in which sense the belief of the first pair differs from the desire of the second. We will have to know more about an agent's various convictions and preferences to detect a certain desire or belief of hers. Within the dispositional account, "the actions leave the intentional states underdetermined" (van Roojen 1995 p.45).

Mark van Roojen attacks Smith's argument against anti-Humeans about motivation, that if beliefs were motivators, accidie would be impossible. As van Roojen notes, this argument indicates the view that "the absence of an effect shows the absence of the disposition". Although van Roojen's aim is to undermine an argument against anti-Humeans, he notes that this is an idea that Humeans should have reason to object as well (p. 48). We can perhaps infer your desires from your behaviour when 'all things are equal', but often they are not equal. Desires and beliefs can be plausibly attributed to you in many extraordinary circumstances. On what grounds are such attributions founded? The mere regularity of your behaviour would obviously be insufficient.

In order to understand your actions as products of beliefs and desires, we must attribute some amount of rationality to your behaviour. The intentional states we endow you with should make sense of what you do. However, to fix rationality constraints is not an entirely descriptive matter. Generosity make us assume your behaviour to express desires and beliefs we regard as reasonable to some extent, while at the same time we may have to stretch our imagination in this respect in order to see *any* reason in what you do. We should not require more rationality than necessary, nor is there any reason to deny you sensible interests and convictions as long as they are compatible with your actions. But the balance is not given from the start — it is a matter of practice and good judgement.

Actions can be underdetermined because different combinations of beliefs and desires can rationalise some specific behaviour. In such cases, as Davidson argues, our understanding is guided by a *principle of charity*, according to which we should assume that most of what the other person believes is true. [6]

Quine's key idea is that the correct interpretation of an agent by another cannot intelligibly admit certain kinds and degrees of difference between interpreter and

interpreted with respect to belief. As a constraint on interpretation, this is often called by the name Neil Wilson gave it [Wilson (1959)], the *principle of charity*. (Davidson, 1990, pp. 318)

In a similar fashion, charity delimits the imaginable distance between the aims and goals of interpreter and interpreted.

Charity sometimes tips over into paternalism — a refusal to accept that the things people strive for really are good for *them*. As David Lewis notes, there may be tensions between the principle of charity, that draws me to endow you with true beliefs and apt desires, and the "rationalisation principle" requiring that motivating beliefs and desires also suit your behaviour properly (Lewis 1983). If your outlooks are somewhat different from mine, the more I model your beliefs and desires on my own, the less of your behaviour will seem intentional to me.

My point here is epistemological, and I do not want to suggest that these limits also determine what really must be in your mind when you act intentionally. So this admission is not to give in to the temptation to follow Dennett's advice and abandon the snarl by replacing the ontological question with an epistemological. How far from my assertions and aims can your outlook be before I must fail to see even a minimal reason guiding your action? That is a question about epistemological constraints on my interpretation. Regular experiences of attempts to predict and explain other peoples' doings are important conditions for my rationalisations of your behaviour. Attempts at viewing myself from this perspective count as well. Together with imagination, such practice will make up my interpretative capacities to a large extent. Though the point that there *is* a limit of this kind is epistemological, the question of where it will be drawn is largely an empirical matter.

The question of where I draw the limit is *largely* empirical and dependent upon my previous experiences, but not wholly. As a reader of your behaviour, I am also able to *imagine* a gap between your motivating reasons and mine, even when, on the face of it, you act unreasonably and I have difficulties in determining any intentional explanation. In those cases, I must decide whether my scheme of interpretation is applicable to you at all for the moment. In doing so, I might be more or less liberal.

One of the questions I need to settle in order to distinguish the inner states that cause and rationalise your behaviour, is how peculiar a desire could be, before I approach the limit for something that possibly could rationalise your behaviour. If I know some of your ends, and some of your beliefs about how to reach them, certain types of behaviour of yours can mostly be ruled out as unintentional. Some instrumental desires are to be expected on the assumption of a minimal amount of rationality; others are incompatible with your goals and cannot be there to rationalise what you do, then.

When it comes to intrinsic desires, matters become more complicated. Some desires for things in themselves are such that my empathic limitations leave me without ability

to understand them. Take retribution. I know that in a rash of anger, I may for a moment desire that an injury is balanced with another. Many people apparently have a non-violent and stable sense of justice that involves this kind of balance for its own sake all the time. In calm moments, I am unable to recall this attitude. Behaviour out of such a desire, even if it is my own, as viewed from a third person perspective, approaches the limit for what I am able to grasp. Nevertheless, I know that there may *be* an intrinsic desire of this kind. Just as there might be weird intrinsic time-biases, like the preference for great pain on Tuesdays before mild pain any other day. Should I impose *any* rational restrictions on what you can desire as an end?

Michael Smith wants to tie 'direction of fit' intimately to a functional or dispositional concept of desire. On my interpretation, this way of thinking appears to fit well in with a Humean norm of practical reason, of the kind normally associated with the BD model. In a sense it is correct that the 'directions of fit' approach meshes, i.e. harmonises, with the dispositional analysis of desires. This is merely because it expresses a norm of rationality, which is part of a tool for identifying intentional states on the dispositional account. An intentional state's direction of fit tells us something about the role it *should* play in relation to external stimuli and other intentional states. To state that desire's direction of fit is world-to-mind is to make clear that we should not expect intrinsic desires to be adjusted to evidence. Desires do not aim at truth.

The thinnest possible sense of practical reason is the one required to make your action intentional. To adopt the direction of fit metaphor is to allow for a certain amount of relativism in this thin form of rationality — to let your ultimate aims be your own business in this respect.

Hence, the metaphor expresses a Humean norm of practical rationality conventionally associated with the BD-model — in Hume's words, that reason *ought* only to be the slave of the passions. The fact that the metaphor is inescapably normative does not make it improper, but it makes it disqualified as a premise in a conclusive and non-question begging argument for the necessity of desires in motivation. In other words, the characterisation of desires as states, with which the world must fit, is not an essential part of our (Humeans and non-Humeans alike) pre-theoretical thinking about motivation. So, the possibility of categorising motivational states on account of whether they are supposed to fit the world, or the world is supposed to fit them, does not constitute independent non question-begging evidence for the BD model. Nor does it yield a substantial descriptive account of the nature of the model's two categories.

Van Roojen is right in claiming that the BD model ("the Humean theory of motivation") does not provide us with an *independent* argument for subjectivity or relativism about practical reason (1995 p.53). But to accept the direction-of-fit characterisation of the BD model's two categories is to endorse an instrumental view of practical rationality, which is weakly relativistic on its own.[7]

Still, insofar as the idea of fitness-directions appeals to our linguistic intuitions, it indicates that we are inclined to psychologise in BD model terms. If we find it plausible to categorise any motivational state as having one of the two directions, or at least as constituted by (theoretically) detractable states of this kind, this gives the BD model and the Humean norm of practical rationality an outstanding position within common-sense theory of action. At least since Hume, philosophers have been drawn to conceptual analyses of this kind. (Since Aristotle, according to Brink.) This is some evidence for the empirical assumption that people generally are apt to think about behaviour in BD model terms. In turn, this gives us a reason to believe that the scheme really has proved effective in predictions and explanations. I am inclined to think that these kinds of evidence are the best we can hope for, when it comes to the ultimate question: Does the theory truthfully depict the ontology of motivation?

Summary of 1.3.2: The popular characterisation of beliefs and desires in terms of their different "directions of fit" is unavoidably normative. Appeal to this metaphor will not provide us with a non question-begging argument for the BD model, as some have thought. The metaphor harmonises with the BD model because it reveals a Humean norm of practical reason. That norm is not a part of the BD model's internal structure, but it fits well in with the tools we need to impose on others in order to identify desires on the BD model's dispositional analysis of such intentional states.

1.4 Focus on 'Desire'

The key concept to investigate in order to mark out the BD model from its rivals is 'desire'. The BD model treats the idea that desires are necessary to produce actions as a conceptual truth. This elevation of desire is what many of its critics doubt. The model lets desires and beliefs make a joint contribution to any reason-governed behaviour. Most BD model theorists would probably argue that even basic non-instrumental intentional actions require some belief about what the action will result in.[8] On the other hand, someone might argue with Hume that the impulse from a desire "had it operated alone, would have been able to produce volition", and still maintain that the BD model is essentially correct (1739 2:3:3). The question of whether agency always requires belief will be left open. The traditional controversy is over whether beliefs could be sufficient to motivate intentional action, not whether they are necessary.

My ambition in most of part I is to make explicit the philosophical assumptions about desires, which are embedded in the BD model. I will avoid presenting any explicit views about the concept of belief. However, a characterisation of the model's 'desire'-

concept can not be without consequences for its notion of belief. The functional role given to desires by the BD model will yield an interlocking conception of the role left to beliefs. It is e.g. obvious that if desires are necessary as driving forces in all motivation, then beliefs must, on their own, be motivationally inert. This does not mean that the model's definition of 'desire' implies a corresponding *definition* of belief. In other words, although the following presentation of the BD model's notion of desire commits me to some assumptions about the functional role of beliefs within the BD model's framework, it does not imply any full-fledged analysis of 'belief'.

When you assign a desire to, say, your cat, in accordance with the BD model, you endow it with an intentional state, a real inner driving force with content. The possibility of other types of intentional states with a non-propositional content is not excluded here, but desires are directed towards states of affairs. In other words, desires are propositional attitudes. It is compatible with the BD model to analyse the content of desires functionally. The claim about content does not presuppose that desires are present "before the mind" but the BD model nevertheless commits us to the idea (under attack by Daniel Dennett, his predecessors and followers) that content somehow is represented within the system.

Your criteria for assuming that your cat desires something are, ultimately, behavioural. The cat behaves as it does because of its desire, and its desire is what makes sense of its behaviour. But the BD model differs from pure behaviourism in giving desires the status of real causal forces. It understands desires in terms of dispositions for behaviour, as causes of actions, and as intentional states rationalising actions. In the following, I will argue that these claims really can be made to fit together.

Summary of 1.4: 'Desire' is the BD model's key concept. Although the functional role given to desires will entail an interlocking view of belief's function, it is not necessary to commit the BD model subscriber to any full-fledged analysis of 'belief'.

---

[1] Brandt suggests that the von Neumann/Morgenstern formula for assigning utilities on account of of choice behaviour, could be used to measure the "valence" of an outcome, while valence at the same time is the resultant of the agent's aversions and desires, realistically understood. (1979 p.51.) This was brought to my attention by Magnus Jiborn. Nils-Eric Sahlin is of the firm opinion that Donald Davidson's theory of motivation (by many regarded as the first formulation of the BD model — at least if the standard interpretation of Hume should appear to be incorrect, as it has been argued; see ch.7) in effect expresses a standard decision-theoretical analysis. (Lecture at Lund University in the eighties, confirmed in personal conversation 1999)

[2] E.g. in the introduction to Gärdenfors' and Sahlin's widespread anthology and textbook *Decision, Probability and Utility*, the editors stress that the common core of the contributions is "that there are two main types of factors determining our decisions. One is our *wants* or *desires*. /…/ The other is our *information* or *beliefs* about what the world is like and how our possible actions will influence the world. /…/ The main aims of a *decision theory* are, first, to provide models for how we handle our wants and our beliefs and, second, to account for how they combine into rational decisions." (1988 p.1). After discussing von Neumann's & Morgenstern's technique for utility assignment, Alvin Goldman states that although "social scientists often disclaim any connection between their concepts and psychic states, it is nonetheless clear that their models presuppose a common sense want-and-belief model of human behavior as their underlying intuitive foundation." (1970 p.137)

[3] Hausman's comment is a conclusion of his criticism of defining 'preference' in terms of choice behaviour, and I find this conclusion just as applicable to Dennett's version of behaviourism.

[4] Aulisio (1995 p.353) credits Smith for designing the *phrase* "direction of fit" in (1987) and tracks the expression "relation to fit" to Searle (1979). Smith (1987 p.51) quotes Platts' use of "direction of fit" in (1979), and appears to ascribe this terminological invention to him. Although Smith makes no explicit claims about the history of the phrase, he appraises Platts (who refutes the distinction) for making Anscombe's idea "sound so plausible". Searle employs the expression "direction of fit" in his "A Taxonomy of Illocutionary Acts" from 1975, as the quotation above shows.

[5] Michael Smith assigns P1 in the present formulation to Donald Davidson, with reference to "Actions, Reasons and Causes" from 1963. (Smith 1994 p.92) However, in that paper, the explicit formal principle which comes closest to Smith's P1 reads:

"C1. *R* is a primary reason why an agent performed the action *A* under the description *d* only if *R* consists of a pro attitude of the agent towards actions with a certain property, and a belief of the agent that *A*, under the description *d*, has that property." (Davidson 1963, p.5).

Though this formulation is *roughly* equivalent to P1, it is even more clear that the principle cannot be concluded from Smith's three premises, if it is put in Davidson's original form. Neither (a), (b) or (c) entails anything about pro-attitudes being directed towards a certain property of the action, for instance.

[6] I am grateful to Wlodek Rabinowicz for reminding me of the importance of Davidson's view on the principle of charity in this context, and for his suggesting a relevant quote.

[7] I think that van Roojen too quickly assumes that this weak form of relativism about practical reason easily leads to *ethical* relativism as well. His point is that the instrumental account implies that people with true beliefs in similar circumstances will have reason to do different things. But there are numerous other ways of condemning what a person does, than by labelling it irrational. Such condemnations may of course in themselves be expressions of intrinsic aims.

[8] See e.g. Persson 1981, sections 6.1-2.

2. Metaphysics of Desire

2.1 Reducible Dispositions and Real Desires

It is central to the BD model that desires are regarded in terms of dispositions for behaviour, and at the same time as distinguishable causes of behaviour. That combination of claims is common in philosophy of action as well as in moral philosophy. Donald Davidson and Richard Brandt explicitly advocate it, and, just to mention one more example, Philip Pettit and Michael Smith also accept this combination. (Davidson 1980, Brandt 1979, Pettit & Smith 1990.) Preference-utilitarians, like R.M. Hare and Peter Singer, often seem to depict desires as action-dispositions in some cases, and as real inner driving forces in other.

Taken together, the two claims appear to imply that dispositions cause their displays. In *The Moral Problem*, Michael Smith writes, e.g., that the dispositional account of desires "does not commit us, as Humeans, to the thesis that desires are to be conceived of as the causes of action. For it is a substantial philosophical thesis to claim that dispositions can be causes". Smith stresses, though, that his own view is that "dispositions, *and so* desires, are indeed causes" (1994 p.114 my emph.)

My primary ambition is, I may remind the reader, to present what Austin calls a linguistic phenomenology, i.e. to expose the thoughts revealed in our daily talk, at present concerning dispositions and causes in relation to desires. We think of desires in both dispositional and causal terms, in line with the conventional BD model. At the same time, in many cases we regard it as nonsense to bring forward a disposition (like brittleness) as an explanation of its manifestation (like breaking under certain circumstances). A secondary ambition is to give a consistent interpretation of that apparent clash.

Even if it is coherent to talk about dispositions and desires like this, there could be other plausible objections to the ontology revealed in daily speech. On the other hand, it seems relevant to examine linguistic practices in search of a more substantial ontological standpoint. D.M. Armstrong, who seems to pay little attention to ordinary language in other metaphysical debates (1968, p.14) thinks that the status of dispositions can be dealt with by studying how "scientists and others often speak" and examining what is "linguistically correct".

2.1.1 Dispositions Do Not Cause their Displays

Some elements in the characterisation of dispositions are uncontroversial, e.g. that a disposition is determined by some triggering conditions and an

effect characteristic of the specific disposition. In a philosophical standard example like "the ice is brittle", the assumed triggering condition is, roughly, "the ice is being tread upon" and the effect "the ice breaks". In the dispositional account of "I want a cup of coffee" the triggering cause is supposed to be something like "I come to believe that I can get a cup of coffee" and the effect is "I tend to get a cup of coffee" (time-variables, ceteris paribus clauses etc. excluded).

Dispositions may truthfully be assigned to objects in the absence of triggering events and triggered displays. Therefore, in a sense, dispositions have a "teleological" or "intentional" character: They point beyond themselves, towards something non-existent. For this reason, Gilbert Ryle and other philosophers associate some realistic views of dispositions (polemically) with occultism. It seems implausible that the physical world could be inhabited by directed powers.

Ryle himself appears to view most disposition-statements as nothing but "inference-licenses". They are pure hypothetical predictions, free from commitments about existent states. Applied to desires this means that when we assign desires to people, we are not making claims about their internal driving forces. We are merely making predictions about their behaviour in hypothetical circumstances. And this view of dispositions is what paves the way for Ryle's own eliminative view of desires and other alleged inner causes of behaviour.

There is, however, a greater difference between disposition-statements and hypothetical predictions in general than Ryle's account admits. When we e.g. say that something is brittle, we are committed to claims about its internal state or structure. As Prior, Pargetter & Jackson argues in "Three Theses About Dispositions" (1982), it does not make sense to speak of dispositions without presupposing causal bases. This is a version of their argument: Suppose we try to characterise the dispositional property "fragility" by exemplifying: "If the ice is fragile it will break when it is walked upon." Now, suppose our world is deterministic, and a certain ice displays its fragility by breaking when it is walked upon as in case 1. Then we can not imagine that the ice does not break when it is walked upon in case 2, provided that the internal properties, molecular structure etc., of the ice are similar, and all other things (like the weight of the walker etc.) are kept equal. If it does not break in the second case, although the external features of the situation (the weight of the walker etc.) are exactly similar, we would assume that the internal properties of the ice had changed. And the point is that such an inference about its changed internal properties is the kind of justification we could have for assuming that there has been a change in fragility.

This argument appears to work for an indeterministic world as well. Suppose we know that all things are equal, including the internal properties of the ice, and the ice breaks in case one. Now our task is to decide how fragile the ice in case two really is, i.e. we must determine the risk of

breaking for ice of that kind with that internal structure, within the limits of risk-determination in an indeterministic world. Even then it seems unreasonable not to take into account what happened in case 1. The apparent soundness of bringing up case 1 must stem from our conception of fragility as dependent upon a causally relevant internal structure, which happens to be fixed in this test case. So, in order to be fragile, an object must have an internal property, such that it breaks under the kind of conditions specific for this specific disposition.

Perhaps one should add that the causal base of the disposition is supposed to be non-relational or internal to the object possessing the disposition. We assume that there is something about the ice itself guaranteeing the truth of the conditional statement. The Devil's resolution to shatter a certain ice every time it is walked upon would not be sufficient to make the ice fragile.

According to Hugh Mellor's article from 1972, "In Defence of Dispositions", it is true that "dispositions require some independent basis for their ascription between displays - but the basis need only be another disposition." I am prepared to agree. The requirement that the disposition's supporting properties are internal to the disposed object does not exclude the possibility that they, in turn, are dispositional. I am merely prepared to conclude that we regard it as nonsensical to speak of dispositions as causes of *their own* manifestations. That position could be compatible with supposing that the state causally responsible for the manifestations of the disposition of a certain object (say, the state α responsible for the breaking of brittle ice) is *another* disposition (like the disposition of some of the micro-components of the ice to disentangle when affected by other micro-structural changes in the environment). The molecular bonding of fragile ice can perhaps not be described properly without use of dispositional notions. The latter disposition would then be a fact about what can be conditionally predicted of some components of the ice, on account of the inner features of those components. *That* dispositional fact about inner features of the ice could explain its brittleness, but that is not a fact about a state formally identified by its relation to the breaking of brittle ice.

Similarly, Richard Brandt has characterised the state underlying wants and aversions as "a readiness of certain neuron-sets to fire" (1979 p.25). That characterisation need not be formally troublesome. The important thing is that the underlying properties referred to are internal to the object in the sense that they are assumed to be specifiable, in principle, without reference to external displays of the very disposition attributed to the object. In the examples mentioned; ruptures in the ice or my getting a coffee break.

This possibility would mean that changes in dispositions might make a genuine causal difference. Note that this is not only meant in the indirect sense that change in dispositions of an object implies change in the intrinsic

state, which is a causal condition for the manifestation of the disposition. Dispositions are not only causally *relevant*, they can perhaps also be causally *efficacious*. Simon Blackburn points out that it is not even evident that this kind of explanation must have some final non-dispositional endpoint (1991, p.196). In line with this, Tim Crane has made clear that such an endpoint is excluded by the view (which he defends) that all properties must be analysed dispositionally.

Another reminder may be in place here. My proposals concern our daily speech. When someone claims that a thing possesses a dispositional property, he is not only committed to a conditional prediction, but also to an assertion about an intrinsic causal base of the disposition's displays. However, we do not presuppose that the base he has in mind must be non-dispositional. Blackburn's and Crane's regresses could raise other problems, but they are not at odds with our daily use of dispositional explanations.

According to one common view, the disposition is the intrinsic causal condition for a disposition's manifestation. The "is" can not be understood as "is necessarily identical with". Suppose that fragility of ice in our world is based upon molecular structure α. Now we can always imagine a world in which all the relational properties of the fragility of ice (breaking under the right kind of circumstances etc.) were attached to β rather than to α. If the fragility of the ice and its causal basis were necessarily identical, we would be forced to say that fragility of ice is another property in that world than it is in ours. That seems odd (as Prior, Pargetter and Jackson point out, 1982) since we think of names of dispositions, like names of other properties, as denoting the same type of thing in all possible worlds. The disposition and its base can therefore not be identical in a sense making it self-contradictory to separate the disposition from its specific base. Though every disposition necessarily is related to an intrinsic causal base of a kind specific for the disposition in question, the relation between each type of disposition and its specific type of base is contingent.

Those who nevertheless believe that the disposition is reducible to the causal base of its displays must therefore suppose that the identification between brittleness and property α is contingent. This is e.g. assumed by D.M Armstrong (1968, ch.6) and Ingmar Persson (1981, p.44). I must confess that I am still uncertain about how that is to be understood, exactly.

One suggestion, by U.T. Place, is that the type-identity between brittleness and micro-structure α can be called contingent if the two terms refer to the same property, but this co-reference is linguistically opaque, and a matter for science to illuminate. Water is in this sense contingently identical with H_2O – or perhaps *was*, before the knowledge was incorporated into language, and the terms became synonymous to most speakers.

> In science, type identities which are contingent hypotheses when first formulated become necessary truths when the conventional criteria for assigning instances to universals begin to change so as to incorporate the empirically discovered 'real essence of a natural kind' into the meaning of the words and expressions of natural language. (U.T. Place 1996 p.59)

Armstrong's own and more cautious label "contingent identification" is more proper than Place's "contingent identity" for what Place has in mind. Literally speaking, the contingent relationship here is not the identity between properties, but the correspondence between linguistic practice and natural demarcations.

Identity-statements, which are true and informative, contain non-synonymous terms referring to one and the same thing. If the words pick out one separate specimen (as opposed to a type), their co-reference might in many cases be called contingent in the sense that the roles or properties used to pick out the individual need not necessarily come together. It is a contingent fact of this kind that "Silvia" and "the Queen of Sweden" refer to the same entity. (Since there is one and only one individual such that this individual is Queen of Sweden, and this individual is Silvia, it would nevertheless be misleading to say, as Place's explication seems to imply, that the *identity* between Silvia and the Queen of Sweden is contingent).

But statements identifying *types*, like "water=H_2O" are usually supposed to express necessary truths. Brittleness and α can hardly be identical in this speech-hypothetical sense. This is what Place wants to show as well, against Armstrong. Even if science proves brittleness to be paired with microstructure α, and that knowledge becomes a part of natural language, we will still not have any difficulties in imagining that it is based on β rather than α. Place's view seems to be that this is the *only* way in which identity between types can be contingent. With that presupposition, there will be no meaningful concept of identity admitting the 'essence' of things to be such that two types of properties are identical in our world but separable in another possible world. If such hypothetical co-reference was what Armstrong's use of 'contingent identity' was supposed to express, the point above would suffice to show that his hypothesis is wrong.

However, Armstrong makes clear that the "contingent" element in his identification does not just mean that the co-reference is a scientific hypothesis, waiting to be part of common knowledge. "It is not like the *a posteriori* identification of the heat of a substance with the motion of its molecules, an 'identity of property constitution' where, [I agree] with Kripke, the identity is necessary." (Armstrong 1996 p.39)

According to Armstrong, brittleness can be picked out via causal role — triggering conditions and expected effects — or by exposing the microstructure of the disposition. In a world with our laws of nature, its causal role is filled by a certain microstructure. Our notion of genes is adduced as an analogy. Genes are, according to Armstrong, dispositions to

develop certain physiological or psychological traits. And we know that genes are identical with sequences of chains of DNA. Similarly, it is "a contingent truth that in this glass the brittleness role is played by the microstructure, i.e. that the brittleness of this glass is (is identical with) this microstructure" (1996, p.39).

We are supposed to identify brittleness with microstructure α because the two terms refer to a state with a certain causal role. The role of brittleness/α is to explain the manifestations of brittleness. Armstrong's argument for this identification is a generalisation about how we talk. He argues:

> that it is linguistically proper, for instance, to say that brittleness *is* a certain sort of bonding of the molecules of the brittle object. The ground for saying this is simply that scientists and others *often speak* in this way, and there seems to be *no objections* to such speech (1996 p.14 my emph.)

In a similar way, Hugh Mellor writes that "the proper rôle of dispositions is to explain their displays: its fragility is what is supposed to explain the breaking of a dropped glass" (1974, p.117).

I do not believe that Armstrong's and Mellor's linguistic observations apply generally. Mellor appears to have changed his mind on this subject as well[1]. Scientists and others rarely speak in this way. Neither do I think that such speech would normally pass without objections. We do not, normally, refer to dispositions as explaining their own manifestations. More important, perhaps, is that we do not *accept* such explanations when they occasionally confront us in daily life. "The glass broke because it was fragile" or "I did it because I felt like it" are not considered genuinely explanatory (more than in very exceptional cases). Suppose science tells us that what makes ice break when it is walked upon is molecular bonding α. Would it not feel more natural to say that science then has found out what makes ice fragile, rather than to say that science has found out what fragility really is? To add an example from the social sciences: The Stanford psychology professor Walter Mischel says in his widespread *Personality Assessment:*

> Trait labels should not be confused with the antecedents and maintaining conditions of the behaviors to which they refer, nor with an accurate description of the behaviors themselves. In fact, this confusion does occur whenever trait descriptions are offered as explanations for behavior — for example, when inefficient, disorganized responses are "explained" by simply calling the performer neurotic. (1968 p.68)

It seems correct that genes are identical with (sequences of) chains of 'DNA' and that it is a contingent fact that DNA fulfils the causal role of genes. However, as U.T. Place points out, the example does not convincingly show that we point to dispositions in order to explain their displays. Genes are normally thought of as intrinsic properties, which are responsible for

dispositions (such as the disposition to develop certain physical traits) — and were so regarded even before we knew their biochemical constitution.

Armstrong touches upon the most apparent objection to such speech, when he explicitly admits that in identifying dispositions with the states that are causally responsible for their manifestations, "we /.../ expose ourselves to Moliére's ridicule." To put it bluntly: If the ice is fragile, it has, by definition, some non-relational feature that causes it to break — but if the fact that opium possesses *virtus dormitiva* cannot explain why it makes us sleep, then fragility cannot explain why the ice breaks. Nevertheless, Armstrong argues, it is linguistically correct to identify the fragility of an object with the causal conditions for its breaking.

Moliére's joke in *Le Malade Imaginaire* appeals to most sensitive language-users. When we come to understand that *Virtus Dormitiva* simply means sleep-producing powers, then we immediately see how Moliére's pharmacist, or perhaps the science he represents, is ridiculed: His explanation is empty. Were dispositions commonly thought of as legitimate explanatory factors, we would probably not appreciate the joke so easily. Admittedly, the joke also works because we recognise the type of fraud in question from science as well as from daily life. The point is that we recognise it *as* fraud.

An important reservation must be mentioned here. All proposals where dispositions are offered as explanations of their displays are not fake and uninformative like the pharmacists. In Moliére, the joke is that the pharmacist's answer gives no information above what is presupposed in the question. In this case, that opium tends to make us sleepy and that this has to do with its intrinsic properties. In other words, that opium possesses the dispositional property of being sleep-inducible. But in many circumstances we may have reason to attribute a disposition to an object in order to exclude rival *external* explanations of its behaviour. Since statements about dispositions are not merely inference-tickets or conditional predictions, but also indicate an internal causal basis, they are useful in cases where we can not take it for granted that what happens to the object is due to its inner constitution.

Should I find that your ugly vase (a birthday present from me) is broken when I come to visit you, I might want to know if its fragility was the important factor in its breaking. Should you kick my lower shin under the table, I would be interested in knowing whether you really wanted to make my leg hurt. "It broke because it was fragile" or "I kicked you because I wanted to" would be explanatorily forceful in proportion to the number of available alternative explanations in terms of other things than the vase's or your mind's constitution. Have I got reason to fear that explosives or hammers have been involved in the shattering of my gift? Is there a chance that your attack on my shin was the result an accidental twitch? If such alternatives do not seem farfetched, it might be useful to exclude them by pointing to the internal properties of the object in question.

The *Virtus Dormitiva* objection shows that dispositions rarely explain their manifestations in a valuable way, but it does not prove conclusively that dispositions cannot cause their displays. It is analytically true that the cause of my sleep is what causes me to sleep. If α refers to the cause of my sleep, it is therefore uninformative to cite α as an explanation of my sleep, although it is true that α caused me to sleep. So even if the dispositional term "sleep-inducible" referred to the property causally responsible for the presupposed effect, it would refer to it in an uninformative way, in line with Moliére's joke.

Nevertheless, our way of understanding the joke undermines the main argument for identifying dispositions with the bases of their displays. That argument presumed ordinary speakers and scientists to find a report about an object's disposition as an informative causal explanation of whatever the object is disposed to do. But that is not how we think about such reports. We do not find a genuine explanatory value in those cases where external explanations are excluded to begin with (as in the question to the pharmacist). Even minimal information about the internal features of the ice — like that it *has* micro-components — teach us more about why it tends to break when we step on it, than information about its fragility (given that external explanations are precluded.) As Campbell and Pargetter puts this point:

> [W]hen we say that a fragile object breaks *because* it is fragile, we do not mean that *fragility* causes the object to break. Rather, the "because" indicates that the fragile object has a nature which is such that it is causally responsible for the object's breaking (Campbell and Pargetter 1986 p.161)

So, in everyday speech there is no support for thinking that dispositions explain their manifestations. Their explanatory value cannot be offered as an argument for assuming that they cause their manifestations. On the contrary, we treat such reports in a manner indicating that dispositions do *not* cause their displays. Although disposition-attributions entail attributions of internal causal bases, ordinary language observations give us no reason to assume that dispositional labels *denote* these bases.

Prior, Pargetter & Jackson defend the position I attribute to ordinary speakers here, i.e. they deny that dispositions are explanatory factors. They argue the other way around. While I claim that the apparent explanatory impotence of dispositions is a reason to assume that we regard them as distinct from their causal bases, Prior, Pargetter & Jackson argues that the distinctness of dispositions is a reason to assume that dispositions are impotent. Firstly, they establish that every disposition must have a causal basis. Secondly, they show that since names of dispositions are rigid designators, dispositions cannot be identical with their causal bases. Since one event cannot have more than one sufficient operative condition, it follows from their first two assumptions that dispositions do not cause their manifestations. The causal base of the disposition is the operative condition

for the disposition's manifestations. Hence, dispositions do not explain their manifestations.

Their argument presupposes that the identification of basis with disposition is seen as necessary. Otherwise the point about the rigidity of property names is insufficient to prove that dispositions are distinct from their causal basis. Furthermore, it could be questioned whether the defender of independent causally powerful dispositions would have to assume that they must be causally *sufficient* to produce their manifestations. With Tim Crane, one might suppose that the disposition's displays require the presence of both base and disposition, besides characteristic triggering circumstances. "Given the disposition — fragility — and the circumstances, then the microstructure will bring about its effects". (Crane 1998 p.221)

2.1.2 An Elimination of Dispositions

There are two sides to our talk about dispositions. It expresses conditional predictions and it points to the internal causal basis warranting that prediction. Judgements about dispositions can not be replaced with pure hypothetical predictions. Strict assertions about inner properties can not replace them either. I have argued that we do not think of the role of dispositions as that of causally explaining their displays. Thereby I have also excluded a type of realism about dispositions, which U.T. Place and Tim Crane (like Mellor 1974) wants to re-establish; That "brittleness" refers to a distinct, non-categorical, irreducible and causally potent property, which *besides* microstructures and triggering conditions explains the cracks of the ice. In other words, that dispositions really exist in our universe as the kind of entities Ryle pokes fun at — objects striving or pointing towards non-existing goals.

However, I have not excluded the realistic hypothesis defended by Prior, Pargetter and Jackson. They argue that "brittleness" refers to an irreducible second order property, devoid of causal power, *supervening* upon the causal basis of its displays. 'F is supervenient on α' might state as a conceptual truth that two entities cannot be exactly similar with respect to α without being similar when it comes to F. Suppose fragility of ice depends upon microstructure α. Since it is *imaginable* that α comes without fragility, the disposition does not supervene upon α in this conceptual sense.

It may be conceptually true that if an entity is fragile, then it has *some* inner property, in virtue of which it has the property of being fragile. It is probably only this weak supervenience claim Prior, Pargetter and Jackson has in mind. But would that underpin an assumption of this kind: "No two objects could be identical in all respects concerning basic properties, and yet one being fragile and the other not." (Cambell and Pargetter 1986 p.155)

It is conceptually possible that α, which actually is responsible for a certain object's fragility, fails to make that object fragile. 'F is supervenient on α' must therefore, in this case, be taken as an ontological assumption to the effect that, although it is theoretically possible that α comes without F, it is impossible in some other sense, perhaps metaphysically. The need for clarification on this point arises because fragility is supposed to be an *existent* property, distinct from its base. If "fragility" named no existent property at all, but functioned as a Rylean inference-licence, warranted by facts about internal causal conditions, then it would be easier to make clear why, and in which sense, both objects must be fragile if they have similar basic properties.

A sufficiently weak supervenience claim could probably be made to fit with what I have assumed so far. However, although I regard the arguments Prior, Pargetter and Jackson present (for dispositions necessarily having intrinsic bases and against identifying dispositions with these bases) as sound, I believe that these arguments would be compatible with a more polished ontology.

In other words, I am tempted to go back to a position close to Ryle's. Why should we think of dispositions as distinct, real and irreducible? Perhaps Ryle is right in viewing disposition-statements as reducible to hypothetical predictions, with the important addition that disposition-statements are distinguished from hypothetical predictions in general, by a certain implicit presumption about which kind of facts the truth of the statement is allowed to be warranted by. I.e. facts concerning the internal properties of the object that is disposed in the described way. As Ryle plausibly argues,

> there is nothing scandalous in the notion that a statement may be in some respects like statements of brute fact and in other respects like inference-licenses; or that it may be at once narrative, explanatory and conditionally predictive, without being a conjunctive assemblage of detachable sub-statements. /.../ Nor is such a statement one of which an objector might say that part of it was true, but the other part was false. (p.141)

This analysis would, then, turn dispositional statements into semi-hypothetical or mongrel categorical statements, in Ryle's terminology. They are handy labels on a combination of complex conditional predictions and hints about where the causal warrant of these predictions can be found. The view is in a way more realistic than Ryle's, since it stresses that you are committed to a view about the object's inner constitution when you attribute a disposition to it. On the other hand, it is less realistic than Armstrong's notion. It does not say that the name of the disposition *denotes* this inner state. Strictly speaking, the suggestion is eliminative, since it presumes that everything we say when we talk about "brittleness" in principle could be put in terms of non-dispositional terms. (This elimination of dispositional terms would not be affected by the possibility that the base

of a disposition is another disposition. Disposition-assignments do not describe the internal base, they just assert that there is one.)

The eliminative position, in combination with my hypothesis about an inherent ambiguity in disposition-talk, could also explain our intuitions on a relevant phenomenon noted by Peter Menzies, who regards the case in question as inconsistent with the view that dispositions are to be identified with their bases. Tim Crane appeals to Menzies' point in arguing for dispositions as independent causes of their own displays, while Frank Jackson and David Lewis discusses the same example in arguments *for* disposition=base. In Crane's words:

> It turns out that the categorical basis of the opacity, thermal conductivity and electrical conductivity of [some] metals are the same: it is to do with the way free electrons permeate the metals. But the effects of opacity, ductility and electrical conductivity are distinct. (1998 p.219)

And Crane goes on, quoting (a yet unpublished paper by) Jackson:

> The transmission of a telephone call down a wire is explained by its electrical conductivity, not by its opacity. Good results from cooking in a copper-based saucepan are explained by copper's high thermal conductivity, not by its opacity or electrical conductivity. (1998 p.219)

Jackson, like Armstrong and Lewis, thinks that these effects are all really explained by the same property, while Crane prefers to say "that in this case, the dispositions themselves are the genuine causal agents. This means that dispositions can be causes distinct from their categorical bases, if they have any." (1998 p.220)

To begin with, it is important to note that my account admits talk about dispositions as causes – though not as causes of their own manifestations. I claim that assignments of dispositions, like 'this saucepan is disposed to conduct heat' might be elliptical ways of expressing conditional predictions along with hints about the causal warrant of those conditionals. It could nevertheless make sense to say that the thermal conductivity of your saucepan was a causal condition for the good results of your cooking. A condition for that conductivity might, in turn, be stated in terms of how free electrons are disposed to pass through the material, and so on.

The identity-theorists dilemma is that if he says that these effects are all explained by one property, he must regard the separateness of thermal conductivity, opacity and electrical conductivity as an illusion. These dispositional terms would have to be co-referential, and that seems implausible. On the account I have proposed, we would, like the identity-theorist, have to regard the displays of all three types of dispositions as explained by one common type of condition. But our distinct uses of the three dispositional concepts would in no way be threatened by that possibility. We employ them in order to stress that there is some fact about

the material, which warrants certain conditional predictions — about what will happen when we heat it, or if we charge it with electricity. Though these conditionals are distinct, they are warranted by the same internal condition of the material. They happen to be simultaneously true about any material with that internal constitution, if the physical story is correct.

That appears to be in line with how we think about such a case. It appears plausible to say that what we have learnt from Menzies' description is that if a certain material is conductive of heat, it must also be opaque and conductive of electricity (although displays may be absent). When we come to know that one and the same base warrants all three assignments, we infer those two dispositions from the first. Crane's approach appears to do little justice to the latter intuition. According to Crane's hypothesis, a disposition to conduce electricity appears to be necessary *besides* microscopic properties, to cause a piece of metal to conduct electricity. Neither base, nor disposition, is sufficient on its own (1998 p.219). But in that case, we will not be allowed to infer an assumption about electric conductivity from our observation of thermal conductivity of a certain metal. Not even after the physicist has taught us that the inner constitution of the metal is such that if heat passes through it unhindered, then so does electricity.[2]

Reductive analyses of the notion of dispositions tend to heel over, either towards the teleological side (the conditionally predictive function — Ryle), or to the categorical side (the inner state-assertive function — Armstrong). Both sides are essential elements in our talk about dispositions. Furthermore, I now want to argue, even in daily speech we waver between expressions stressing one or the other of these two sides. Sometimes we really use "fragile" and "soluble" as if they were "physical predicates on a par with others, and the dispositional form of the words is just a laconic encoding of a relatively dependable test or symptom" (Quine, 1992 p.76). Just as common, however, is a way of speaking in which our *primary* purpose is to bring forward an assertion about a foreseen symptom, though an implicit reference to the physical conditions for that symptom to obtain is offered in the bargain.

2.1.3 Reasonable Hypostatisation

Suppose we pass a skater on a pond and you remark that the ice is brittle. Your primary purpose is then probably not to explain anything. Instead, you want me to note a possible risk. Unless the skater is very light, or very careful, the ice is likely to break under her. Should we pass a man in an ice-hole and you tell me the same thing about the brittleness of the ice, I would take your comment to be meant in an explanatory sense. It is reasonable to think that you are presenting a causal hypothesis to me. The man is no winter bather, as I might have thought. The hole is there primarily on

account of internal properties of the ice, which are such that no extraordinary circumstances are required to make it crack. Judgements about dispositions are put forward to explain, or to predict. When used explanatorily, their value is proportionate to the number of rival external explanations in the context.

The joke about virtus dormitiva works, I argue, to the extent that external alternatives are excluded from the start. If they are, nothing is added by asserting that the object is constituted so that it can be expected to behave in the way we need an explanation of. But if other explanations are within reach, it might be important to mark where the cause should be sought.

Conventions of normal speech situations are such that I have a right to expect that the things I am told in some way are relevant and informative to me. At least I should expect speakers mastering those conventions to tell me things they *believe* to be of interest to me. This expectation explains why truisms and insinuations are effective methods of cheating without lying. If you tell me that B showed up sober and in time for work yesterday, you make me think that this is an exception. Otherwise, why would you tell me? In a similar way, the conventions of verbal communication give me the right to assume that there really are rival external explanations of the object's behaviour, when you point out that it has a disposition to behave like that — provided that it is clear to me that your utterance is supposed to be explanatory rather than predictive.

Language offers markers revealing whether the primary purpose of a certain judgement assigning a disposition to an object is explanatory. Roughly, I mean that the hypostatising or "objectifying" way of referring to dispositions is a marker of that kind. We may speak about *the* brittleness, the want, the desire or the virtus dormitiva as if those dispositions were independent entities inside the object. That is useful if we want to stress the need for exclusion of external explanations. We objectify dispositions when we want to stress the causal basis of the displays of the disposition, rather than its inherent conditional prediction.

Hypostatisations are often regarded as symptoms of ontological confusion. When Daniel Dennett presents his by now well-known analogy between 'fatigues' and "beliefs, desires, pains, mental images, experiences — as all these are *ordinarily* understood" (1979 p.xx), his ironic parallel serves to illustrate his thesis that nonsensical ontological commitments are embodied in our traditional ways of framing problems about the mind. In our way of asking, we hypostatise or objectify unreal entities. The analogy is simple and somewhat amusing. This depends, I believe, on the fact that we recognise hypostatising talk about dispositions from ordinary language. However, Dennett's diagnose need not be correct, not even in the imagined case with an established use of "fatigue". We do not have to assign an absurd ontology to speakers using these expressions. Suppose an insensitive linguist uses her child in an experiment. She consistently refers to his tiredness in terms of number, intensity and location of fatigues. "Fatigue"

will become part of this child's natural vocabulary, alongside with "tired". Even in this case I think that someone who listens to this child will be reluctant to assign a confused ontology to him. Instead it would be natural to interpret his use of "fatigues" rather than "tired" as indicating that he does not only want to tell us how we should expect him to behave. He feels, probably, that there is a condition for his behaviour to be found internally, perhaps even with a more specific location, like in his legs or in his head, for instance.

Natural dispositional concepts differ from each other in this respect. Some terms mostly carry with them this (pseudo-) objectifying function. In other cases, like 'fragility', this interpretation is rarely applicable. However, the context will usually give sufficient guidance. Suppose I tell one of my daughters to be careful with a glass and she asks me why. Because it is fragile, I answer. In that case I am attributing a dispositional property to the glass, the reducible property of breaking under certain presupposed conditions, on account of its intrinsic character. Suppose instead that I ask her why the glass broke and she answers by saying that it was because the glass was fragile. We could understand that as a way of hypostatising a causally potent distinct entity, of the occult kind disliked by Ryle. To eliminate that interpretation I could ask her "Do you mean that *fragility* was what caused the glass to break?" To me it does not seem unlikely that she will find this question silly. Perhaps she will come up with a clarification along the lines mentioned, perhaps "No, of course not; I mean that there was something wrong with how the glass was made, so that it broke too easily".

Hypostatising uses of dispositional terms will feel natural depending upon how common external explanations are. When an object breaks it is a very common procedure to assume that an important condition is to be found in the object's inner constitution. Information about brittleness will only be of interest in exceptional cases. It is a different matter when it comes to human behaviour. The minimal information that I did something because I wanted to would be substantial in many cases. E.g. where you could suspect that my action was a mere reflex, that ignorance made me make something I did not want to, or that external factors influenced the effects of my behaviour in an unforeseen way. For that reason, "desire" is a term more naturally used in hypostatising sense, stressing the internal causal basis of that which is to be explained.

Summary of 2.1: Judgements about dispositions assert conditional predictions and stress that these predictions are warranted by facts about the object's inner constitution. Natural language lets us choose which of these functions we want to stress when making such judgements. The resulting ambiguity is usually contextually dissolved. Furthermore, the hypostatising use of dispositional terms provides an elliptic way of pointing to the internal causal basis of the displays of a disposition. That kind of objectification is

quite legitimate, provided that external explanations are not already excluded in the context of explanation. In the latter cases, Moliére's ridicule is in place — such explanations are without force.

When we assign desires to people, we are not just making hypothetical forecasts about them. We endow them with existent internal states, causally powerful. 'Desires are causes' does not presuppose 'dispositions are causes', since the hypostatising notion of 'desire' presupposed in 'desires are causes' points to the causal base of the displays of the disposition.

2.2 Causal Tendencies

In the BD model's characterisation of desires, tendencies are supposed to play a role distinct from dispositions. If A desires that p, she is disposed to *tend* realising the state of affairs that p, and if she wants to ϕ, she is disposed to tend to ϕ. Desires are not attributed with an all-or-nothing commitment about the behaviour of the agent, given that the conditions for acting on the desire are fulfilled. B can have a more or less strong desire for p, and this strength should be revealed in the strength of his tendency to get p, when the disposition is triggered. In other words, 'tendency' may do justice to the intuition that it makes sense to speak of strength of desires in absolute terms. If tendencies are distinct from dispositions, they are not to be understood in terms of conditional predictions. And if the notion of strength of a single tendency is meaningful, 'tendency' cannot be read frequencially. Both of these interpretations of the term are in use, and I want to distinguish a third sense of 'tendency' from these two.

It appears common to regard 'tendency' and 'disposition' as near synonyms. Ryle treats the term like that and J.L. Mackie's view (in *The Cement of the Universe*.) appears to be similar (1974).

R.B. Brandt argues that we cannot understand 'tendency', which plays a central role in his analysis of 'valence', in terms of frequency. His reason is that "since we might want to say that a person had a strong tendency to perform a certain act at a time even if, when he has such a tendency, he seldom performs the act on account of stronger contrary tendencies." (1979 p.26) Unfortunately, Brandt never proceeds to account for the non-frequential concept he has in mind. His only positive suggestion is not, I think, very helpful. He claims that we can get at least an initial idea of the meaning of 'tendency'

> from just one of these laws: that an agent will actually perform an action *A* if and only if the tendency to perform *A* is stronger than the tendency to perform any other action *B*. (1979 p.26)

According to Ingmar Persson, there is an important difference between ‘disposition’ and ‘tendency’. A brittle object is always disposed to break easily, although it *tends* to break only in certain circumstances. A state is dispositional when

> its instantiation causes a certain effect when influenced in a certain way. An imputation of a tendency goes one step further in that it entails something about the prevailing circumstances as well, to wit, that they are such as to call forth a response if no countervailing factor turns up too. (1981 p.116)

Persson’s example is illuminating, but I find his explicit characterisation of the distinction too brief. It requires a way of separating the *state* picked out by the dispositional statement from mere ‘prevailing *circumstances*’ which are supposed to be a distinguishing mark of tendency-assignments. That might be problematic, since both are analysed hypothetically. They are seen as conditions for a result, which might or might not be obtained, depending on the occurrence of other conditions. In other words, Persson’s characterisation of ‘tendency’ is not sufficiently demarcated from ‘disposition’. Let me attempt to strengthen that demarcation.

A brittle object, like the coffee-cup in front of me, has, through its existence, a tendency to break. It may also, now, tend to break. Both these truths can be covered by ‘The cup tends to break’. In a similar way, the statement ‘I tend to drink too much coffee’ is ambiguous: It might describe a passive state, identified in terms of frequency or hypothetical prediction of a certain result, or it could denote an occurrent activity. I might have a tendency to drink too much coffee without, now, tending to do it. It can also be true that I, now, *tend* to drink too much coffee although I have no tendency to do it, nor am I actually now drinking too much coffee. So there is an ambiguity in ‘tend to’ which the dispositional account of tendencies overlooks.

> What a modern philosopher would call a ‘dispositional’ account of tendencies blurs this distinction between occurrence and state by failing to allow for the possibility that ‘tending to do something’ can equally well denote either an occurrence or a state. (Champlin 1991 p.127)

The consequences of failing to distinguish these meanings of ‘tend to’ can be substantial. Consider, e.g., the importance of how to understand the term in J.S. Mill’s famous proposal that ‘actions are right in proportion as they tend to promote happiness, wrong as they tend to promote the reverse of happiness’. (See T.S. Champlin & A.D.M Walker 1974). Does he mean, like rule-utilitarians, that actions are right if they belong to a type of acting, which in “normal” circumstances promotes happiness – permitting that individual actions are right in spite of their failure to promote happiness? (“Normal” can be taken statistically or as an abbreviation for a closed set of circumstances, which are specifiable, at least in principle. I.e. the rule utilitarian could make use of either the frequency or the disposition account

of tendencies.) Or is he claiming that the rightness of an action is proportionate to the degree of happiness-promotion it tends to yield in every individual case? The latter interpretation presupposes that it really is possible to give a non-dispositional characterisation of 'tendency'.

The tendency-concept I am looking for is teleological. That intuition may require some defence. T.S. Champlin argues, e.g., that the impression that tendencies are teleological notions depends on confusing 'a tendency towards or away from something' with 'a tendency to do or to be something'. Only the first, logically distinct, sense of 'tendency' involves goal-directedness, he claims.

> The assimilation of tendencies to do something to tendencies towards a goal is what lies behind the incorrect characterization of all tendencies as teleological. (1991 p.131)

According to Champlin, there is e.g. no teleology in a bored pupil's tendency to fidget. The distinction gets lost, he thinks, because many examples of tendencies to do things concern actions which, unlike fidgeting, are in themselves goal-directed. A football-player, who tends to score, performs an action directed towards a goal – but her *tendency* is not what is goal-directed.

Champlin's initial assumption about how philosophers traditionally have viewed the role of teleology is ambiguous: "Tendencies have been said to be teleological notions" (p.122). Does he want to question the conceptual assumption that 'tendency' is a teleological notion? No, as far as I can see, Champlin's arguments are concerned with a less conceptual question: "How much teleology is there in *tendencies*?" (p.122, my. emph.) His arguments about fidgeting and scoring are construed to show that tendencies (save tendencies towards or away from things), literally speaking, are not goal-directed. I am prepared to admit that the attribution of such goal-directedness would force us to think of tendencies as occult directed "powers" or "agents". (Champlin claims that failure to distinguish tendencies to do something from tendencies towards a goal explains why philosophers like Mill and Descartes have tended to treat tendencies as directed forces.)

But that admission does not force me to abandon the view that there is a *concept* 'tendency' which is teleological. Statements about tendencies implicitly indicate a foreseen possible result. The result is implicitly understood as an unmodified claim, which the tendency-statement is supposed to weaken. If 'this pupil tends to fidget' is to be understood frequentially, the implicit unmodified result of the tendency is 'this pupil invariably fidgets', if applied to a single occasion, the unmodified claim is 'this pupil fidgets'. The most natural interpretation of this statement is frequential, and in that reading, it seems farfetched to regard the unmodified generalisation to be a goal in any sense. However, there is a difference in kind between statistical tendencies, and tendencies attributed to a single case.

One difference is that there is a certain ambiguity in the non-frequential "tendency", which is excluded in the frequential. 'This football-player tends to score', taken in the frequential sense, implies that 'this football-player invariably scores' is false. However, when 'this football-player tends to score' is applied to an individual case, two interpretations become possible – one excluding success, and the other admitting it. Even the interpretation admitting success is though, "patchy" in the sense that it indicates a *possible* shortfall. When I report from the football field that 'no 10 tends to score a goal', this does not necessarily exclude the possibility that she actually scores, although it necessarily includes the possibility that she might fail. The latter "generous" sense of 'tendency' is presupposed in my suggested definitions of 'want' and 'intention'. Otherwise I would rule out successfully fulfilled intentions by definition.

I do not find it farfetched to think of the unmodified result, presupposed by the modifying tendency statement, as the *goal* of the tendency, when we talk about single case tendencies. That notion of a tendency is bound up with a notion of a triggered causal process, which would produce a certain result, were not the possibility of shortfalls apparent. Why not call that indicated result a goal? The idea is not parasitic upon the teleology of the football-player's action. The (single case non-statistical) tendency of the ice to break under the foolhardy skater is also directed towards a goal in this profane sense. When we attribute that tendency to the ice, we have the result of the completed causal chain in mind.

In other words, talk about tendencies as striving towards goals should be interpreted more generously than in terms of magic forces. The terminology need not express the confusion Champlin attributes to Mill and Descartes, but merely the conceptual claim that 'tendency' implies an assumed result (which might not be obtained). The "goal-directedness" of a tendency is then nothing mysterious. The presupposed goal in the tendency-statement is simply the result the speaker assumes would obtain, given that the causal chain could go on uninterrupted.

This assumption about direction of tendencies should rather be compared to the innocent use of teleological explanations in biology. The expected effects for a plant getting sunlight and water enough to survive and multiply, can e.g. be presented as an explanation of the specimen's form of growth. As Ernest Nagel forcefully made clear in *The Structure of Science*, such explanations are quite legitimate, given that they are viewed as shorthand versions of complicated causal explanations. (1961, pp.401-428)

All uses of 'tendency' indicate that the supposed result need not be fulfilled. In Champlin's words: All "talk of tending to do or to be something is 'gappy' or 'patchy' in the sense that it indicates that somewhere there is a shortfall". (1991 p.127) The frequential notion is used to indicate shortfalls in

frequency, the dispositional allows for insufficiency of conditions. But could there be a meaningful non-occult and non-dispositional account of 'tendency', which nevertheless can make sense of both the teleological and the "gap-permissive" functions of the term?

Consider a person who tries his grandmother's home made medicine, and then surprised reports that 'this stuff tends to cure me', It indicates that the speaker has a foreseen result of a certain process in mind, and that he can imagine a shortfall in the scope of that process. His statement is not frequential, since it refers to what happens on a single occasion. Nor is it conditional – he does not, and would definitely not be prepared to, claim that the medicine has properties such that 'under such and such circumstances, it *would* cure me'. What the applicability of 'tendency' in this non-dispositional sense seems to require is merely that one refers to a process which can be predicted as leading to a certain result – and that the process towards this result can be thought of as being fulfilled to a certain *degree*. If the possible answer to a single question is either 'yes' or 'no', we cannot say about a specific answer, that it tends to be correct. We might, however, say about a student that she *tended* to answer the question correctly if she produced a chain of arguments heading towards the correct solution. Such examples show that the goal-directed and "patchy" elements in 'tend' can be done justice, even when the term is used in a non-frequential and non-dispositional sense. This use of the term presupposes only that it is applied to a process, which can be thought of as producing a specific result, and that the process (but not necessarily the result) can be imagined as *partly* fulfilled.

Consider, again, the act-utilitarian interpretation of Mill's statement 'actions are right in proportion as they tend to promote happiness'. 'This action tends to promote happiness' contains the information that the speaker has a certain thinkable result of the action in mind – happiness-promotion – and that there might be a shortfall in the scope of the process imagined to produce that result. The rightness of the action would in this interpretation vary with two variables: the amount of happiness the complete process is thought of as producing, as well as the degree of completion of the process.

The requirements for application of this non-dispositional notion of a 'tendency' are, as far as I can see, fulfilled in uses like *'tend to cause'*. I will therefore insist on claiming that 'tend', as it figures in the BD models concept of 'desire', can be thought of as non-dispositional. The teleological element in this concept does not, as should be clear by now, force us to think of the unfulfilled process referred to as somehow predestined towards the imagined goal – i.e. as guided by a directed power. Nevertheless it is, I believe, possible to interpret 'x tends to cause p' realistically: As referring to an ongoing causal process, the (unestablished) completion of which will result in p.

Summary of 2.2: The dispositional concept of 'desire' inherent in the BD model requires a distinct account of 'tendencies'. The notion required should be *non-frequential* in order to makes sense of the idea of single case tendencies. It should also be *teleological* in the sense that it explains why we think of tendencies as heading towards some pre-imagined goal. Furthermore, it must allow tendencies to stop short of that forecasted end. When we talk about causal tendencies, we employ a concept of that kind. We refer to an ongoing process that can be predicted, from our point of view, to produce a certain result. The process towards that result can be fulfilled to a certain degree. There is nothing magical about the teleological element in 'tendency' regarding this view, nor does that goal-directed element exploit the distinct intentional notion of being directed towards a goal, as some have objected to the teleological picture of tendencies.

---

[1] Mellor's change of view about the role of dispositions is indicated by their complete absence in his analysis of causal explanation - in terms of universals and laws as "nomic facta" – in *The Facts of Causation* 1995. In personal conversation at the ECAP-conference in Leeds, September 1996, he confirmed that he no longer thinks of dispositions as explaining their manifestations.

[2] The case with three dispositions having one and the same base would not have to be a problem to the supervenience theory (proposed by Prior, Pargetter and Jackson) either. That theory does not have to exclude the possibility that more than one dispositional property supervenes upon a certain micro-structural base. Like in my eliminative view, it would have to say, of course, that there is one common factor, which explains displays of different dispositions like electric and thermal conductivity as well as opacity. But that appears to be quite in line with what Menzies and Crane took the initial observation to establish.

3 Content of Desire

We talk about desires in terms indicating that they refer to things outside themselves. Desires can be *satisfied* or *realised*, and we sometimes make them *come true*. They are regarded as satisfied "if and only if the propositional complement of 'desire' is true" (Stampe 1994 p. 246). We speak of them as if they were inner representations of the various states of affairs they will contribute to under the right kind of conditions. Different desires tend, causally, to result in different states of affairs, but they are related to these states of affairs not only causally, but also referentially.

A desire that p might never manifest itself by giving rise to any tendency towards realising p. Such a desire exists as an inner state warranting conditional predictions about the agent. But in which sense can its abstract unrealised *content* "be there" even in the absence of external displays?

Furthermore, content appears to be of causal relevance. The fact that one desire is about p and another is about q is exactly what is supposed to explain the differing causal powers of these states. A representation with a certain content may be a cause; Printed tokens in front of you, representing an English sentence, are among the causal conditions for your present perception. Representations can be causes, but it seems a mystery how semantic abstract properties, like the proposition they may represent, could make a causal difference. E J Bond argues, e.g., that beliefs "might be regarded as causes, but one cannot so construe belief *contents"*(1983 p.23).

So, are these assumptions about content compatible with the BD model's non-phenomenal and dispositional concept of desire?

3.1 From Object to Content

David Hume firmly declares passions (inner states performing the action-guiding role of BD model desires — on the most widespread interpretation of Hume) to "have no more a reference to any other object, than when I am thirsty, or sick, or more than five foot high" (1739 2:3:3). And Stuart Hampshire claims that wanting "unlike, for example, regretting - is not an essentially thought-dependent, and therefore an essentially human, concept" (1975 p.37). His examples of such thought-independent desires are lust, hunger, thirst and other urges intimately connected with bodily needs and sensations. "When I am starving, my desire to eat does not depend for its existence on any particular conception I have of this activity: it depends solely on the stomach" (ibid. p.47).

Both declarations are ambiguous. They could be saying that desires represent nothing *as being the case*. Such a denial of the assertive function of passions would suffice for Hume's argument about passion's immunity to reason's allegations about truth. It is also quite in line with the BD model to deny desires

that role. However, the most natural reading, in both cases, is that these philosophers refuse desires to be *about* anything, i.e. they object the view that desires have content. But even if Hume and Hampshire both really deny that desires have content, they cannot mean that desires have no intentional character. Any account of desires must admit that desires are directed, i.e. that they are *for* something.

Hampshire's and Hume's positions stress that desires, essentially, guide and produce actions. (Arguments for assigning this view to Hume are presented in section 6). To attribute a desire to A is to view her as being in a state which will trigger causal tendencies under specifiable conditions. Tendencies, I have argued, are teleological in a down-to-earth sense. When we talk about the tendency of an object in a single case, we have a certain result in mind, towards which the object is heading. More properly speaking, the object itself need not be moving towards the imagined goal, but there must be a sequence of causally linked events, for which we regard the object's constitution as an important causal condition. I should emphasise that the process is on its way to a pre-envisaged endpoint only as seen from a spectator's perspective. No representations of the goal have to be attributed to the object or to any other essential element of the process, for it to be teleological in this sense. As I made clear in the former section, their goal-directed appearance is therefore not a confused generalisation from those tendencies which are behavioural, and where it is already presupposed that the object itself has a goal in mind.

Since desires by definition are tied to causal tendencies, there is one sense, then, in which they, at least from the third person perspective, have objects. The agent who desires is not excluded from viewing himself like that. He might diagnose himself by forming hypotheses about his own causal constitution, and make conditional predictions about himself, although, as I have argued before, he is not better off than people who know him well when it comes to the reliability of such assumptions (on the contrary).

Do we need to assume desires to have objects in any other sense than this? The object of the desire is the result it tends to produce, when triggered. There is in that case no need for assuming a representation of that result anywhere within the agent.

As Tim Crane points out, the direction of the desire can not be a relation between the internal state and its actual result, since that would exclude desires for things that do not or can not happen (1996 p.211). The tendency view under consideration would handle that requirement. Since tendencies are "gappy" — they admit the possibility of stopping short of the imagined result — but nevertheless are teleological from the spectator's perspective, an account identifying a desire's object with the result it tends to cause will allow effective (i.e. triggered) desires to fail *to reach* the realisation of their objects.

There are several imaginable kinds of failure of a desire to realise whatever it is directed towards. Some desires never result in any tendencies towards their object at all, since the conditions for triggering tendencies are never present. More dominant desires may always prevail, or the agent may simply be certain

that the object of desire is far beyond reach. An advantage of the dispositional account is that it allows us to attribute desires in the absence of their manifestations, just like we ordinarily do. However, even in these cases, we could from a third person perspective say that the disposition is directed towards an object. That might simply be a conditional prediction about a causal tendency towards an imagined result.

Failure resulting from misrepresentation appears to be more difficult to harbour within this simple causal account of desire's object. Suppose Fido desires a dog biscuit in your hand, and jumps in order to get it. When he catches it and begins to swallow, it turns out to be a pill he would not eat unless you deceived him. Although his desire causes a completed causal tendency to get the pill, it seems fair to say that what he wants is a dog biscuit, and that he does not desire the pill in any sense. The object of his desire is not what the tendency is directed towards. We infer Fido's desire for a biscuit from a tendency to get what Fido *believes* to be a biscuit. But how can the causal base of a disposition to get a biscuit have the property of being causally sensitive to the (mistaken) belief that something is a biscuit? That seems to presuppose representational qualities in the desire.

Let me make use of a distinctly different approach in order to expose more clearly the view that intentional actions must be caused by states with representational content. Frederick Stoutland presents a completely non-representational model of the content of reasons for action. Roughly, his analysis says that R is a reason for x, if x is a response to R (under the proper description). When A does x because of R, R *might* be a propositional attitude, but also an event, an external object or whatever. When the stop sign makes me stop, my reason is the stop sign, *not* a representation of that sign. When the biscuit in your hand makes Fido jump, his reason for jumping is that there is a biscuit in your hand. It is simply a prejudice, according to Stoutland, to believe that only representational states could function as reasons. Reasons in Stoutland's sense are nevertheless agent-relative, since action-responses are always triggered due to background conditions, among which the agent's beliefs and commitments may be an important part. When someone acts on a false belief, like Fido, his belief will *become* a factor of explanatory value, but otherwise there are no reasons to cite his intentions, goals etc. in reason-explanations of what he is doing. (Stoutland 1998 p.61)

(The present issue is whether reasons for action have to contain inner representations of propositions. Therefore I will disregard, for the moment, Stoutland's overall ambition to argue, in the tradition of von Wright and many others, that action-explanations are not causal at all. In his view, intentional actions are not effects of their reasons, i.e. the external things they directly respond to.)

Stoutland's approach provides an analysis of a reason as whatever an act is a response to. This means that it goes firmly against the idea that content must be internal to the agent. Content need not be represented in any state of the agent, but Stoutland does not exclude the possibility of propositional attitudes

functioning as reasons. When it comes to Fido's act, a propositional attitude is what his act responds to. The act is to be explained by Fido's false belief.

Most objectionable in Stoutland's analysis is that it seems quite ad hoc to assume that someone's "belief becomes an explanatory factor /…/ just in case she was mistaken about the situation originally appealed to as justification." (1998 p.61) If we have to invoke propositional attitudes to explain misdirected attempts, how can we disregard their role in causing successful attempts? We can easily imagine two cases where every condition generating the action, including the agent's inner states, are similar up to the point where the expected result is about to be realised. Information about false beliefs may often be less evident than information about true beliefs, and the explanatory value of the information may be greater in that sense. But if the causal stories are similar, except in that Fido gets a biscuit in the first case and a pill in the second, Fido's propositional attitudes must be just as responsible for producing his behaviour in the first case.

Furthermore, it seems difficult to draw Stoutland's line between background conditions and explanatory factors in a non-arbitrary way. The question of whether we want to view e.g. someone's norms as explanatorily valuable, or as belonging to the causal background, would not depend on their actual role in producing behaviour, but on how extraordinary they are, given our beliefs and presumptions. That is another argument for assuming that if propositional attitudes are necessary to explain the unsuccessful cases, then they should be invoked in the successful ones. In other words, I regard Stoutland's attempt to eliminate representational content from reason-explanations as inconclusive.

The difference between Fido's desire to eat a biscuit and his desire not to eat a pill can be characterised in terms of the various kinds of states of affairs the two desires would tend to cause under different conditions. In order to state that difference, we must refer to Fido's possible beliefs among the imagined triggering conditions. But to admit that much is to let desires have propositional content — in a functionalist sense. Belief-sensitive action-guiding inner states have a content purely determined by, or constituted by, their role in relation to specific beliefs and acts. "Desires have determinate content because of their dual connection with belief and action." (Stalnaker 1984 p.19) When combined with the right beliefs, a desire that p tends to bring about the state of affairs it is a desire for; 'it is the case that p'. When triggering beliefs are faulty, desires might tend to realise a state of affairs, which the agent mistakenly believes to be the one he desires. This description implies that desires have representational qualities. Desires respond to correct as well as incorrect representations of the world, and their satisfaction consists in realising a state of affairs as represented in the desire.

Dispositional differences reflect real internal differences in the object's constitution, in accordance with what was established in the former section. Therefore, on this functional definition of propositional content, it makes sense to speak of the objects of desire as being represented *within* the system. An

occurrent but unmanifested desire has a content which can be picked out via its functional role, in terms of different conditional predictions, but which also reflects an underlying causal base of the manifestations used to identify it. This means that, although the content of a desire is identical with a certain causal role, that role exists in virtue of a content-specific existent and causally efficient internal state of the agent. "Content-specific" does not imply that each type of content has its type of internal state. The relation between functional role and internal base is that of type to token.

This means that content is internal to the agent in a sense to be clearly separated from the radically internalistic Cartesian view that "one's mind (with its contents) neither involves nor depends upon anything external" (Sosa 1995, p.309). The idea defended is that internal *and* external features determine content of desire. Which desires a person has will depend upon his behaviour under different imaginable circumstances. The conditionals required to characterise his desires will necessarily tell us things about how he would respond to different aspects of his environment.

The approach is nevertheless internalistic in a weaker sense. Facts about the inner character of the agent are implicitly supposed to warrant the truth of the conditionals implicit in desire-assignments. When you point out that A desires p, you want the listener to grasp that there are noteworthy facts about the agent herself (rather than about her former, present or expected circumstances), which are especially relevant to the truth of the conditional forecasts about her behaviour. As Ernest Sosa suggests in an article on the possession of concepts,

> the distinction between internalism and externalism is not just a matter of whether the property or the family of properties in question is reducible to conditionals about how *x* would behave in certain conditions. This is not enough, for the truth or falsity of such conditionals might itself be relative to a presupposed environment — in such a way that the truth of such a conditional is determined not just by the intrinsic character of *x* but also by the character of the external environment. (1995 p.323)

In other words, although the conditional characterisation of a behavioural disposition must refer to something external, it can in itself be more or less internalistic. That will vary with our assumptions about what the truth of these conditionals requires. When we assign desires to someone in accordance with the BD model, we abstract from the actual external circumstances in which the agent finds himself. Desires are defined in terms of inputs and outputs, and their content is determined by the agent's predicted responses vis-à-vis changes in the environment. Nevertheless, our assertions about these dispositions are "depending for their truth on no external grounds concerning *x*'s environment, past, present, or future, at the moment when *x* has such a disposition." (ibid. p.323)

This functional characterisation of desire's content says that the content of a certain desire belongs to its dispositional features. The content is not a disposition on its own, but an inseparable part of each desire's specific dispositional

character. If we spell out all the conditional predictions, which make up a desire for p, we would not only say something about what it is for it to be a desire. We would also say all that needs to be said about what it is for a desire to be about p.

We might think of a certain desire with specific content in terms of a certain pattern of behaviour in relation to beliefs, i.e. as the very role or functional state relating specific possible inputs and outputs to each other. Furthermore, we could want to stress that there is an inner (perhaps neural or mental) state playing that role, i.e. that internal state warranting the predictions which *define* the role. We might use the language of role functionalism ("functional state identity theory" defended by Block 1990 and Crane 1998) or of realiser functionalism ("functional specification theory" defended e.g. by Armstrong — see Bransen & Cuypers1998 p.11). Both ways of speaking are useful options inherent in dispositional talk.

But if my eliminative suggestion in the former section is plausible, none of these identifications reflect the full and true story. The distinction between desires as roles and desires as realisers of roles rests upon a view of "roles" as distinct from the intrinsic properties of the object. An object's role is in this sense like a hypostatised set of conditional predictions about it. (In this respect the object's "role" differs from e.g. an actor's "role", which is a task actively assigned to him, in order to make him play a pre-established part.) It is misleading to identify desire + content with a certain causal role in that sense. That would be to forget the internal causal warrant of the predictions hypostatised. On the other hand, to assign specific content to a certain inner state is not *only* to attribute to it certain intrinsic qualities, but also to add some conditional predictions about the object in possession of the state.

This analysis differs from the idea that there are two incompatible perspectives on agency, which is defended by Jennifer Hornsby (1995) and by Thomas Nagel in *The View from Nowhere* (1986) among others; the impersonal and the personal. Hornsby argues not only that these points of view are mutually exclusive, but that actions are "absent simply, from the impersonal point of view." (p. 185) "Those who give and seek 'action explanations' do not regard the matter impersonally or externally, any more than the agent herself does when she deliberates about what to do." (p.180).

The account suggested here proposes an impersonal or "non-individualistic" (Jackson and Pettit 1995) viewpoint in declaring that the content of important action-guiding entities like desires can and must be picked out via references to things external to the agent. We must explain acts in terms of things just as accessible to other people as to the agent herself. But that does not contradict the idea that explanations of actions must be internalistic in the sense explained. These explanations are attempts to fix the individual agent's inner qualities, as independently of his actual external circumstances as our imaginative capacity permits us to consider.

Nothing in this picture indicates that there is something wrong with thinking that empathy, 'taking the other's point of view' or 'putting yourself in the

other's shoes' are indispensable methods of understanding why people do what they do. Behavioural patterns are complex, aims and goals changeable and dispositions can counteract each other. To pick out the right points of intersection between possible beliefs and actions, we would have to imagine the agent in a great variety of circumstances, where subtle nuances and quickly passing aspects of the situation may make all the difference. Language is blunt when it comes to catching the motivationally relevant richness and detail of real life choice situations. To imagine the facts from the other's perspective is probably a more efficient way of forming an opinion about what the agent would respond to, than to attempt describing them and their relations to action-alternatives. Furthermore, one should not underestimate the common ground for decisions. Many goals are probably common to all humans and an even greater number are common to those we share culture and many experiences with. 'What would I do if I were in his place?' might therefore be a proper starting point for attempts at conditional predictions about someone else.

It might therefore still be proper to say that action explanations in a sense are "for those who share with that person a point of view on the world" (Hornsby 1995 p.180). The extent of that admission should not be exaggerated, however. Its point is merely that viewing myself in your shoes could be a useful methodological device, if I should make an attempt at understanding why you do what you do. Unlike Hornsby's and Nagel's analyses, the view defended here assumes no insoluble Kantian antinomy in people's ways of picturing or describing their actions.

There may be *some* facts in principle accessible only to you, like facts about your conscious states or about which motivational states you are aware *of*. But the driving forces that produce your behaviour are not necessarily phenomenally present, nor are they necessarily present in your active deliberation in some other sense. If you base your own explanations on introspection, chances are great that phenomenally salient features of your present state of mind will prevent you from making an unbiased estimation of the forces motivating you. That is one of the things which might explain why people close to us often detect our motivators and predict our behaviour more successfully than we are able to do from the agent's perspective.

Summary of 3.1: Two non-representational suggestions about the objects of desire were examined and rejected. A desire's object is not identical with whatever the triggered act would tend to cause, nor with the external things the act responds to. These views have difficulties in characterising motivational failure, due to misrepresentation (in a manner that is not ad hoc). Furthermore, they fail to catch our ordinary way of talking about the content of desires in a semantic sense — desires "come true", are "realised" etc.

The dispositional notion of desires carries with it a functionalist notion of desire's content. Such a characterisation assumes that the content of a specific desire can be framed in terms of publicly available facts, like possible inputs and outputs and their relation to beliefs. An assignment of a desire is nevertheless

internalistic since it claims to expose those conditionals which are true in virtue of facts about the agent's inner states.

3.2 The Causal Relevance of Content

A functionalist characterisation of content saves the BD model from begging the question against physicalism, and it explains how the BD model's desires can be essentially non-phenomenal. Inner states with representational qualities could, for all we know, be neurological imprints, "sentences" in the brain (Fodor 1981 esp. ch.7), affecting the readiness of certain neuron-sets to fire (to recycle, again, Brandt's guess about how desires are physically stored). In principle, this sort of functionalism about content of desire begs no question against Cartesian interactionism either. But apart from conventional objections to such dualism, it would require an elaborate and psychologically dubious hypothesis about sub- or unconscious mental states to maintain the BD model's non-phenomenal conception of desires within a view of desires as essentially *mental.*

Although it is clear that an inner state with representational qualities, which warrants certain predictions, could be a cause (just like marks of ink with such qualities can be causes), it remains to be explained how *that which is represented* could make a causal difference. Stoutland's intuition that reasons are whatever we respond to appears, in this case, plausible to some extent. When you regard a certain reason as rationalising your ϕ-ing, you explain your ϕ-ing by referring to different things — objects, situations, and so on — relevant to that kind of behaviour. From such information about external facts, we come to note the state of affairs you aimed at. You *express* relevant beliefs and desires in your reason-explanation, but you *cite* the things they refer to as being your reasons. It is not necessary to add that you have certain beliefs and desires *about* these things, since that is trivial. When a desire is misdirected or a belief is false it would be less trivial to point out that your reasons were believed or desired by you. (That is also in line with Stoutland's intuition that beliefs and desires *become* explanatorily important when they are mistaken.) Like E.J. Bond, Stoutland concludes that if 'reasons' refers to the propositional *contents* of the alleged causes, then reasons cannot be causes (Bond 1983 p.23, Stoutland 1998 p.44).

One way of attempting to bypass their worry might be to admit that the traditional characterisation of BD model explanation as 'reasons as causes' is somewhat elliptic, and that a more adequate label would be: 'causal explanation in terms of beliefs and desires containing the agent's reasons'. It is compatible with admitting that reasons are causally inert abstract entities "to claim that *the thinking of certain thoughts* the contents of which are reasons could be causes."(Persson 1992 p.111)

That is correct, but within the dispositional view under examination here, this move would not adequately meet the challenge. Beliefs and desires are supposed to make a causal difference *in virtue of* their content. It is no causal coincidence that beliefs and desires about p cause ϕ-ing, (rather than γ-ing etc.). They do so because p figure in their content. To use Dretske's analogy, unlike the shattering effect on glass of "a soprano's upper-register supplications", the effect of beliefs or desires vary with their content (1988 p.79). The BD model's dual claim about causal explanation and justification hangs on this.

What remains to be explained, then, is how abstract entities like contents can be causally relevant. Through what mechanisms could the truth of 'this desire is about p' become a causally relevant fact about the desire? Note the gap between *relevance* and *efficacy*. "Trying to exhibit the causal efficacy of meaning itself would be like trying to exhibit the causal efficacy of mankind, justice, or triangularity" (Dretske 1988 p.80)

Crane expresses strong doubts about the distinction between *something* being causally efficacious, and its being causally relevant. He does not believe that "there is an adequately worked-out notion of causal relevance which is distinct from the ideas of causation or nomic sufficiency, and which is not simply based on striking but ultimately unexplained metaphors". (1998 p.203) Apparently, he suspects that causal relevance, properly explicated, will collapse into causal efficacy. I find that suspicion understandable, given that we talk about causal relevance as a relation between those kinds of entities which *also* might stand in relations of causal efficacy to each other. We would perhaps choose to say that a certain historical event was causally relevant to another (rather than that it caused the other), if we wanted to stress that the contributing effects of the first on the second were complex, indirect, or very small. 'Causally relevant' might just be a wide term covering various efficacious causal connections. I am inclined to agree that it would be difficult to make clear how, e.g., an *event* could be of causal relevance to another event without assuming that it has effects at all.

Armstrong and Crane both assume, on different grounds, that dispositions cause their own displays. Suppose they were right in that the manifestation of a disposition is an effect of that very disposition. Since desires are dispositional states, and the content of desires is essential to which manifestations they might have, the question of whether content is efficacious or merely relevant would in that case, on account of what was said above, seem to be more of a terminological quarrel than a real issue. Crane would then have a case against those who identify content functionally or dispositionally but still stress that the possibility we need to explain "is not of meaning itself being a cause, but of a *thing's having meaning* being a cause" (Dretske 1988 p.80). My point is that the distinction between the possible relevance and the possible efficacy of content (within this functionalist framework) will be threatened only if dispositions, roles and their manifestations are presumed, from the start, to be the kind of entities between which there can be causally efficacious relations.

The eliminative view of dispositions I suggested before implies that disposition-statements, taken literally, always both assign inner states to the object, and express conditional predictions about it. But there is no irreducible dispositional property. Nevertheless, as I have *already* indicated in the former section, I am prepared to agree with Crane that it makes sense to think of one disposition as a causal condition of another disposition, as well as a causal condition of a certain event. Dispositions may be causally efficient in that sense. The fact that an object A will ϕ under circumstances C, might explain other conditional predictions, e.g., why A will γ under D, or why B will ϕ under C etc. The the book page in front of you will reflect some waves of light and absorb others when illuminated. Therefore it will look black and white to you if your eyes work properly. Because of that, you will be able to read what I have written as long as the surface is illuminated, and so on.

In "The Efficacy of Content" Crane never explicitly distinguishes the question of whether dispositions can be causes, from the question of whether dispositions cause their own displays. It is of course sufficient to show that dispositions cause their manifestations, as Crane attempts to do, in order to prove that they can be causes. However, it is not necessary, since we could argue that dispositions, as well as roles and other facts about what can be conditionally predicted, can be causally efficacious without allowing dispositions to cause their own displays.

Crane's own form of functionalism is a role functionalism, where the disposition (in this case the belief or the desire) is a functional state, "a relatively abstract state of being such as to have certain causes and effects" (1998 p.207). Dispositions are real but identical with roles. As I mentioned before, the alternative is a realiser functionalism, like Armstrong's, according to which dispositions (like beliefs or desires) are internal states performing roles described in terms of causes and effects.

It is tempting but misleading, here, to think of an object's role as if it was a pre-set task the object is set upon fulfilling, a job rather than a position in which the object is situated. People may for different reasons *take on* roles, much like actors. A prudent and dutiful person who is appointed port master will soon get into the character and actively perform the role of a port master. Note that the subject-role of a port master is distinct, also, from her appointment as a port master. She could be appointed port master and still not choose to perform the things characteristic of the port master's role, and vice versa.

The fact that her role is that of a port master may be a causal condition for many *things*; e.g. that local fishermen pay her some respect. However, the role is not a causal condition for the various acts *constituting* her performing the role of a port master to a lesser or greater extent — her decision to get *into* that role is the condition for those acts. In turn, her port master appointment is what makes her take on that role. Similarly, when an actor is cast in a certain role, his role is clearly a causal condition for some of his actions. But the actor's role is a task assigned to him, which he is ordered to fulfil. It consists in his being paid to play

or perform a part at least partly as determined in advance by the script, the director's intentions etc. Such appointments may have direct causal effects. But the actor's role is then not simply a hypostatised set of conditional predictions about how the actor probably will react under different conditions, but a device in the form of a fictional character, actively used to produce a certain behaviour. In other words, that kind of subject-role performance is not what I am talking about when I say that content is determined by causal role.

My view is that dispositions and roles are causally relevant for their displays, and that they may be causally efficacious for other things. "Brittleness" is an elliptic label for conditional predictions and hints about their causal warrant. Brittleness is nevertheless causally relevant to its displays, since 'the ice is brittle' informs us that there are causally efficacious properties of the ice, on account of which we may expect it to break under specifiable triggering conditions.

Crane's strategy is to show "that we can explain the efficacy of content without having to distinguish between causal efficacy and causal relevance, if we adopt the conception of dispositions". The conception in question is a non-reductive analysis of dispositions as role states, essentially identified in terms of certain outcomes, being causally efficacious precisely because they are directed towards these outcomes.

Tactics could be reversed. One might start with Dretske's intuitively plausible assumption that we take it for granted that "*the fact that something has meaning* [is] a causally relevant fact about the thing" while we are inclined to think of attempts to "exhibit the causal efficacy of meaning itself" as doomed. (1988 p.80). Another attractive standpoint concerning belief's and desire's contents within the BD model framework is that content can be characterised as a part of the dispositional character of these states, i.e. as determined by a certain causal role. Such a functional characterisation of content comes in the bargain, if we accept a dispositional analysis of desire. But if a certain hypothesis about dispositions forces us to abandon either the functional view of content, or the distinction between causal relevance and causal efficacy of content, the wisest thing might be to let go of the disposition-theory in question.

Summary of 3.2: A desire's content is causally relevant in virtue of its being an element in the desire's dispositionality. We expose the content of a desire by describing various causal relations between triggering conditions and behavioural displays. However, content is not a separable and causally efficacious role *state*, it is a functional feature of the agent, which is warranted by facts about the agent's internal character.

Functionalism about intentional states and representations is a well-established philosophical tradition in which philosophers like Armstrong have produced a variety of suggestions about how to reconcile the causal and the justificatory allegations typical of the BD model. My primary ambition here is to give a hint about the type of account of content that would do justice to the BD model's

notion of desires. Even if my specific suggestion about how to understand desire and content in dispositional terms should appear to be incorrect, I am convinced that the key to such a reconciliation is to be found within a dispositional or functional approach.

3.3 Non-linguistic Content

We normally assume that ‘desire’ “is not restricted in its application to language-users, who can think of the object of their desire as being of a certain kind” (Hampshire 1975 p.37).

First, it must be noted that when we apply the term “desire” to some non-human animals, we sometimes use the belief-desire terminology metaphorically. Or even mistakenly: in cases where the apparently intentional behaviour of the animal can be exposed as automatic. Some people would e.g. like to say that the bird who feigns a broken wing in order to distract a predator *wants* to get the predator to believe that she is easy prey, and thereby draw its attention from her nest. The automatic character of such behaviour may be revealed by its rigidity — in situations where the feigning-response is triggered or continued although its function obviously is pointless, for instance. Or, to take an example where the absence of intention is more evidently revealed. If you move a branch of seaweed from the water's edge on the Danish shore of the Sound further up on land, the insects in it *(Gammarus Locusta)* will jump eastwards *in order* (or so the spectator might be tempted to think) to get back to the water. However, if you bring the branch of seaweed from the west shore fresh by boat to the Swedish shore of the Sound, the west-shore animals will still jump eastwards when disturbed (hence further away from the sea), even if the branch lies close to the water.

The BD model does not identify desires with patterns of behaviour. Therefore it will always leave room for scepticism about when it is proper to regard an organism as an intentional system on account of its behaviour. Observations of rigidity and adaptiveness will, however, often give us good reasons to suppose that an organism has beliefs and desires – or that it does not. In this case, it seems reasonable to suppose that the insects are revealed as having no desire to get to the water. Their behaviour is a triggered reflex, rewarded by natural selection in these populations' frequent change of generations. (The point of the experiment may, I imagine, be to illustrate that they have evolved sensitivity to a certain point of compass or direction of light.)

On the other hand, though much animal behaviour reasonably can be shown to be non-intentional, whether *all* non-human animal action will fall under this description depends on the plausibility of Hampshire's assumption that conceptions must be linguistical. To begin with, I do not think that the latter view receives much support from common sense observations. Why could not, for

instance, a new-born infant have a conception of e.g. stomach-pain, and, accordingly, a desire for that kind of pain to terminate? When our dog, after having tried two kinds of food, consistently chooses one of them before the other, just on sight of the two cans, is it not reasonable to admit that she has a preference, based on a conception of the contents of the can?

The functional characterisation of content gives us no reason to suppose that only language-users have beliefs and desires with representational content. (see e.g. Dretske 1988 p.75-76) The adaptivity and complexity of an animal's behaviour are our main clues for finding this out. Linguistical behaviour is but one of many sources of understanding.

The interdependence between a) the idea that non-human animals are incapable of wanting, b) a propositional analysis of desire and c) the linguistical view of beliefs and conceptions is spelt out in an illuminating way by R.G. Frey, a well-known opponent of animal rights. Frey presents a reversed version of Hampshire's argument. His doubts about the case for animal welfare are partly based on the assumption that desires are dependent upon beliefs and that (following Quine in (1960)) beliefs must be analysed in terms of *sentences* rather than *propositions.* (Frey 1980 pp 53 ff) His conclusion is, then, that animals are incapable of having desires.

Frey supposes that desires are dependent upon having beliefs that some things are *true*. This is thought-dependence in a stronger sense than the BD model's (that desires have propositional content). For criticism of Frey on this point, see Egonsson (1990) p. 129-130. In order to arrive at his conclusion against animals' desires, Frey would not have to commit himself to the claim about capability to believe that things are true. The discriminating factor is language, not ability to assert. Unless the idea of non-verbal cognitive content can be made sense of, most non-human animals are excluded from desiring. That is in itself a strong intuitive argument against assuming that propositional content must be linguistical.

Summary of 3.3: It has been argued that it cannot be part of our concept of desires that they are "thought-dependent" or have propositional content, since we assign desires to creatures without language. It was noted, first, that in some cases we get evidence showing that it was a mistake to think that a certain animal behaviour reflected desires. Second, and more important, is that the functionalist approach shows how desires can have a propositional content in non-linguistical terms. The fact that few of us regard infants, birds and non-human mammals as incapable of desiring supports that characterisation of content.

4 Signs of Desire

4.1 Desire and Sensation

As Michael Smith points out in *The Moral Problem,* many opponents of the BD model criticise it because they assume a phenomenal notion of desire. Mark Platts thinks, e.g., that any one who claims that all actions stem from desire, either puts forward a vacuous and boring claim, or a theory which is bluntly false, phenomenologically speaking. The possibility of desires as distinct explanatory factors without necessary phenomenal qualities is not among his alternatives. (1979, p.256) Adherents of the BD model sometimes seem to take this for granted as well. While claiming that it is conceptually true that desires cause acts, they may at the same time suppose that when an agent desires p, it is also necessary that the thought of realising p "occurs to him, occupies his attention, fills his consciousness" (Goldman 1970, p.86). This dual claim appears to be inherent in David Hume's picture of passions as well, as we shall see.

The BD model would be implausible on such a conception. Just like Platt's notes, most of us would soon find the model falsified if it presupposed that all our actions spring from experienced feelings. Then, why is that view of desires still common? A simple explanation may be that the term "desire" is ambiguous. "Desire" is not only employed when we refer to the roots of intentional actions, i.e. to those intentional states making sense of what people do. "Desire" is also sometimes used as a near synonym to terms like "urge", "drive", "yearning", "instinct", "lust" and other expressions of a more emotional or sensual character. But the BD model involves no claims about the motivational role of bodily needs, instincts or emotions. Let me illustrate the fallacy of equating the two notions of desire in some greater detail.

"Wanting - unlike, for example, regretting - is not an essentially thought-dependent, and therefore an essentially human, concept /.../. Desire presupposes only the capacity to act and to feel" (Hampshire 1975 p.37). Some desires "may come into existence independently of any conception", Hampshire states. As examples he presents "sexual desire, lust in any of its common forms, the desire of the hungry man for food or the thirsty man for drink and some other desires that arise from bodily needs".

However, the ambiguity in "desire" easily infects discussions concerning desires related to bodily needs. The English word "desire" is sometimes employed to refer to bodily sensations or feelings of bodily origin. It may, for instance, be used as equivalent to "lust" (OED). The point is simple, but I believe that this ambiguity is an important source of misunderstanding. It is sometimes difficult to disregard completely the flavour of physical sensations we normally associate with the word "desire".

It is probably true that sensations and feelings of this kind may arise independently of any conception of the cause of the feeling, or the means to get rid of it. For instance, a socially deprived person, like Kaspar Hauser, may perhaps feel signals of his bodily needs without being able to direct these feelings towards any special state of affairs. Neither is it certain that he will form any tendency to act upon them, since there are no necessary relations between these sensations and being disposed to act in certain ways.

The point is made by Richard Swinburne:

There are, however, desires which are accompanied by sensations, e.g. hunger, the desire for food which is accompanied by 'pangs' of hunger, i.e. by sensations which, we believe, the satisfaction of the desire will remove and which we desire to remove. But the desire for food is not the same as and need not at all involve the occurrence of unpleasant sensations (1985 p.434).

When Swinburne elsewhere describes desire as a slightly rebellious phenomenon in need of suppression and control by the more reflective parts of the agent's mind, his view becomes more understandable if “desire” is taken in the primitive sense. “Desire, in that case, inclines a man to act contrary to his beliefs about worth, including his moral beliefs.” (1986, p.115) This way of thinking would seem reasonable if it just meant that, for moral and prudential reasons, we can not always put the satisfaction of desires based on our bodily needs first.

To consolidate his thesis that “there may in principle be motivation without motivating desires”, Thomas Nagel refers to a distinction between 'desires motivated by reason' and 'unmotivated desires'. A desire to put a dime in a soft-drink machine may e.g. be motivated by a desire to drink together with certain beliefs. It is a trivial truth, according to Nagel, that if “desire” is taken to include motivated as well as unmotivated desires, then “whatever may be the motivation for someone's intentional pursuit of a goal, it becomes in virtue of his pursuit ipso facto appropriate to ascribe to him a desire for that goal”. He denies that an unmotivated desire (like thirst in the soft-drink example) is always. present in the motivational process. Motivated desires are always present, Nagel admits, but they are not the independent initiating driving forces the BD model takes desires to be. According to the BD model, actions are either produced by instrumental desires, which in turn, at some stage, have been caused by unmotivated desires and means-ends reasoning. It also allows for actions to be caused by intrinsic desire and beliefs directly. So, unmotivated desires always figure in the explanatory background. Such desires are the necessary driving forces of any action.

Nagel puts forward hunger and thirst as typical examples of unmotivated desires. Moreover, hunger and thirst are identified with desires for food and drink (1970 e.g. p.29 & 32). The intuitive plausibility of Nagel's examples of cases where motivation is supposed to work without motivating desires is, I think, undermined by the distinction

between felt “desires” and desires in the BD model’s sense. Unmotivated desires should be distinguished from bodily sensations, instincts, needs or urges. “Hunger” may refer to pangs of hunger, or to the desire for food. In some cases a sensual state related to physical needs is a causal antecedent of our unmotivated desire, like when a dry throat gives rise to a desire for drink. But it is clear that ‘desires’ in that primitive sense are unnecessary in motivation. All of Nagel's crucial examples concern desires closely connected with bodily sensations. Intuitions in favour of his view may therefore be based on considerations about the independence of intentional action from bodily sensations, which is irrelevant to the BD model’s view of motivation.

Let me now return for a moment to Hampshire's initial argument against the thought-dependence of desires. The distinction between desires as dispositional states with representational content, and ‘desires’ as bodily sensations does not exhaust the possible uses of “desire”. Hunger, lust and many other urges appear to be non-representational. One explanation of this impression may be that we associate these ‘desires’ with bodily needs, or sensations associated with such needs, both of which are non-intentional (or at least non-representational) states. But assume that we do not speak of hunger and lust as the sensations produced by bodily needs. There is then still another, non-representational, sense in which these almost universal drives can be directed towards objects.

A common view of *sexual* desire seems to be especially difficult to fit into the two categories suggested so far. Could not Kaspar Hauser sexually desire something or someone (the first human being he meets, for instance) without having any idea about what it is that he desires? Might he not just *want* her, though he does not have any idea as to what kind of state of affairs that could fulfil his desire? In that case, there would be a sense of “sexual desire” referring to an internal state with a direction, a desire *for* something, but still a state which does not require that the agent has any conception of the desired object.

Jerome A. Shaffer presents an analysis of sexual desire that meets this description. His characterisation implies that, even if we do not confuse sexual desire with mere feelings or sensations of arousal, sexual desire is *not* a subspecies of desires (in the qualified sense). “B sexually desires A” is normally not just another way of saying “B desires to have sex with A”. According to Shaffer, sexual desire is similar to desires in general in that it is tied to satisfaction or frustration. But unlike desires, sexual desire is not fulfilled or realised by some “pre-envisaged state of affairs” (1978 p.184). B need not have any idea about what the satisfaction of his sexual desire for A could consist of — what the “getting” of A might be (he may sexually desire A without being sincerely interested in “having” A at all). A is the “object” of sexual desire only in the sense that she is seen as a source of satisfaction. And “satisfaction” does not here mean anything like realisation of the propositional content of the desire — but is rather used as an overall term for certain types of bodily events and sensations which

"are centered in the genital area and radiate out from them". (p.186) To sum up, sexual desiring (in a certain sense) resembles desiring in that it implies that something or someone is desired. But the mere occurrence of the sexually desired object is not what satisfies the sexual desire. Sexual desire "is not a desire of any special amorous practice /.../ In a general way [sexual] desire is not a desire of *doing.* /.../[Sexual] desire can not posit its suppression as its supreme end nor single out for its ultimate goal any particular act" (Sartre 1958 p.385) This non-propositional view of *sexual* desire is compatible with a propositional analysis of desire.

Shaffer's proposal shows the possibility of distinguishing between three notions in connection with sexual desires. Firstly, lust as a mere sensation of feeling sexually aroused – a feeling without direction. Secondly, lust as a state directed towards an object seen as a source of satisfaction, but a state which may come without any representation of the state of affairs that would satisfy the desire, and also without any disposition towards realising such a state. Thirdly, lust as a desire in the BD model's sense, i.e. a dispositional state with a representational content — 'having sex' — such that the agent under certain conditions will tend to realise that proposition.

I believe that the second notion, which Shaffer wants to expose as specific for our sexual motivation, can be applied to other types of motivations related to bodily needs as well. "Hunger" may refer to the desire to have food, or to the felt signs of an empty stomach, but also to a directed state that does not represent its satisfaction in terms of some state of affairs. It is e.g. imaginable that a hungry infant perceiving its mother's breast simply "wants" it, without imagining itself as "having it" at all. No self-consciousness would be needed to be in such a state. Similarly, we might imagine an otherwise normal person with a limited neural damage making her forget everything about how to eat. In the light of cases described by Oliver Sacks in*The Man Who Mistook his Wife for a Hat and Other Clinical Tales* and other books, that does not seem to be a too unrealistic example. She would feel pangs of hunger, and furthermore, it is likely that her attention would be directed towards food on the table, if it smelled and looked appetising. Still, she would not represent herself as eating that food, or be in a state such that the typical kind of circumstances would trigger her eating. The important thing here is that the type of directed state she is in must be distinguished from BD model desires.

Summary of 4.1: The BD model does not presuppose that desires must be felt, experienced or accompanied by any kind of sensations. It is, however, understandable that many of the model's critics have thought so, since the term "desire" is often employed in a primitive sense, typically referring to sensations related to common bodily needs. Furthermore, there is still another notion of desire in use, which refers to a felt state directed towards an object — as in "he sexually desires her". The important difference between this notion of desire (which can be applied to other areas than the

sexual) and desire in the BD model's sense is that the agent's state in this case is not set upon the realisation of some state of affairs.

4.2 First Person Knowledge of Desire and Action

In "Desire" Richard Swinburne states that the "real trouble" with a dispositional definition of desire is "that it would be odd to suppose that a man's desires were totally unknown to him" (1985 p.432). Therefore he supposes that "my desire *is* my mental set which gives rise (barring exceptional counter-evidence) to the belief that I will act" (p.433 my emph.). This mental set is also characterised as "an organized readiness to do the act when I believe that I have the opportunity to do it, a readiness *of which I am aware*" (p.433 my emph.). Note that Swinburne's suggestion is proposed as an alternative to the dispositional analysis of desire. It is fair, therefore, to suppose that he is not only claiming that desires entail the possibility of first person predictions, but that first person predictions are part of what *constitutes* a desire.

The first sentence seems to identify my desires with my predictions about my behaviour, the second with my assumptions about my dispositional states, i.e. about my desires. Knowledge of desire is presented as essential to having desires. Swinburne makes explicitly clear that if my prediction is false it is still a better guide to my desires than my action-dispositions are. "I believe that I would choose the éclair, but in fact I would chose the rice pudding. Which do I now desire? I suggest that ordinary usage favours the answer 'the éclair'" (1985 p.433) (Swinburne's proposed identification of desires with predictions comes close to another identification which is the target of Anscombe's "direction of fit"-criticism in *Intention;* intentions with predictions.)

Sometimes we desire to do things we suspect that we will *not* do when we get the opportunity, and ordinary usage must also be subtle enough to allow for self-deception. It is a fact that people often are mistaken, even without being victims of the wilful act of self-deception, about what they desire. Swinburne's éclair-example could be a quite unproblematic illustration of this. Swinburne presupposes links both between desire and action and between desire and introspective knowledge. In a similar way, David Gauthier writes that we "must distinguish a behavioural dimension of preference revealed in choice, and an attitudinal dimension expressed in speech". He thinks that a conception of "preference" concentrating solely on revealed preferences would be too impoverished a concept. (1986 p.27) It is correct that the explanatory value of a "revealed choice" concept of preference is lesser than the BD model's more substantial notion. But both signs mentioned by Gauthier – what we say and believe *about* our desires and what we then actually do – cannot be conceptual criteria of desire. It is a brute fact that they often come apart. Speech is also intentional

acting, and as such it is an instrument for desire fulfilment and a criterial manifestation of desire. But the desire expressed verbally, or autobiographically described, need not be identical with the desire that moves you to those speech acts.

There are many cases where people desire things but do not have any beliefs about what they would do in a certain situation, e.g. when they have not been giving it any consideration at all. For example, I think it would be all right to say that I desired, even before this example came to my mind, to have a meal this evening, although I had not been thinking anything about what I would do if it should appear to be impracticable. “After all”, Pettit and Smith argue on p.574 in “Backgrounding Desire”, “the evidence of intuition and introspection - the phenomenology of deliberation - is squarely against the hypothesis that desire always has a foreground presence. We are no more inclined to think that the deliberating agent always considers his desire-states than we are to imagine that he always considers his states of belief.”

That description is incompatible with requiring desires to be objects of our occurrent beliefs, but someone could reply that my desire for p at least implies the disposition for thoughts that I desire p[1]. When my attention was drawn towards the matter of supper, I came to think, occurrently, that I desire supper this evening. But is that true of any desire by definition? This is one of the possibilities Pettit and Smith would want to exclude as well, since they stress that “foregrounded” must not be conflated with “phenomenally present”. “A desire may be in the background and be consciously possessed. And a desire may be in the foreground, as in implicit deliberation, without being consciously considered.” (1990 p.568) Even this weaker assumption, that desires must be objects of dispositional beliefs, would be phenomenologically implausible. Someone might ask me to consider whether I desire p, I may search my mind and direct my attention to the matter, and still not know whether I desire p or not — even if my behaviour eventually reveals to the spectator that I actually desire p.

Philip Pettit's and Michael Smith's main argument against the view that desires must be present in the foreground of an agent's deliberation is also concerned with changes in motivation (1988 p.576). I believe the point they are making can be explained like this. Most desires can be satisfied whether or not they are still held by the agent. In contrast to e.g. the desire to smoke whenever I feel a craving for a cigarette, most desires are not, to use Parfit’s term (1986, p.151), conditional on their own persistence. Suppose it was a necessary feature of motivation that intentional action involves both the belief that the action will realise a certain property and the belief that the property is desired by the agent. Then it would be difficult, they argue, to explain cases in which I act upon a desire, which I predict will have vanished by the time it is to be fulfilled. I may e.g. submit an article for publication because I have a desire to air new ideas. At the same time I may know that I will have lost interest in airing those ideas by the time the paper is published. “The prediction will not”, Pettit and Smith argues,

have stopped me acting on the desire to have the paper published, because what nourished that desire was not the prospect of relieving the desire to air the new ideas in the future but simply the prospect of airing them. We believe that any claim to the effect that desire is always foregrounded in decision-making will run into problems of this kind. (p.577)

It is probably possible to modify the epistemic criterion of desire in a way enabling it to cope with self-deception, foreseen changes in motivation over time and other difficulties of similar kinds — at the cost of simplicity. (One might e.g. assume that my present desire to air ideas at a time when I predict that this desire has vanished must be foregrounded in my deliberation at the time when I submit the article for publication, i.e. that the prospect of satisfying — rather than “relieving”, which is somewhat demagogical in this context, as Wlodek Rabinowicz pointed out to me — that present desire in the future motivates my action now.)

But other difficulties may arise if desires are taken to imply knowledge of desire and action. Swinburne’s first characterisation appeared to assume that knowledge of desire partly *defines* what it is to have a desire. Other philosophers have more explicitly placed the agent's beliefs or even verbal reports about his desires among the defining characteristica. (Audi 1973, Pears 1975, Williams in *Moral Luck*. 1981 p.48). But if 'I believe that I desire p' is part of the meaning of 'I desire p', it seems difficult to explicate what it is for me to believe that I desire p. This type of circle would undermine the definition.

A similar circularity may arise from defining desires in terms of predictions. The type of action forecasted by the agent must be supposed to be done at will. Otherwise it would obviously be too wide. Our ability to predict our own reflexes must be irrelevant in this context. On the other hand, at least within the BD model framework, where intentions belong to a subclass of desires, this will also make definiens refer to definiendum.

Within the dispositional view of desire, there must be an intimate connection between knowledge of desire and predictions of behaviour. Since the desires that motivate you have no necessary phenomenal presence, but are essentially distinguished by their relations to beliefs and behaviour, your sole advantage over other people when it comes to knowing your own desires is that you know what the exterior looks like from your point of view. Smith and Pettit could be right in assuming that a desire in the background *may* be consciously possessed (that is not an issue I will debate) but the important thing is that its being consciously possessed has nothing to do with its motivational force. Introspection is therefore notoriously unreliable as a method for detecting desires. To form an opinion about your desires,

you will, like other people, have to think about how you would react under different conditions.

Some well-known types of self-references are improper for formal reasons. I cannot honestly say that I always tell lies, and I cannot really believe that all my beliefs are mistaken — if I believe that, I must also believe that it is false. However, someone else could truthfully point out, without any inconsistency, that I never tell the truth, and she might also know that none of my beliefs are true. As Frederic Schick convincingly shows, first person knowledge of choices and future actions belongs to that group of inconsistent self-references (1999). I cannot think of myself as making a future choice among options, and at the same time be sure about how I will act. The idea of an option is the idea of "an action I neither yet think I will take or that I won't", as well as an action I believe I would perform I chose to. (p. 7).

This is the main point of his argument, put in informal terms. Suppose you know now that I will choose p rather than q tomorrow. If p is an option to me, it follows that I cannot now believe that I will perform p, though it follows, also, that I believe that if I choose p tomorrow, then I will perform p. But if *I* also believe, now, that I will choose p tomorrow, then I must believe that I will perform p. In that case, p is not an *option*. I.e. if I believe now that I will choose p tomorrow, this belief cannot be true. And if it is true, I cannot believe in it. "Others may know this, but we ourselves can't. To that extent our knowledge is bounded, and bounded not by our mental limitations but by our self-discipline" (p.11). Furthermore, as Schick also stresses, we have little to lose if we avoid attempting to make such predictions. This is a point I will stress in a different context further on.

We can indeed deny ourselves the beliefs that I say are improper. We lose nothing of any importance if we avoid such beliefs. Still, we sometimes ascribe such beliefs, if not to ourselves then to others. Sometimes we even endorse ideas that oblige us to do that. We trip ourselves up when we do, so it is well to be cautioned against it. (p. 11)

The impossibility of foreknowledge of one's own choices gives us part of an explanation of the fact that first-person motivational analysis is less reliable than third person analysis. Suppose you attempt, now, to form a judgement about your present state of motivation. You have to take stock of all your motivationally relevant desires and beliefs. Assume, for the sake of argument, that considering your beliefs is unproblematic. How do you detect your desires? Since desires are, essentially, dispositions for action, to judge what you desire must be much like predicting your behaviour. So when you have considered your beliefs, in order to find out what you desire, you have to decide how these beliefs relate to action. The nature of that relation is determined by your desires. The trouble is that any decision about which action that is fitting, given what you believe, in itself must be just that: a decision. Your own

judgements about how you will act given your beliefs are *parts* of your motivation. If you include those decisions in the body of motivation you survey, you disqualify yourself as an agent, i.e. as someone with a choice between options.

My pessimistic generalisation about first person motivational analysis (and its formal explanation) is an important element in my anti-rationalistic approach to akrasia and morality in part II. Empirical laboratory evidence underpinning it is referred to in section 8 (Gopnik 1993, Nisbett & Wilson 1977).

Like the linguistic conception of desire's content, the idea that desires must be foregrounded would have practical consequences for non-human animals. One of R.G. Frey's arguments against animals' desires is e.g. based on such an assumption. He puts forward the following argument: Suppose someone claims that even if animals cannot have belief-dependent desires, there is a certain kind of "simple" desires, like Fido's desires for bones, "which do not involve the intervention of belief" (1980 p.101). As Tom Regan has pointed out, it is doubtful whether anyone would think that it makes sense to attribute a desire for something — a bone — to someone — Fido — while denying that the agent believes anything at all concerning the desired object — like, for instance, that it is a bone. (Regan 1982 p.277)

However, now Frey argues a) that ability to be aware that one desires implies self-consciousness (lacking, it is supposed, in most non-human animals), and b) that it is pointless to attribute desires to creatures who are alleged to be capable of having only unconscious desires. "Where no desires are conscious ones/.../", Frey asks, "what cash value can the use of the term 'desire' have?" (Frey 1980) Tom Regan's criticism of Frey focuses one implicit assumption of the deduction — that 'Fido desires but is unaware that Fido desires' is supposed to imply 'Fido's desire is an unconscious desire'. This inference is invalid, Regan argues, since it overlooks a distinction between "being aware of our desires" as objects "of non-reflective consciousness" and "being aware that we have desires", i.e. to have our desires as objects of "reflective consciousness".

The meaning of "unaware of a desire" and "unconscious desire" is not entirely clear in the context. If read as synonymous, Frey's inference would be valid. Furthermore, it seems a bit strong to characterise the fact that a desire is not an unconscious one, as its being an *object* of consciousness, as Regan does when he assumes that "the desire for the bone can be an object of Fido's simple consciousness" (1982 p.278).

But why is awareness *of* desires essential at all? Could it not be the case that Fido just wants the bone, without being aware or conscious of his desire for the bone? Frey believes that the idea of unconscious desire in humans makes sense "only because we first make sense of conscious desire" (p.104). But for reasons mentioned earlier, awareness of desire cannot be among the conceptual criteria of having desires. We

cannot rule out the logical possibility of creatures, who desire but never think that they desire.

Admittedly, it would be implausible to analyse desire in a way, which makes awareness of one's desires conceptually impossible. But the attribution of desires to beings incapable of self-consciousness need not involve such conceptual claims. Animals' desires are not alleged to be inaccessible in principle — they are not supposed to fall under a different notion than peoples' desires. It is simply a contingent fact that some creatures never can be conscious of their desires (or of their beliefs, for that matter).

It might be the case that Frey's argument about the cash value of “desire” should be interpreted differently, so that it exploits another distinction than the one Regan emphasises. Namely the one between 'being conscious *of* one's desires' and 'having conscious desires' (i.e. be able to experience or feel one's desires). What he in effect says, then, is that it would be pointless to attribute desires to animals incapable of experiencing (i.e. feeling) their desires. Such an argument could be met in two ways. To begin with, this interpretation *would* make Frey's inference from lack of self-consciousness to lack of desire blatantly invalid. From 'Fido cannot know that he desires' does not follow 'Fido cannot feel his desire'. It would be like inferring ‘Fido cannot feel his broken leg’ from ‘Fido cannot know that his leg is broken’. In fact, there are well-founded reasons, based on analogies (neurology, behaviour, evolutionary advantage) for assuming that other higher mammals are able to feel much the same things as humans do. But most important is that, for reasons mentioned, feelings of desire are not *criterial* of having desire. Therefore it is logically possible that there are desiring creatures who for some contingent reason are never able to feel their desire.

Summary of 4.2: The BD model’s dispositional notion of desire is incompatible with the assumption that desires entail knowledge about how we will act, or about our desires. Those assumptions are also phenomenologically implausible. We often desire things without thinking about how we will act, we do not place desires in the foreground of our deliberation and we do not even always have dispositions for occurrent beliefs about our desires.

An agent cannot consistently believe that he will make a certain choice. Other persons may view him as conditioned to make certain choices, but he cannot. This fact undermines the reliability of our beliefs about our own desires, since the identity of desires is determined by what can be conditionally predicted about the behaviour of their bearers.

4.3 Desire and Tendency to Get

“The primitive sign of wanting is trying to get” is an oft-quoted statement made by G E M Anscombe in *Intention* (1957 p.68). Donald Davidson assumes as a necessary truth that “if an agent desires to do x more than he desires to do y and he believes himself free to do either x or y, then he will intentionally do x if he does either x or y intentionally.” (1980 p.23) In *Freedom of the Individual* Stuart Hampshire writes, “'A desires to do X' is indeed equivalent to 'other things being equal, he would do X, if he could'” (1975 p.36). An important presupposition in Hare's theory of prescriptive language is that there is a “close logical relation /.../ between wanting and doing something about what one wants” (1963 p.71). If 'evaluation' is used to cover “mere desire”, then Davidson is right, according to D F Pears in *Motivated Irrationality*, in supposing that there is a necessary two-way connection between valuing and doing (Ch. IX & X). Ingmar Persson's *Reasons and Reason-Governed Actions* consists partly in an elaborated defence of the view “that it is a conceptual truth that in a class of conflicting wants the strongest one is the one that expresses itself in behaviour” (1980, p.101). Alvin Goldman emphasises that “it is a logical truth that wants tend to cause action” (1976, p.112). Harry Frankfurt thinks that the “concept designated by the verb 'to want' is extraordinarily elusive” and that wanting to x is compatible with, for instance, wanting to refrain from x-ing and not “really” wanting to x. He nevertheless admits that to have a “genuine” want is “to be inclined or moved to some extent” to perform the act in question (1971 p.9). All of these declarations express a central theme of the BD model.[2]

The BD model’s conceptual link between wanting and doing can be split up into two: The Forward Connection (wanting implies doing) and the Backward Connection (doing implies wanting), to use D F Pears' terminology (1984). Though these views are closely related, it is conceptually coherent to hold one of them without embracing the other.

None of the quoted philosophers defend a conception of desire, such that there is a straightforward and simple connection between desiring and doing. The ceteris paribus-clause and the hypothetical “if he believes himself free to do” (Davidson) are presupposed, at least implicitly, in these contexts. Desires are not supposed to imply actions, not even to imply action-tendencies — the BD model assumes that desires entail dispositions to act.

Hampshire thinks that “the open, or catchall, phrase 'other things being equal' cannot be replaced by a definite condition, or closed set of conditions, without destroying the equivalence” (1975 p.36). Nevertheless it is possible, for instance from Hampshire's own text, to specify four main types of reservations. If B desires that p, he will necessarily realise p provided that:

• he is capable of realising p,

- his beliefs about how to realise p are correct,
- he has no stronger desire which could be frustrated by realising p, and
- he believes that he is capable of realising p.

Each of these reservations is in itself sufficient to explain why someone desiring that p still might fail to realise p. None of the conditions is a matter of all or nothing. We may be more or less hindered to get what we want, more or less certain about our capabilities and about which means that are proper to our ends etc. It is reasonable to think that most real life choices are influenced by one or several considerations of this kind. Actions result from complexes of desires and various beliefs about means and ability. This means that choice behaviour only under very idealised circumstances would reveal a person's desires straight off.

Why are precisely these four provisos natural? Capability and correct relevant beliefs are evidently required in order for action to give evidence of desire. It is not unusual to be misinformed about the objects of one's desires, or otherwise incapable of realising them. Concerning the third exception Hampshire writes that it "is of the nature of desire that a desire or interest may at any time be prevented from issuing in action by a conflicting desire or interest". (1975 p.36) Most real life choices are between complex outcomes towards which our desires are diverse. R. B. Brandt suggests that

> [if] the valences of the expected outcomes are mixed, the most I would suggest as a possibility is that if each negative product can be matched with a larger positive product there will be a residual action-tendency to perform the act; and if each positive product can be matched with a larger negative product there will be a residual action-tendency not to perform the act. (1979, p.65)

How can the fact that a desire may remain unexpressed because of conflicting desires be made compatible with the idea of a forward connection? In line with Brandt's suggestion, we could describe such hidden desires in terms of counterfactual statements like: "If p, which is an element of the expected outcome, had been the sole motivating factor, there would have been an action-tendency of strength S." Or perhaps better: "If p, which is an element of the expected outcome, had been absent, the action-tendency would have increased/decreased in strength with S-S'." The very point of talking about dispositions for action-tendencies (rather than about action-tendencies directly) is to make room for unmanifested desires.

The possibility of conflicting desires concerning an expected outcome may perhaps appear problematic to the view that what we intentionally do always is desired. If x does p, against which she has a strong aversion (sees her dentist, for example), because p is a necessary condition for q, which she desires strongly, is it not simply

untrue to say that she does something she desires to do? Our linguistic conventions seem to be ambiguous on this point.

However, I do not believe that this ambiguity is inherent in the concept of desiring. It is rather the multitude of degrees of specification in daily speech, which creates the possibility of classifying actions in widely different ways. If the agent were to specify precisely which objects she tries to get or shun, the ambiguity would cease. If the backward connection shall be upheld, the description of her intentional action (what the agent tries to do) must fit the agent's opinion about everything that she thinks the action might lead to in the long run. (Her intentional action, necessarily supported by a desire of hers, is in this sense not only to see her dentist, but to see him in order to get her denture fixed and avoid future suffering etc.) When an agent's intentionally performed action is demarcated in this broad way, it is always true that the desire resulting from all her desires and aversions concerning different parts of her action, in the moment of acting, is a desire to perform the action.

Real-life explanations are rarely complete or fully specific. ‘You desire p’ might sometimes mean merely that you *normally* would desire p. I.e. that whatever your present dispositions point to vis-à-vis p, your otherwise typical inner set-up would produce a tendency to get p when other things are equal. The expression could also inform us that you ineffectively desire p. This means that there is an actual and real inner state of yours, such that it would produce an action *if* other things were equal. E.g. if you were capable, believed that you were capable or had no other dispositions pointing in incompatible directions. Thus, there are philosophically unproblematic senses in which the BD model allows you to do things you do not desire to do.

To avoid misinterpretations, I should stress a terminological point here. Unless explicitly stated otherwise, “occurrent” (when used about desires) is throughout this work used in Brandt’s sense, without implying phenomenal presence (1979, pp.27-28.). An occurrent desire for coffee can be assigned to me when I am in a state such that if the belief that I can get coffee comes to my mind, I will tend to have coffee. I have been discussing desires mostly in this sense. That is to be distinguished from a *normal* desire for coffee, which is the state I am in when I just had enough of coffee for one afternoon, and therefore would abstain from coffee for the moment, even if I came to believe that it was cheap to get. So that just means that under most circumstances, I would have an occurrent desire for coffee. Occurrent desires are not, however, necessarily *effective* desires, i.e. they are not necessarily triggered.

Another, perhaps even more common way of using “occurrent” implies that if B’s desire for p is occurrent, the thought of realising p “occurs to him, occupies his attention, fills his consciousness” (Goldman 1970, p.86). To Goldman, an occurrent want is a mental event or mental process, a “going on” or “happening” in consciousness. A “standing” want, in Goldman’s terminology, is a disposition to have occurrent wants, a disposition “asserted to someone only if he has a number of

occurrent wants for a period of time". My use of the term does neither entail phenomenal presence, or dispositions to experience mental events of that kind. It is distinct from Goldman's standing wants, as well as his occurrent wants.

As Amelie Oksenberg Rorty argues in "The Social and Political Sources of Akrasia", social institutions and economic systems often encourage and foster practices between which there might be great tensions. "[W]hile condemning aggression, they also praise 'aggressive initiative.' While admiring selfless devotion, they also reward canny self-interest. Except in extreme cases, rewards and sanctions do not form a clear and guiding pattern." (1997, p. 652) Weak-willed actions are commonly taken to be theoretical threats to the idea that evaluations express desires, or to the idea that desires entail dispositions to act. Rorty's observations indicate that we should not be surprised to find that the ordinary agent in a given moment is disposed to realising irreconcilable or antithetical goals and that momentary contingencies determine which dispositions that are manifested. Furthermore, as Olav Gjelsvik notes, the traditional philosophical debate often treats akratic acts against an empirically untenable background assumption about unchanged preferences over time. Empirical evidence shows, e.g., that even without additional input of information or affective pressure, the gradual reduction of time-distance to an expected event tends to change our desires and dispositions to make choices concerning that event. (Gjelsvik 2000) In other words, the conventional philosophical worry about akratic behaviour within the BD model springs, at least to some extent, from an underestimation of common complexities, tensions and instabilities within an agent's network of driving forces. What is originally mainly a psychological and perhaps socio-political problem is perhaps thereby unnecessarily turned into a philosophical difficulty.

Another point of importance in this context concerns explanatory and predictive value. Many philosophers have concurred in Thomas Nagel's view in *The Possibility of Altruism,* that the wide dispositional notion of desire is a "logical ghost" (E.J. Bond 1983 p.13). They think that this notion makes the BD model true at the cost of trivialisation, "since anything that moves us (at least to intentional action) is likely to count as such a desire" (Scanlon 1998, p.37). It is true that if the BD model assumed, as a conceptual truth, that there was a straightforward and easily detectable link between real life choices and real desires, then it would be quite uninformative to cite desires in order to understand people's behaviour.

But the BD model commits us to a stronger claim about desires. It does not merely say that desires exist only as inferences from actions and that if there is no action, there can be no desire. Desires are to be attributed to agents if they are moved to action, but also when they *would* be moved, were it not for counteracting beliefs, desires or other circumstances of any of the four kinds mentioned above (see e.g. Foot 1979 p.149). So desires are supposed to exist as real causal forces even when they are

unmanifested. Information about such desires will help us predict and understand people.

In an idealised situation where we have independent evidence showing beyond doubt that none of the reservations mentioned are applicable, we will find an explanation like "I did it because I wanted to" completely uninformative. If the context makes it evident from the start that the agent knows what she is doing and that the most notable condition for her behaviour is her inner constitution, rather than accidental external circumstances, we would find such information useless. Our intuitions would be in line with the BD model then. But even the meagre report that someone wanted to do what she actually did is normally of explanatory value, since people mostly would only tell us this in situations where we otherwise are expected to believe that the action was unintentional, misconceived or accidental in some other way. Furthermore, *if* they tell us this as an explanation of what they do, the conventional way of speaking about dispositions gives us a *right* to assume that they want us to understand that such external explanations could be expected in the context. Otherwise it would not be informative to exclude them.

The fourth typical reservation concerning a desire's giving rise to behaviour states that the agent must believe that he is capable of realising p, if the desire shall initiate a tendency towards p. This suggestion should not be given a stronger interpretation than necessary. The weak justification needed to make my ϕ-ing intentional requires merely, I would suggest, that I regard ϕ-ing as a means to the realisation of some state of affairs desired by me, and that this means-ends belief is among the causes of my ϕ-ing.

Sometimes, e.g. in simple immediate actions, the desired state of affairs might simply be the one in which I am ϕ-ing. The acknowledged instrumentality of what I am doing must not be thought of in terms of a causal relation between my action and its effects. My ϕ-ing is caused by my desire for some p and a belief that ϕ-ing in some sense leads to p. In some cases, "leads to" means that p is a possible effect of ϕ-ing, i.e. a causal consequence following upon the action, but an effect which nevertheless affects the action-description. An example: One of the actions I am about to perform right now is to publish this book. This action goes on as long as my beliefs and desires to do so supports behaviour tending to realise the publishing of the book. These motivating reasons will probably, as far as I can see right now, continue contributing to that kind of behaviour at least until it is too late to withdraw the book from the publisher. My action of publishing the book goes on just until then, but that description of the intentional act can not be fixed before some further effects are known — like if the book really gets published. Even if I will regret having it printed immediately after the point of no return, I will by have performed the intentional act of publishing the book. So in this case, ϕ-ing precedes and causes an instantiation of p, i.e. of the proposition contained in my rationalising desire.

In other cases, the realisation of p may not be thought of as caused by, but as instantiated *in* my ϕ-ing. I may view my action of eating a certain dish as constituted by the desired eating of something filling. Or I could regard my public uttering of certain words as an instance of making an apology. If I see my ϕ-ing, or an element in it, as an instance of p in this way, this is enough to fulfil the instrumentality requirement of my motivating reasons. Michael Bratman agrees with Gilbert Harman (in discussing Kavka's so called Toxin puzzle, to be commented in section 5) that "if one does somehow come to have a new, reason-giving intrinsic desire to drink, that may make it instrumentally rational to drink." (Bratman 1998 p.27) "Instrumentally" could here be understood in line with my weak characterisation. My belief that drinking is an instance of a state of affairs, the realisation of which my present desire is directed towards, is in this case a means-ends belief in a non-causal sense.

As I noted before, it is not necessary that my desire for that state of affairs is present as an object of my deliberation. Since my motivating reasons for action can be unknown to me, it is quite possible that I ϕ intentionally without having the belief that I am ϕ-ing, i.e. without knowing what I am doing – in a literal sense. I need not picture myself *as* ϕ-ing, or as coming to ϕ, in order for ϕ-ing to be caused by my desires and instrumental beliefs in the right way. Although I cannot intend to ϕ unless I regard ϕ-ing as contributing to something I desire, intentional ϕ-ing does not *conceptually* require that I have the positive belief that I am able to ϕ. It is enough that my ϕ-ing is caused by desire and means-ends belief. I do not even find it empirically unreasonable to suppose that many of my actions are caused by such assertions and desires, without the aid of any positive beliefs about capability to perform those actions.

Admittedly, it does seem difficult to imagine a person, who intentionally ϕ-s (he believes that ϕ-ing will contribute to the realisation of p and his behaviour is caused by the desire for p, which is triggered by this belief) and at the same time is positively convinced that he is unable to ϕ. Such beliefs are typical inhibitors of the triggering of desires. But I am not sure that it would be *incoherent* to describe him as ϕ-ing intentionally. If you believe that you are incapable of ϕ-ing, it is unlikely that you think of your ϕ-ing as an effective instrument for anything, and the type of instrumental belief that could rationalise your ϕ-ing will probably never occur to you. That is improbable, but as far as I can see, there is nothing in the conditional assertion 'my ϕ-ing would with some probability lead to p' to *entail* 'I am capable of ϕ' (or 'I will ϕ' for that matter.) Therefore it is at least a conceptual possibility that you ϕ intentionally while believing that you are incapable of ϕ-ing.

It would be a mistake to assume that it is conceptually impossible for a person's ϕ-ing to be *caused* by a conditional means-ends belief that ϕ-ing might lead to p, and a desire for p, in the presence of his belief that he is unable to do it. The important question here is not whether that kind of causal and conceptual relation is conceptually

possible, but whether his pessimistic conviction by definition disqualifies his reason from being the kind of rationaliser that allows us to view him as ϕ-ing *intentionally*.

This illustrates the question of how thin the rationalising or justificatory function of BD model reasons should be taken to be. To view a creature as acting intentionally is to apply a belief-desire scheme of concepts to it. This scheme enables us to see means-ends rationality in its behaviour and in that weak sense, the BD model identifies intentional behaviour with rational behaviour. However, the model allows intentional behaviour to be irrational in a variety of different thicker senses. Most of us regard the person who intentionally ϕ-s, in spite of his mistaken conviction that he is unable to do so, as an agent who is acting irrationally, given his own beliefs. Many adherents of the BD model want to exclude that possibility by definition, and say that this type of irrationality is enough to disqualify his act from being intentional.[3] That is perhaps mostly a matter of terminological conventions.

For practical reasons, I think that the notion of intention should appeal to the thinnest possible concept of rationality. There is no reason to endow agents with more rationality than necessary. This strategy is also a way of forestalling a common objection to the BD model, put forward e.g. by Annette Baier against Davidson, that the model on some characterisations make agents unduly rational, and that it is counterintuitive in that sense. Furthermore, building the thicker concept of rationality into intentions (excluding intentional actions contrary to pessimistic beliefs of the kind mentioned) would make the BD model leave a distinct category of behaviour out of the account.

Suppose we have a quarrel in my garden, and when you are out of further arguments, you kick down my solid-looking garden shed as an effect of your sudden desire to do me some damage. Your act is caused by your desire to make me sorry, and by your belief that your kicking down the shed would have that effect. Throughout your act, you are also convinced that no one could make what looks like an unusually stout little building fall over by kicking it. It might be psychologically rare for behaviour to be caused under such circumstances, but it is possible. Your action differs distinctly from reflexes, misdirected attempts and other typical forms of unintentional behaviour, since it is caused by means-ends beliefs and desires with the right kind of content. So, if what you do is not intentional, what kind of action is it?[4]

Summary of 4.3: The BD model makes it a conceptual truth that actions are outcomes of desires and that desires produce dispositions for behaviour. It entails, therefore, one sense in which it is trivially true that people always do what they want to do. However, the BD model's realistic conception of desires allows desires to exist but remain unexpressed for a lot of different reasons. The model admits greater diversion, contrariety and unsteadiness among an agent's desires, than many of its critics appear to have thought. With that in mind, it is easy to see that the model permits people to

act against their own desires in a variety of ways. Nor does the BD model undermine the normal explanatory and predictive force of citing an agent's desires. Even in cases where an agent simply cites a desire to do what she did as an explanation of that behaviour, we would, normally, at least learn that she finds the situation to be such that we might have expected some external explanation instead.

When a belief of the type 'my ϕ-ing will contribute to the realisation of p' triggers a desire for p so that these states tend to cause p, an intentional action takes place. In the absence of appropriate means-ends beliefs, desires will remain unmanifested, as causal warrants of certain conditional assumptions about the agent. An instrumental belief's triggering of a desire to realise p can also be blocked by stronger conflicting desires and by various beliefs, such as the belief that I am incapable of ϕ-ing. Unlike stronger conflicting desires, the latter kind of belief is, however, not an inhibitor for conceptual reasons, i.e. it is not conceptually impossible to have desires to realise p by ϕ-ing triggered in spite of such beliefs. It may be a psychologically odd form of [5]practical irrationality, though.

4.4 Another Objection to Desires as Mere Inference-Licences

I have made clear that the BD model's notion of desire is dispositional and realistic at the same time. To assign a desire to someone is to claim something about his inner life. But adherents of the BD model often characterise the dispositional notion of desiring simply in hypothetical terms, as a relation between a possible belief and its effects on action, without stressing that this conditional relation must be supposed to be backed up by facts about the agent's inner states. I will now add some further comments about this way of viewing desires.

Hampshire claims that if A believes that he could do X, then "'A wants to do X' is indeed *equivalent* to 'other things being equal, he would do X, if he could'" (1975 p.36 my ital.). The implication of this is that when we assign a desire to someone, then we are just making a conditional prediction about him. Michael Smith appears, momentarily at least, to presuppose something similar when he thinks that the dispositional analysis of desires meshes with the "directions of fit"-approach. The direction of fit of an intentional state concerning an object p is, namely, then characterised (roughly, and somewhat simplified, Smith admits) purely in terms of its counterfactual dependence on a perception of p, and the effect of that perception on the action-tendencies of the subject (1994, p.115).

This characterisation runs the risk of turning the dispositional notion of desire into nothing but an inference-licence. That would be in line with the behaviouristic ambition of the "revealed preference" approach contrasted with the BD model earlier.

The eliminative effects of that move alienate the notion of desire from any realistic allegations. It reduces desires to what we regard as (inconclusive) evidence of desire, when we apply the BD model. To paraphrase Hick's disclaimer about the theory of demand again, the BD model really depicts people with some "pretence to see inside their heads".

One minor objection to the inference-licence view, in defence of the more realistic conception, is this. In terms of conditional predictions, there appears to be no intrinsic difference between a situation in which A desires that p, and one in which she would desire that p, *if* she came to believe that she could obtain p. But ordinarily we would only regard the first as a state of A where p is *occurrently* desired by her.

It could also be argued that such a conditional prediction might be true of beings who are totally ignorant of the object of the desire which thereby is assigned to them — like the desire of a new-born infant to end the conflict in the Middle East. It might, as a conceptual possibility, be true that if it occurred to her that she was capable of reaching the object, she would try to do this. Though perhaps false, the attribution would not be meaningless. In other words, the notion is far too wide.

In *Interests, Utilitarianism and Moral Standing,* Dan Egonsson argues that this type of objection really is harmless (although his own preferred definition of desire, in terms of internal states, appears to be unaffected by it). In order to make the hypothetical predictions true of creatures incapable of having a clue about what they desire, we would have to assign (counterfactually) several other capabilities to them as well, such as ability to act upon the belief. Therefore it will, at most, make sense to attribute *hypothetical* desires to them, not "*actually* dormant wants"(p.79). One might therefore object to my assigning political ambitions to the new-born child, that there are holistic restrictions on formation of belief and desire, which require that I assign a whole network of beliefs and desires to the infant before the isolated counterfactual hypothesis becomes true.

But if 'other things being equal, it would obtain p, if it could (and believed that it could)' really is supposed to be *equivalent* to 'it desires p', then I am inclined to think that those countermoves are blocked. Such restrictions on what someone could do, given a certain belief, presuppose that the agent's constitution, given a certain desire, sets limits to the conditional predictions we can make about it. The pure "inference-licence" approach allows me to assign a desire on the assumption that the child would act upon that belief in a world as close to our own as possible, *except* for all the changes necessary for the child to be capable of acting. That is permitted, unless I assume an analysis of the dispositional notion of desire, which lets attributions of desire entail things about the agent's constitution.

Summary of 4.4: An independent objection against defining desires as mere "inference-licences", without commitments about the agent's inner states, is that such

a definition appears to blur the distinction between occurrent desires and dispositions to form such desires under certain conditions.

---

[1] Wlodek Rabinowicz suggested this possible counterargument.

[2] In spite of many attempts to prove otherwise, there is a prevailing suspicion among many philosophers that it is simply a fallacy to combine causal and justificatory claims like the BD model does. See e.g. Gilbert Ryle in *The Concept of Mind*, G H von Wright e.g. in *Norm and Action*, Anthony Kenny in *Action, Emotion and Will* and E.J. Bond in *Reason and Value*.. It is reasonable to think "that such writers as Donald Davidson, William P. Alston, and Alvin I. Goldman have achieved at least their negative purpose of showing that the major arguments against a causal theory of reasons are unsound" (Audi 1993 p.35). Audi contributes further to that tradition of refutations in *Action, Intention and Reason*. (1993) So does Méle in *Springs of Action*(1992 and several authors in Heil's and Méles anthology *Mental Causation.* (1993)

Although it ought to be clear by now how the dispositional account combines the causal and the justificatory claim, it might nevertheless be necessary to forestall the two most common objections of this kind.

To begin with, some simply assume without much argument that if we explain behaviour in causal terms, then we turn actions into mere happenings. E.g. Chisholm (1966), Taylor, R (1970), Melden (1961) and Bond in *Reason and Value*, p. 23-24. They find it evident that causality eliminates agency. But, as far as I can see, this way of reasoning simply begs the question against the possibility of explanations which both are causal *and* rationalising. For further and effective criticism of this view, see Davidson (1963, p.19).

A more substantial ground for reconsidering the possibility of a causal interpretation of reason-explanation is the discrepancy between the law-like character of intentional explanations and, on the other, the contingent relations between the things (beliefs, desires and their effects) they denote.

Within the BD model, it is e.g. a conceptual truism that strength of desires is proportionate to behavioural tendencies. How could there be an ontological counterpart to the logical relation between 'strength of desire' and 'tendency to cause'?

In a discussion of Kenny's argument's against a causal analysis of emotion, J. R. S. Wilson explicates where this way of thinking has gone wrong:

"1 necessarily (or non-contingently) any A is related (or connected) to a B
2 therefore any A is, necessarily, related to B
3 therefore any A is necessarily-related to a B.
1 clearly tells us nothing about the nature of the A:B relation. But it would be possible to move from 1 to 3, which seems to say that the A:B relation is of a certain kind, namely that it is a necessary or non-contingent relation, and from this conclude that therefore it cannot be a causal relation.
This is obviously invalid. Even if necessarily a father is related to some child, nothing follows from this about the father: child relation, and in particular it does not follow that it is not a causal relation." (Wilson 1972, p.22)

A closely related objection invokes Hume's requirement that only independent events can be causally related. Desires, as they figure in reasons, are "subjective states, /they/ have *no objective or independent existence.* They exist wholly within the intentional world of the subject who thinks or feels

them. Hence they cannot qualify as causes." (Bond 1983 p.23) A similar idea seems to be at the core of Thomas Nagel's refutation of desires as driving forces in *The Possibility of Altruism:*

"That I have the appropriate desire simply *follows* from the fact that these considerations motivate me; if the likelihood that an act will promote my future happiness motivates me to perform it now, then it is appropriate to ascribe to me a desire for my own future happiness. But nothing follows about the role of desire as a condition contributing to the motivational efficacy of those considerations. It is a necessary condition of their efficacy to be sure, but only a logically necessary condition. It is not necessary either as a contributing influence, or as a causal condition." (1970, pp.29-30)

If the presence of a desire is a logical consequence of a reason's motivating, then that desire *cannot* be among the causes of motivation, Nagel argues further on. For Nagel this shows that desires "exist" as *nothing but* an inference from the fact that an act has been done. I.e. that it is a logical fallacy to think that desires could be *forces* of some kind. This way of reasoning seems to commit him to a kind of verificationism concerning desires: If desires are unthinkable independently of their evidential effects, then they are reducible to these effects.

Admittedly, like every psychological state, desires are, in Davidson's words, "constituted the states and events they are by a location in a logical space" (1982, p.304). They do not, then, have "independent" existence. But even if every intentional act necessarily is related to a desire, and every desire necessarily is related to a behavioural tendency, it simply does not follow that desires cannot be the causes of action-tendencies. Wilson's argument makes this clear.

"Causality and identity are relations between individual events no matter how described. But laws are linguistic; and so events can instantiate laws, and hence be explained or predicted in the light of laws, only as those events are described in a certain way." (Davidson 1970 p.215)

[3] I believe e.g. that Audi (1992) and Persson (1981) both would exclude as incoherent the possibility that intentional actions can be irrational in the sense that they are performed against the agent's conviction that he is unable to do the thing in question.

[4] For similar reasons, I am inclined to dismiss the distinction between wanting and wishing, which is defended as substantial e.g. by G E M Anscombe (1957 p.67), and Ingmar Persson (1980 p.104). They reserve the term "want" for states of motivation directed towards objects the agent considers to be within his reach. To begin with, this seems counterintuitive, since it turns the common experience of desiring what one believes to be impossible into a misconception.

The distinction between wanting and wishing could be based *solely* on the idea that wishing takes as its objects "propositions the realisation of which the wisher definitely places outside the range of his power" (Persson 1980 p.104). In that case, I believe that considerations of the kind mentioned nevertheless show that wishes (conceptually) may perform the causal and rationalising role needed for intentional action to take place.

Another possibility is that 'wish' is taken to designate a different *type* of psychological state, not only a state with different objects. To me, this assumption has a low prima facie credibility. As Persson notes, "In non-technical discourse, "to want" seems sometimes indistinguishable from "to wish" (p.93). Anscombe thinks of wishing as a state without any necessary connection with motivation. "A chief mark of an idle wish is that a man does nothing - whether he could or no - towards the fulfilment of the wish" (1957 p.67). Persson seems to be subscribing to a similar view when he states that "a wish is the expression of an emotion (of hoping that p, being sad that -p etc.)" (1980 p.185). If wishes are certain

types of emotions, two distinctions are run together here. The second concerns phenomenal character while the first is about objects of the states in question. It seems illegitimate to stipulate that these categories must coincide.

Suppose we reserve “wanting” for dispositional states directed towards objects within reach and “wishing” for a certain type of emotion, directed towards unreachable objects. We will then get two unnamed categories to account for. The dispositional state I am in when I would ϕ, if I came to believe that it could lead to p, which I desire, and the state I am in when I experience an emotion, intrinsically exactly like the one normally connected with wishing, but directed towards an object I consider to be within reach.

	object within reach	object beyond reach
disposition	“desire”	?
emotion	?	“wish”

5 Intention

Motivating reasons are thought of as causes of actions when we think in BD model terms. These causes are reasons for the behaviour they produce in virtue of their making sense of this behaviour as well. I.e. they rationalise it and cause it. Reasons are capable of doing so because they comprise desires alongside beliefs. In order for reasons to rationalise actions, it is not sufficient that they depict the world; they must set goals as well. The causal power of reasons, as well as their justificatory force, is partly determined by their content.

The BD model's account of *intention* "needs only desires (or other pro-attitudes), beliefs, and the actions themselves. There is indeed the relation between these, causal or otherwise, to be analysed, but that is not an embarrassing entity that has to be added to the world's furniture" (Davidson 1978 p.87). (The BD model's 'desire' is suitably broad to replace 'pro-attitude' and Davidson's parenthetical remark is therefore superfluous. Like Michael Smith, I do not find that terminological point substantial. —1987 p.55). My intention to ϕ is not metaphysically separable from the state in which my desire for some p is triggered by my coming to believe that ϕ realises p, so that these states together tend to realise, causally, a state of affairs determined by the content of the belief and the desire. In other words, intentions are reducible to belief, desire, behavioural tendency and the right kind of causal relation between these entities.

5.1 Intention as Executive Desire

W.P. Alston expresses a BD model accommodation of intentions. His reductive proposal states that an intention to ϕ is an 'executive' desire to ϕ: "a desire that has come out victorious over any immediate competitors and that will therefore trigger off mechanisms leading to overt movement provided that the relevant mechanisms are working normally." (1974 p.95) Alston identifies the intention with one element of the belief-desire complex that causes behaviour. One might also choose to regard the whole motivating reason — operating beliefs and desires — as the intention. Both these identifications, though somewhat arbitrary, are in line with the BD model, and we often appear to talk about intentions in that way, as if the intention is an entity which underlies behaviour.

I see no objections to any of these suggestions. On the contrary, these are the most plausible solutions concerning how to understand our use of the noun 'intention'. We sometimes *form* intentions and we think that the entity thus coming into existence *is* something. In line with the account sketched earlier, we might understand these expressions as referring to the internal state causally responsible for the ongoing manifestations of the triggered desire. On my reading of Alston's proposal, emphasis should be laid on the assertion that an intention is identified with a desire that *has come out*

victorious. Intentions do not exist antecedently; desires become intentions only insofar as they are manifested in behaviour. Deliberative forming of intentions is an active procedure affecting the triggering of desires. In this sense, intentions can be thought of as "coming into existence," sometimes as a result of deliberation.

It is important that the identification of intention with executive desire does not mislead us to think that intentions in principle could be separated from actualised action-tendencies, and characterised in other terms. Davidson's "pure intending," literally understood as "a state or event separate from the intended action or the reasons that prompted the action" is a contradiction in terms, on my BD-reductive view. Desires or belief-desire complexes of a certain specifiable *kind* are therefore not needed to trigger intentional action. The attractiveness of the reduction lies partly in that it frees us from having to specify the essential qualities of intentions in any other way — i.e. in any other way than in terms of a certain relation between belief, desire and behaviour.

A reduction of this kind does not distinguish the strength of a desire from its tendency-triggering character. This is its main flaw, according to Alfred Mele. He argues that many intentions, which are executive states, are formed in significant part on the basis of evaluative assessments of the objects of our desires. Second, he states that "there is considerable empirical support for [this thesis:] The motivational force of our wants (broadly constructed) is not always in line with our assessment or evaluation of the "objects" of our wants, that is, the wanted items" (1992 pp.163-164). In short, Mele resists reducing intentions to beliefs and desires because he thinks that we may intend to do other things than our strongest relevant desires point to, partly because intentions are formed by evaluative assessment rather than by desires.

An important premise for Mele's conclusion is that evaluative assessment and strength of desires can come apart. Beside various imagined cases, Mele's most important evidence for that assumption are Walter Mischel's well-known behavioural experiments with children. These experiments also play the main role in R. B. Brandt's arguments for 'adequacy of representation' as a separable motivating factor. (Brandt's suggestions will be discussed further on.) The constant feature of Mischel's experiments is that the children, after explicitly ranking different kinds of snacks, are told that they can have the lower ranked item on request, any time. They have to wait a while to get the explicitly preferred snack, and they cannot get both. Other elements in the situation are varied; Rewards of different kinds, slides of the preferred food being shown, instructions about how to think of the food (as chewy and sweet marshmallows, or as cotton balls etc.) preferred food in sight, etc.

"To put the experimenter's conclusion simply, attention to the consummatory features of the snacks increased the children's motivation to request the lower ranked snacks, even though these very features were the basis for the children's ranking of the snacks." (Mele 1992 p.164, see also Brandt 1979 pp.61-63) It is worth noting that a less surprising conclusion,

pointing in the opposite direction, can also be established from some of Mischel's experiments, as described by Brandt and Mele. When the children can watch pictures of the higher ranked snack, they become *more* inclined to wait for it. Pictures appear to guide their attention to valuable features of the snack, without making the children so peckish that they choose proximate gratification. Only when their attention is directed to real food on the table before them, or they are told to think of the pictures in a certain way, i.e. as real food, they become more disposed to take the immediate but lower ranked snack.

Both Brandt and Mele regard the somewhat surprising conclusion (that some types of attention to features of the preferred snack make the children take the lower ranked snack) as an illustration of "essentially the hoary philosophical problem of 'weakness of the will'" (Brandt p.60). In *Irrationality*, Mele argues that these cases illustrate how heightened attention to certain features of the object (say, like getting to smell the higher ranked snack) affects the motivational force of the agent's desires. (1987 pp.84-93) In line with his account in *Springs of Action*, he would say that the self-controlled person succeeds in sticking to the intention he settled with to begin with, on the basis of his initial evaluative assessment of the alternatives. (That could be corroborated by the fact that older children and children with higher IQ were less tempted to choose the lower ranked available snack.)

A weak-willed agent, in Mele's view, is more motivated to realise p than to realise q, in spite of his decisive judgement that it is better to realise q. But on such a characterisation, the irresolute child in Mischel's experiment need not display akrasia. Genuine akrasia requires disparity between the evaluative rankings of alternatives and motivation towards those very alternatives. But in all the cases described, the children initially appear to express their evaluations of *snacks*, while their later behaviour reveals their motivation towards *options* — proximal snack or delayed snack.[1] That does not show how evaluations of alternatives come apart from motivation to get those very alternatives. Further discussion of this follows in ch.8 in relation to normative problems about akrasia.

Suppose that this feature was eliminated from the test situation — the children could be told to rank the snacks, then to evaluate the option of getting the fairly good snack any time, and the great snack in twenty minutes. Assume that they initially set higher value to great snack later than to fairly good snack sooner, and that their resoluteness thereafter varies with age, IQ and various inputs as described. It might nevertheless, to begin with, be questioned whether the children's motivational change vis-à-vis the options really *must* have been produced by attending to the very *same* consummatory features, as those the initial ranking was based on. Slide-shows of the preferred snack or food on the table could have made them think of formerly disregarded features of the lower ranked item as well. This would mean that relevant information had changed their rankings.

Mele's conclusion might also be objected to in another way. Younger children and children with a lower IQ might simply be less stable in their

evaluative assessments. There may be various psychological explanations of such a co-variance. (Such explanations need not *necessarily* indicate that resoluteness in deliberative decision-making is an intellectual virtue. Perhaps age make us more rigid for good and for worse. One could imagine that such a development would have *been* an evolutionary advantage in times where humans were born in widely differing conditions, but then mostly stayed in a certain environment during their lifetime. And performance on IQ-tests may co-vary with evaluative resoluteness as a personality trait, without intelligence being a condition for resoluteness. Resolute persons could be more firmly motivated to display a good result on IQ-tests, evaluative resoluteness could happen to be paired with ability to concentrate on one task for a long period, etc.) The point is that these experiments do not prove that intentions formed on the basis of deliberative evaluative assessment of options may come apart from an agent's strongest motivating desires. They may equally well be taken as illustrations of how dominant desires, *thereby* evaluative assessments, and *thereby* intentions, are sensitive to various subtle changes in the agent's representations of the options at hand.

With reference to Bratman's non-reductive analysis of intentions as planning devices, Olav Gjelsvik remarks that this perspective on intentions "presupposes dynamic consistencies as a norm." (2000 p.121) One might perhaps say that the non-reductive view of intentions imposes time-neutral rationality restrictions on motivational changes, so that options are supposed to be seen as invariant over time, in spite of probabilistic or evaluative changes in the agent concerning the goods brought along in those options. The agent's own imposition of such norms may in itself function as a tool of deliberation and in that sense play a role in motivational change. The supposed identity of options through time may be instrumentally required. But in that case, this norm works as any other element among the attitudes and convictions of our motivating desires. It may be a normal and instrumentally valuable norm, at least in rational adult humans, but to assume that such a device could be a distinct meta-motivational element (an element preserving the identity of options over time) *over and above* the motivating reasons that operate in each moment would be to force the agent to transcend his temporal limitations.

All intentions are not formed under the influence of deliberation. Even when deliberate weighing of alternatives, practical inferences and other varieties of practical reasoning precede them, these procedures will only partially affect the reason-based triggering of action-tendencies, i.e. the intention. Intentions are not automatically inferred from deliberative procedures. This means that if evaluative assessment is understood in the foregrounded deliberative sense (rather than in, e.g., a motivational pro-attitude sense), and it plays a role in decision-making, then it is not at all unlikely that the ranking in such assessments in themselves differs from motivational force ranking. As I have stressed, the BD model allows foregrounded evaluations and desires to deviate from motivating desires. It is clear that the ranking before my mind need not match the ranking I

express in overt behaviour. In that respect, Mele's premise is sound. However, in that interpretation, the premise would not underpin his conclusion, since intentions are not products of deliberation, but of motivation.

Intentions can be seen as victorious desires. In a somewhat trivial sense, the BD model will anyway admit a distinction between strength and efficacy of desire. Victory can be a walkover, due to disqualified competitors. Let me use another analogy. Assume that two substances each are capable of reacting with oxygen and that the resulting compounds, in each case, have a similar effect, i.e. the compound eats its way through metals. Substance α is more strongly corrosive than substance β, which means that it makes the metal corrode more and faster. I.e. the triggered tendency for which α is a causal condition is more thoroughgoing than the tendency β may contribute to. Now we may imagine that the conditions for triggering these effects differ slightly, so that when we pour both substances on a metal plate, low humidity blocks the triggering of α's (but not β's) disposition to make holes in the plate. Still, α's disposition to do that is stronger than β's, in terms of the exactly similar types of tendencies they both would cause when triggered. The executive dispositional property is not necessarily the strongest one.

The realistic dispositional notion of desire asserted before lets us distinguish between strength and effectiveness of desires in a similar way. The BD model's conceptual commitment to the claim that the strongest desire is the one expressed in action must therefore be further qualified. It is true only on the assumption that we talk about strength of desires within a group of conflicting desires, which all are supposed to be triggered by similar external conditions. The triggered desire is not necessarily the strongest one concerning a certain type of outcome, when viewed in relation to a broader set of desires. (Again, this distinction between strength and effectiveness is of course much less substantial than the gap Mele attempts to establish.)

5.2 Deliberative and Future-Oriented Intentions

Recall the child who waits for the higher ranked food, in spite of input that makes other children choose sooner snack before better snack. The BD model's picture of the resolute self-controlled child could be that this is a child for whom the initial executive dominant desire persists and continues supporting the tendency to wait, for some (any) reason beyond the child's control. Some people are just less apt to change their mind after having made it up. Genes and upbringing may make our motivational sets more or less unstable. Resoluteness varies between persons, even in cases where no additional information is added. (See Gjelsvik 2000 on Ainsley's work, which is evidence of this, and my discussion in ch.7)

However, the BD model could also do justice to a more deliberative picture of self-control. I may e.g. resort to external pre-commitment devices

in order to block foreseen temptations to change my intended course. When I started telling people that I am about to finish this book, that was one of the actions prompted by my desire to finish it, together with my beliefs about the means to achieve this end. (I knew I would be less inclined to change my mind at the cost of appearing wishy-washy.) That intentional pre-commitment was an overt causal tendency in the direction of the goal represented in my desire.

Even when such devices are entirely internal, they are parts of the intentional action decided upon by the agent. Such internal instruments for sustaining a course one has settled upon may include investing energy in thinking about alternative plans and strategies for the goal, deliberately avoiding to direct one's attention to distracting matters, etc. Some people apparently often do things like that, especially when they have formed intentions directed towards the non-immediate future. We can imagine that the resolute child is aware that some kinds of attention are likely to trigger other tendencies of hers and that she therefore avoids attending to such features, etc. The BD model allows us to depict her like this, although such efforts are no necessary parts of intentional actions, whether they are future-oriented or immediate.

Like Bratman and Davidson, Mele views the difference between "future-oriented" and "immediate" ("distal" and "proximate" in Mele's terminology) intentions as significant. Another similarity between Bratman's and Mele's non-reductive views of intention is that they both appear to regard our understanding of intentions concerning actions in the non-immediate future, i.e. distal intentions, as primary for our understanding of intentions in general. *Planning* is e.g. regarded as an important element in intentions. Intentions incorporate executive attitudes towards plans according to Mele. In a definition-like passage he says: "Intentions are executive states whose primary function is to bring the world into conformity with intention-embedded plans."(1992 p.162, Bratman 1996, Davidson 1978).

Audi remarks in defence of a BD-reduction that Mele's objections to the reductive account appear plausible because "Mele is conceiving intending as arising from something like assessing options." (1988 p.244). The positive analysis of intention Mele subtly elaborates in *Springs of Action* retains this feature (1992 ch.9-10). To form an intention is to "settle things" (1992) or to settle "a first-person practical issue" (1988 p.241) Audi is clearly right in ascribing this deliberative view of intention-formation to Mele. It is essential to Mele's main argument against the reduction that intentions and motivationally dominant desires often come apart — and that they do that because intentions, unlike desires, are based on "our assessment or evaluation of the "objects" of our wants, that is, the wanted items."(1992 p.163.)

When discussing intentions for things in the non-immediate future, it appears quite plausible to view planning, choice of strategies etc. as important elements incorporated in the intention. Furthermore, I have no objections to the idea that distal intention-formation, i.e. decisions about

what to do in the future, typically are preceded by deliberative assessments of options. In that respect, Alston's formulation of the reduction is too hasty. It gives the impression that no intentions could be deliberately affected by other courses of action taken by the agent, such as thinking matters over, dwelling upon ranking of options and the valuable features of those options. The idea that some desire "has come out victorious" leads the thoughts to a somewhat random procedure. Another oversimplification is that the passage identifies intentions with desires that "trigger off mechanisms", which seems to indicate that intentions then typically immediately withdraw. I would prefer to say that beliefs trigger *desires,* and that beliefs and desires in most cases play a continued causal supporting role. A third unfortunate choice of terms in Alston's formulation of the reduction is that the executive desire is supposed to trigger off mechanisms that lead to "overt movement."

None of these three features are characteristic of future-oriented intentions. Intentions concerning what to do at a later time are often formed under the influence of some deliberative assessment of alternatives. Such intentions are usually not only triggers, but upheld as sustaining motivators of the initiated action-tendency. The tendency towards realisation of a future desired state of affairs does not, either, necessarily begin with overt movement. Instrumental reasoning of various kinds, including, possibly, reasoning about my own anticipated future intentions, may well be parts of the initiated tendency.

What I am getting at is that BD model reductionism about intentions, properly expressed, can accommodate Mele's observations of features typical of distal intentions. Let me illustrate further, to make that clearer: These signs were typed intentionally. They were the result of overt behaviour caused and rationalised by my beliefs and desires. My intention to write precisely these words was formed and executed almost simultaneously, without any intermediate planning or other types of reasoning. That kind of action would fit well in with Alston's reductive account. Even on Mele's view, "proximal intention plays roughly the triggering role identified by Alston."(Mele 1992, p.173) The fact that I intended to write these words was not a fact about a separate state of mind, formed under the influence of deliberation. It simply consisted in that I did what I did under the influence of my contemporary beliefs and desires.

A while ago I formed an intention to write down some comments concerning intentions. My writing of the words above partly fulfilled the intention. Suppose something had stopped me from executing that intention — it would never result in any writing. How could I then claim that intending is equivalent to having desires triggered into actions? The answer is that actions can include a variety of non-overt activities directed towards the realisation of the pre-set goal: Being pre-occupied with manners of expressing myself is one such activity. The triggering beliefs and desires sustain and motivate these activities as long as the motivational situation is unchanged in other respects. In that sense the intention exists as long as the process continues. But it is in the nature of tendencies that they can fail to complete their process.

Sometimes, also, I form a few sentences in my mind, and try to figure out which one is best, *before* deciding upon what to write. I.e. I deliberate a little, before settling things. But then, how can my intention simply be the result of the strongest desire in my mind, among those triggered by what I come to believe for the moment? Here, the answer is that the BD reduction involves no restrictions concerning how the triggering conditions are obtained. I may choose to direct my attention to matters especially relevant to a certain course of action (in this case perhaps think about some potential readers, etc.), and intentionally influence my decision in that indirect way. My intention still adds nothing to the story. When desires are triggered and cause the behaviour they rationalise, that behaviour is intentional.

Although the BD reduction can do justice to these features, which are most typical of intentions directed towards the non-immediate future, it does not give them status of criteria for intentions in general. They are not even essential to future-oriented intentions.

5.3 Pure Intending

Among the non-reductionists about intention, Mele lists Donald Davidson, due to the views he expresses in "Intending" (1978) and in his replies to critics in Vermazen's and Hintikka's anthology (1985). Davidson has, initially, no problem with reductionism in descriptions of immediate intentional *actions*, where intentions are formed and executed simultaneously. His doubts about reductionism arise from the assumption that the existence of "pure" intending must be recognised. (1978 p.89). Pure intending "is not necessarily accompanied by any action." (p.88). Although Davidson finds it possible that pure intending has some essential features not shared with all other cases of intending, "it would be astonishing if that extra element were foreign to our understanding of intentional action." Future-oriented intentions are especially relevant here, he thinks. "It seems that in any intentional action that takes much time, or involves preparatory steps, something like pure intending must be present." And when that much is admitted, a stronger claim can be defended. "Once the existence of pure intending is recognized, there is no reason not to allow that intention of exactly the same kind is also present when the intended action eventuates." (1978 p.88) So, like in Mele's account, distal intentions are supposed to make a strong case for non-reduction.

Why do we have to admit the existence of pure intendings? Davidson's evidence is that some intentions can be had, and formed, "without conscious deliberation or overt consequence." I accept that, but I do not think that this observation "leaves no doubt that intending is a state or event separate from the intended action or the reasons that prompted the action."(p.89) Note that Davidson's evidence of pure intendings is cases where intentions are described as having no *overt* consequences, and being based on no *conscious deliberation*. Elsewhere he employs other modifiers: Pure intendings are to be "abstracted from *normal* outcomes" (my ital.) and a

certain analysis of pure intendings is dismissed because it fails to give an account of an action that is "familiar or observable" (p.90). What all these expressions implicitly suggest is this: The various motivational states we regard as instances or proofs of pure intending are nevertheless examples of intentions tied to reasons, but not to deliberative, foregrounded or phenomenally present ones. They are also bound up with actions, but with non-overt, non-familiar or non-observable actions.

Davidson examines the possibility that pure intending *is* an action. Since intending is not a change or an event, it cannot be something the agent does, he argues. Therefore, the thesis must be that "the action is forming an intention, while pure intending is a state of the agent who has formed an intention."(p.89) So, what he goes on to discuss is whether the *formation* of a pure intention in itself could be an action. His objection to that account is that "the purported action is not familiar or observable, even to the agent himself" (p.90). He also discusses theories to the effect that intentions are actions like speech acts — promises or commands. Davidson is roughly right, I believe, in claiming that "to point out that promising and commanding, as we usually understand them, are necessarily public performances" is enough to discredit these theories (p.90).

As I have already hinted at, there is a more natural but less literal understanding of the idea that intentions can be abstracted from actions. The solution is implicit in Davidson's own choice of words. Beliefs and desires cause and rationalise action-tendencies that are neither overt, nor based on deliberation. In many cases these motivating reasons persist and support the course of behaviour until the initiated tendency is completed, in other cases they disappear before that. When you intend to realise something that "takes much time, or involves preparatory steps," your attempts to reach it often, to begin with, consist merely in your mental preoccupation with means and strategies.

In this sense, there appears to be something plausible in Hume's picture of "the will" — read here as synonymous with "intention" — as "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." (1739 3:1:1 my ital.) These impressions may be regarded as auto-simulations, with a somewhat demagogical term borrowed from computer oriented psychologists. That suffices to get the kind of causal relation between reason and action, which allows us to endow you with an intention. Enough has been said, I believe, about how the BD-reduction of intention handles distal intentions, deliberative intentions, and (allegedly) pure intentions.

I am not sure that Mele's non-reductionist label on Davidson is correct. Davidson's own positive account is closer to reductionism than it appears. Let me indicate why. Davidson's analysis gives intentions the role of conclusions from Aristotelian practical syllogisms. Such conclusions are sometimes thought of as actions, sometimes as mere judgements. Now, practical reasoning may result in actions in some cases, but in many cases it results merely in an intention to do something in the future according to Davidson. When the conclusion of a piece of practical reasoning is an action,

that conclusion *can* be expressed by a value-judgement referring to the action. E.g.: “This action of mine, this eating by me of candy now, is desirable” (1978 p.96).

But when a practical inference produces an intention to do something later, i.e. a pure intention, which might not end with the desired result, that conclusion is *just* a judgement, says Davidson. When actions are of “brief duration, nothing seems to stand in the way of an Aristotelian identification of the action with a judgment of a certain kind — an all-out, unconditional judgment.” In the case of pure intending, “the intention simply is an all-out judgment” (p.99). There is no doubt that a judgement is a *form* of propositional attitude for Davidson, and that what he has in mind here is not an assertive attitude, but a pro-attitude, belonging to the same genus of pro-attitudes as desires (p.97, and p.102).

When Christopher Peacocke discusses the possibility of interpreting Davidson's “better judgment” in terms of satisfaction of present desires, he understands the judgement as expressing a “belief in the agent about which course of action will best satisfy his present desires” (1985 p.56). In a reply to Peacocke, Davidson writes that they both “agree that to intend something is to have an attitude towards a proposition. He calls this a disposition and I call it a judgment; but what I call a judgment is a disposition, and I am happy to give up the word 'judgment'” (1985 p.211).

What he does not make clear in this reply, is *which* kind of disposition he regards these judgements to be. Are they dispositions for behaviour, or dispositions for occurrent belief? If the latter should be the case, Peacocke would still be right in pointing out that “to make such a judgment is not yet to have settled the question of what one will try to do.” However, it can be safely concluded from Davidson’s declarations in “Intending,” that he is not thinking of judgement as an expression of *belief* in this context. “No weight should be given the word ‘judgment’. /…/ I do not suppose that someone who wants to eat something sweet necessarily *judges* that it would be good to eat something sweet.” (p.97) The sole difference between pro-attitudes like desires and pro-attitudes like intentions is that intentions are unconditional and can be expressed by all-out judgements, while desires are conditional and merely correspond to prima facie judgements. Both are dispositions for behaviour, and Davidson’s talk of pure intendings as having no *overt* outcomes indicates that even in these cases, he thinks of an all-out judgement as a disposition which is manifested in *some* sense. On an action-dispositional interpretation of “judgement,” the following declaration could well be taken as identifying intentions with executive desires.

> a judgment that something I think I can do — that I think I see my way clear to doing — a judgment that such an action is desirable not only for one or another reason, but in the light of all my reasons, a judgment like this is not a mere wish. It is an intention. (Davidson 1978 p.101)

The non-reductive air of Davidson’s description springs from the fact that “judgement” could be used to distinguish pro-attitudes seen as mental episodes from pro-attitudes seen as dispositions. It could also be read in the

conventional sense (excluded by Davidson's footnote on p.97) implying truth-claims. Both of these interpretations would be natural.

5.4 Intentions and Predictions

Doubts have been repeatedly expressed in former sections concerning the possibility of making predictions about one's own behaviour, on the basis of one's desires and beliefs. But we appear to tie intentions much closer to such beliefs in ordinary usage. As Robert Audi says:

> Note how odd it is to say things as "I intend to go to your paper, though it is not likely that I will make it," "He intends to visit us for the weekend, though he does not believe it probable that he will come," and "She intends to surprise him by coming early, but believes that as likely as not her coming early won't surprise him. (Audi 1993 p.57)

Audi's own belief/desire reduction of intentions says, roughly, that intentions are dominant wants together with beliefs *that one will do the act in question.* It is clear that the BD model reduction suggested here is different, since it cannot admit that intentions are necessarily (partly) *constituted by* predictions. Beliefs with any kind of content can, in principle, trigger the desires they are appropriately related to, and when this happens, an intentional action takes place.

Nevertheless, Audi is right in describing the linguistic behaviour above as odd, and this impression should be explained. Do intentions conceptually *entail* predictions, do they typically but contingently *give rise* to predictions, or is it simply a natural mistake that needs to be diagnosed further, to suppose that there is something inconsistent about the descriptions above? My answer is that these three claims are all true, when read in the correct sense, respectively.

We do not intend to do everything we desire to do. One reason for this is that we typically do not intend to reach the goals we believe with great certainty to be beyond reach. I desire to levitate, but I do not intend to do it. On the BD-reductive account, intentions can be ascribed to a person only when a desire is triggered. Therefore, a minimal *negative* belief-condition for intentions is inherent in the conditionals specifying the relation between most desires and tendency to get. Usually, we do not intend to ϕ and at the same time believe that we are unable to ϕ. Your belief that you are unable to ϕ is a standard inhibitor of the triggering of your disposition to ϕ.

However, as I made clear in section 4, absence of belief about incapability is not a *conceptual* requirement for intentions. It may be a psychological fact that means-ends beliefs rarely or perhaps never trigger desires in the presence of such negative beliefs. Perhaps it also a fact that beliefs of the type "I will not ϕ" typically inhibits triggering of a desire to ϕ. From none of these assumptions follow that your intending to ϕ mostly or always is dependent upon a positive belief that you are able to ϕ.

A consequence of the suggested reductionist view is that intending and desiring also differ in that desires can *precede* action, i.e. they might exist as causal conditions waiting to be triggered by the right kind of beliefs. It is perhaps too strong to suggest that 'Willing implies delay' (John Donne), but it is at least a conceptual feature of desires that they admit delay. As a contrast, when I intend, my action has begun —by definition. Intention never precedes action. Hume's characterisation of intention as a state occurring "*when* we knowingly give rise to any new motion of our body, or new perception of our mind" is correct in this respect (1739 3:1:1 my ital.). That feature of reductionism appears to be the great stumbling-block to non-reductionists — they view prolonged actions as cases where the intention exists before the action, and therefore as proofs that intentions can be abstracted from actions.

Let me therefore illustrate the point once more: When A intends to insult B, her actions are directed by beliefs and desires which actually tend to cause something A believes will be an insult to B. If she has that intention, her action has begun. Her intention does not precede the action. For different reasons, such an intention might fail. B is too stupid to be insulted, or A's insights in the art of humiliating people are insufficient. A can also be physically unable to execute her decision: B stands in a crowd and A cannot make herself heard, or A suddenly suffers from cerebral haemorrhage and utters other sounds than the words she intends to utter. A might also simply be overcome by a sudden sense of compassion and suddenly want to hug B rather than insult him. In all these cases, it makes sense to say that her beliefs and desires *tend* to cause something she considers an insult, as long as her intention persists. It is therefore a mistake to think that prolonged actions show that future-directed intentions can be separated from actions and motivating reasons. A causal chain is initiated, and it is directed towards an imagined goal, even though the causal sequence is interrupted (for whatever reason) while it still had a long way to go before reaching the pre-envisaged endpoint.

Intentions can be seen as future-oriented in two senses. "I intend to visit you tomorrow" might either be understood as a declaration of an ongoing intentional action, where my arriving at your place is an important final component. Seeing you is the desired element instantiated in my ϕ-ing. But the expression could also be seen as stating my present ongoing tendency to affect my intentional behaviour *tomorrow*, which then is seen as an action distinct from the one I am performing now. On the latter description, the intention-statement nevertheless implies that an action-tendency, ε-ing, is going on now and it is directed towards a state of affairs in which I perform another intentional action, ϕ-ing. In that sense, intentions can be seen as antecedents of actions within the reductive view. But this is not a sense showing that intentions can be abstracted from actions.

The fact that you intend to realise p is in prolonged actions likely to give rise to beliefs that you will realise p. Your intention means that you are already heading towards p in some sense. A tendency towards p has been initiated by the triggering of your desire and you are a part of that tendency.

To the extent that you know what you are doing, you believe that you are about to realise p. We are probably unaware of our immediate intentional actions at many times. But when it comes to actions of long duration it is improbable that you are unaware of what you are doing, since such actions typically begin in your mind, with planning, strategies and even internal pre-commitment devices. If you live far from us, and now intend to visit us next week, you are in one way or another already preparing for the trip, and it is unlikely that these preparations are unknown to you. Should you sincerely tell us that you are coming to see us, we are allowed to assume that you have been giving the matter some thought, and that you believe that you are on your way (metaphorically speaking) to do this. 'A intends to realise p' implies 'A is about to realise p' (provided that 'is about to' is read in the weak tendency-sense, admitting of low probabilities, interruption and failure). This goes for self-referential attributions of intentions as well.

Your statement that you intend to visit us may of course be false, not only because you might deceive us, but also because you are conceptually allowed to deceive yourself in this matter. I.e. it is conceptually possible that you do not know what you are doing, and what intentions you really have. The point is that your belief about your intention commits you to a belief about what you are doing.

Audi argues that cases like the ones quoted in the beginning of this section show that when "intending" is used "stringently," then 'I intend to realise p by ϕ-ing'' implies that I regard it at least probable that I will realise p by ϕ-ing (Audi 1993 p.57). From this he infers as a conceptual truth that a person can intend to do something only if she believes that she will (or that she probably will) do it. (p.65) In my view, it is correct that a person could not sincerely assent to 'I intend to realise p by ϕ-ing' without regarding it as probable to some extent that she will ϕ and realise p by doing so. However, her sincere assent is then an expression of a belief about her intention. She can not have that belief without finding it probable that she will ϕ and thereby realise p. But that assumption does nothing to necessitate the conclusion that she can *intend* to realise p by ϕ-ing only if she believes that she will realise p by ϕ-ing.

Let me qualify that conceptual claim a little. Suppose she views herself from a third person perspective, and tries to form a picture of her intentions. She might believe that she strongly desires p, and also believe that she believes ϕ-ing to be a way of reaching p. Nevertheless, she may well be uncertain about whether she is about to ϕ in order to realise p. On the reductive account, this implies that she is also uncertain about whether she is intending to ϕ. She may believe that she intends to ϕ without believing that her ϕ-ing is going on now, but only if she believes that she now is about to do something else, ε-ing, regarded as instrumental to her coming to ϕ later.

I claimed initially in this section that, if read in the correct sense, three seemingly incompatible assumptions are true. Firstly, intentions conceptually *entail* predictions. This merely means that the description 'A intends to realise p' implies 'A tends to realise p'. Secondly, intentions typically *give rise* to predictions. Planning and other mental activities often begin the

tendency triggered in intentional actions, at least when acts are of some duration. In such cases, we should expect people to be aware of what they are intentionally about to realise. Thirdly, to suppose that one cannot intend to do something without believing that one will do it, is a natural mistake that needs to be diagnosed further. That mistake could be a symptom of conflating any of the first two assumptions, or both, with the less reasonable claim that 'A intends to realise p' *implies* 'A believes that she tends to realise p'.

The two conceptual commitments about intention and prediction inherent in the BD-reduction of intentions can be illustrated with Gregory Kavka's so called Toxin-puzzle (Kavka 1983).

Put briefly, the dilemma is this: At t0, a clairvoyant and trustworthy billionaire offers you a billion dollars, which you will receive at t2, if you form the intention, at t1, to drink nauseating poison at t3. The poison makes you really sick for a day but has no other effects. What you *do* at t3 is no part of the deal — no further retributions etc. are to be expected whatever you choose at t3. So, the benefit of acquiring the intention is supposed to be autonomous; it does not depend on executing the intention. Ranking at t0 is evident. 'A billion dollars at t2 plus sick for a day at t3' is better than no money, and 'a billion dollars at t2 minus sick for a day at t3' is even more desirable. You will also anticipate at t0, that your ranking at t3 will be the same. Why on earth should you then get sick for no benefit at all?[2]

Side-bets and pre-commitment devices that could affect future rankings are forbidden. For the sake of argument here, exclude also the possibility that ranking at t3 could be changed by the mere acquisition of an intention to drink the poison at an earlier time. That might otherwise get you the billion. Wlodek Rabinowicz points out that resoluteness and sticking to previous intentions are in themselves desirable features according to some agents. One might perhaps also argue, as Gilbert Harman has done, that the intention to drink in this case must express an intrinsic desire, which then affects my reasons at t3 for drinking the poison. (Rabinowicz 1995, Gilbert Harman 1998, both referred in Bratman's discussion of the Toxin puzzle 1998). Such solutions are, then, also out of the question by stipulation here.

Michael Bratman discusses whether one can *rationally* form the required intention at t1, while Kavka's difficulty concerns, also, whether one *can* form that intention *simpliciter*. Obviously, if conceptual commitments prevent us from forming such an intention, it can not be rationally formed. Kavka states that "you can not intend to act as you have no reason to act, at least when you have substantial reasons not to act" (1983 p.35). Let me interpret that assumption in line with the reductive account of intentions.

On the face of it, it seems that the BD model reduction of intentions should have difficulties in admitting that there could be an autonomous benefit of the intention, since the intention on this view is inseparable from its execution. However, there is a sense in which one could say that an intention to drink poison could be separated from the drinking of it. You may intentionally start your preparations for drinking it, either overtly, e.g.

by avoiding eating things you believe will make the nausea worse, or non-overtly, e.g. by actively avoiding thoughts about being sick and concentrating on the pleasure a billion will give you. But somewhere along the line, your motivating reasons change, and you will not drink it anyway. We could think of this as a case in which you intentionally ε-s at t1 because you regard ε-ing as a means to realise a desired state of affairs at t2, in which you perform another intentional action ϕ. It might also be described as your intentional ϕ-ing from the start, where drinking poison instantiates the desired element of ϕ-ing from your perspective. In both cases your intentions are unsuccessful and withdraw without giving sufficient continuous support to the action-tendency in question.

On the BD model, it is not an open question whether you actually will drink the poison or not. Your motivating reasons at t3, i.e. your strongest desire and your relevant instrumental beliefs, as stipulated, is then against drinking. 'Strongest' implies decisive of behaviour, according to the BD model. This is the minimal rationality requirement conceptually inherent in all intentional action. Akrasia as clear-cut incontinence, i.e. as acting against strongest motivating reasons, is excluded. Behaviour is the final measure of strength. The initial characterisation of your reasons implies, therefore, that you do not drink at t3.

To begin with, this means that it would be difficult for you at t1 to *believe* that you intended to drink the poison. That would imply believing that you are about to do it, i.e. that a tendency towards drinking is already going on (in any of the two senses admitted). A spectator of your actions, like yourself when you take this perspective, may in many cases doubt whether these goal-directed tendencies will be fulfilled in accordance with the intentional states that rationalise them. However, this case is exceptional on the BD model reading, since success for your intention at t1 is not only insecure; it is excluded.

The BD-reductive view implies that you can not have an intention at t1 to drink poison unless a tendency towards realising p has begun (at least in the form of strategies etc.), and this tendency is supported by a motivating reason, consisting of appropriate beliefs and desires. I am inclined to think that it is conceptually impossible for you to regard yourself as "being about to drink poison" if you believe that drinking poison is no available option at all. Note that this is not parallel to examples discussed before. The question then was whether you could intend to ϕ, in order to realise p, while finding it impossible for you to ϕ. And my answer was that this is at least *imaginable,* since your act nevertheless could be caused by the right kind of desires and instrumental belief — like that your kicking down my shed would upset me. But here, the question is whether you can *believe* at t1 that you are about to realise the state of affairs in which you drink poison and be convinced that you will no do it.

However, intentions can be unknown to agents, and if t0 and t1 are separated by some time (say, a year), you might be able to deliberately manipulate your motivation so that you come to have the valuable intention at t1, without being aware of it. That is imaginable, and probably not even

more farfetched than clairvoyant and trustworthy billionaires are. Then you do things at t1, like mentally preparing for drinking poison, as means towards realisation of poison-consumption at t3 (which is something you desire at t1) without thinking of yourself from the third person perspective *as* someone who intentionally does those things at t1.

In the example where you destroyed my garden shed, my point was that you could believe 'ϕ-ing leads to p', without believing that you were able to ϕ. This could be enough for your ϕ-ing to be caused by your desire for p. In that case, I saw no reason not to label your action intentional. However, I am less inclined to admit a similar possibility here. Even if you could manage to have the intention of drinking poison at t1 without being aware of that intention, I believe there might be other conceptual reasons against the possibility of forming it.

The deliberate forming of such an intention would force you to acquire an instrumental belief you know to be false. The minimal instrumental belief required is that *if* you do the things characteristic of having an intention at t1 to drink poison at t3, this will contribute to the realisation of your drinking poison at t3. *I* could induce the right kind of instrumental beliefs in you at t1, and at the same time be certain that the beliefs of yours are mistaken. I am, though, inclined to think that *you* could ascribe a belief recognised by yourself as bluntly mistaken to your (other) self only if you embodied genuine multiple personalities.

Admittedly, person's can willingly manipulate their beliefs. A person following Pascal's advice may place his bet on God's existence and then deliberately avoid hearing and thinking about things that could undermine his faith, and concentrate on rituals and thoughts strengthening it. "Although you cannot believe by simply deciding to do so, you can come to believe by deciding to cultivate belief" (Mackie 1982 p.202). In Pascal's argument, the ultimate truth of the matter is supposed to be inaccessible to reason, and the desired belief does at least not, it is alleged, contradict other plausible beliefs of yours.

In the toxin case, things are different. I have taken for granted as a prerequisite that you know, at t1, about your anticipated motivational attitude towards drinking the poison at t3. Some philosophers, e.g. Robert Nozick and Frederic Schick, doubt that knowledge is closed under known logical implication, i.e. they deny that if a person knows that p entails q, and he knows that p, then he knows that q (Nozick 1981 pp.203-211, Schick 1991). But I guess that most of us would find the principle of closure sound (though this is a matter beyond the scope of my ambitions here). If it is sound, you cannot know that the motivating reasons you will have at t3 entails that you do not drink the toxin, without knowing that you will not drink the toxin. Then you cannot without blunt inconsistency have the right kind of instrumental belief. And it is even more questionable whether it is theoretically possible to entertain contradictory beliefs knowingly in this way.

Perhaps it is no explicit prerequisite that you know at t1, about your motivation at t3, although that seems to make the case less challenging. But

it is certainly an important presumption that you know the full story at t0, when you are offered the deal. At t0, you know about your motivating reasons at t3, and therefore that you will not drink poison at t3, whatever your beliefs and desires at t1 will be. So you would then have to manipulate your state of mind so that you forget this knowledge and come to believe that your intentional activities at t1 really do contribute to the realisation of your drinking. It seems apparent to me, that at some stage in this deliberative process, you will have to believe in a blunt contradiction, or at least violate the principle of closure.

Your intention to drink poison may not imply a belief that you will do it, but at least it implies that you regard something you do on account of that intention as an instrument for coming to drink poison. In many cases, you may intend to do things in the future but be very pessimistic about your ability to stick to the plans. But here, the example is set up in a way that makes that ability excluded — on the reductive account of intentions. You know at t0 (and perhaps also at t1, depending upon the prerequisites of the case), that whatever your intentions are at t1, at t3 you will intentionally avoid nausea, since your intentions then, like now, merely reflect your motivating reasons (which, by stipulation, at t3 are against poison). In a sufficiently strong sense of "can not" you are committed to the belief that you *can not* do it, due to your foreseen future reasons.

So, I am inclined to say that on the BD-reductive account of intentions, you can not form an intention to drink the poison. This is in line with Kavka's assumption that "you cannot act as you have no reason to act, at least when you have substantial reason not to act". Insofar as the billionaire shares these conceptual intuitions, he is pulling your leg. In other words, the question of whether the intention can be rationally formed will make sense only on a non-reductive account of intention, where intentions can be abstracted from motivating reasons and action.[3]

5.5 Intended Attempts

The worst storm in decades rages outside, and I intend to get home this evening. Or perhaps I should say, "I intend to *try* to get home," since I strongly suspect that the ferry I must use will not sail. Would that choice of words make any significant difference? Is it even improper of me to say that I intend to get home?

D.M. Armstrong holds that when someone performs an action intentionally, this entails her having tried to perform that action (1968 p.151). And John Searle's "intentions-in-action," which largely resemble intentions in the reductive sense defended here, are in more plain English referred to as "trying," according to Searle (1991 p.298)[4]. These suggestions are in line with my BD model intuitions about attempts. Any intentional action is also an attempt, and the degree of probability with which the action is supposed to realise the state of affairs it is directed at, raises no central conceptual questions.

Other philosophers regard differences in degree of success-probability from the agent's viewpoint as philosophically relevant. According to Christopher Peacocke, the possibility of intentionally doing things where chance of success is low calls for revision of the belief condition (i.e. the alleged conceptual entailment of 'I believe I will ϕ' from 'I intend to ϕ'). (1985 p.69) One of his examples of an intention in the distinct "weak" sense threatening the belief-condition is the intention to hit a croquet ball through a distant hoop. (Peacocke's revised belief condition states that the agent must believe that "I will do what I can to ϕ" — a suggestion that must seem very thin to adherents of any belief-condition, and appear problematic also for other reasons, I think.) Audi takes the more radical position that intentions for low probability outcomes are disqualified; We do not say that we intend to realise p "unless we at least believe it probable (in the sense of likely)" that we will do it (Audi 1993 p.57). Otherwise it is more proper to speak of "hope," according to Audi. (Some have held that intending implies *believing* that one will try. That view is mistaken for reasons similar to those mentioned in the former section.)[5]

A weak belief clause could be invoked in accordance with the reductive view defended here. The agent must view what he does as in some sense instrumental to something he desires. He should think of his behaviour as possibly *causing* the realisation of a certain proposition, or as that state of affairs being *instantiated* in his action. Note the difference between belief conditions requiring that an agent must believe that he *will* ϕ, and thereby contribute to p, and this condition which merely says that he must believe that *if* he ϕ-s, this will contribute to ϕ. Can this reductive account do justice to our ways of talking about risky vs. safe intentions without essential revision?

To begin with, I do not believe that we normally say about our fellow croquet-player that he intends to try to get the ball through the hoop. "He intends to do it," or "he tries to do it" is more natural. Another simple observation supporting the conclusion that the distinction is non-substantial is that when we apply it in daily life, it is drawn quite arbitrarily. We can hardly say that I intend to get home provided that I am confident in success to more than fifty percent, otherwise I am merely hoping. There is not even a rough estimation of when the probability is high enough to distinguish 'try' from 'do', as established in our linguistic practices.

Broadly speaking, we diminish people's actions to attempts when they fail to get what they have started to reach out for. Should they unexpectedly succeed in fulfilling very demanding intentions, we may sometimes talk about successful attempts. As Davidson says, talk about intentions to try are seen "as more accurate than the bald statement of intention when the outcome is sufficiently in doubt." (1978 p.92). It seems to me that this is all the distinction needs to indicate. 'I will try' simply adds the information that I regard my chances of succeeding as relatively low (as compared to 'I will do'). The important thing is still that my actualised tendency towards a certain result is triggered and sustained by the right kind of intentional states.

A somewhat more intricate use of ‘I intend to try’ or ‘I will try’ is also admitted within the reductionist account. The notion of intention proposed here admits the possibility of intentionally affecting one’s own future intentions — by directing one’s attention away from tempting alternatives, concentrating on future rewards that are dependent upon sticking to the plan etc. In those cases, my ongoing intentional action affects future intentions, and I may be uncertain, now, about my own ability to govern my motivation.

The epistemic problem might become more difficult when the agent has very low confidence in success. In such situations, it is simply more difficult to figure out what he is doing. Suppose you know what I believe about my chances to get home tonight – they are close to zero. When you see me on my way to the harbour, it may therefore be difficult for you to figure out my intentions. (An analogous difficulty of understanding what I am doing will arise when you have independent reason to believe that a certain desire of mine is extremely weak, and it nevertheless triggers behaviour — on walkover). These are, though, merely practical difficulties that need not enter the account of intention.

To conclude, the distinction between intentions to realise p and intentions to try to realise p is not essential, action-theoretically speaking. To paraphrase a somewhat lofty passage by Hume (on another subject), this distinction is not, primarily, a business for philosophers; it belongs to *Grammarians* to examine what entities are entitled to the denomination of ‘intending to realise p’, rather than merely ‘intending to try to realise p’.

5.6 Unintentional Actions

Some actions are not intentional. Some action-types are *typically* unintentional. Consider misinterpreting. One might intentionally pretend to misinterpret, but true misinterpretation implies failed intentions. Many would say, with Davidson, that it is fruitless to search for a concept of action that does not appeal to intention. What turns misdirected attempts into actions is that they are intentional under some description: “a man is the agent of an act if what he does can be described under an aspect that makes it intentional.” (1971 p.46) That takes care of actions, which are unintentional due to mistaken assumptions. When I misinterpret your words, it is still true that I intentionally listen to what you say, although I fail to realise the desired effect. “I am the agent if I spill the coffee meaning to spill the tea, but not if you jiggle my hand.”(1971 p.45-46) My spilling the contents of the cup is intentional, and that aspect makes my spilling coffee an action.

However, we sometimes categorise as actions reflexes, twitches, instinctive procedures, manifestations of common clumsiness and other types of behaviour, which are not intended under any description. They fall under the notion of agency as different types of unintentional actions. Under the adjective “reflex,” OED lists “reflex action (independent of the will, caused as automatic response to nerve-stimulation)” and the noun “reflex”

is explained as synonymous with "reflex action." When I unknowingly talk loud to myself, or even when I accidentally stumble and fall, you may think of me as *doing* something. In these cases, there is no appropriate intentional aspect to the behaviour. When I talk to myself, it is not the case that I intend to tell somebody something and fails. I do not intend to make myself heard or form any verbal structures at all. This behaviour of mine may pass without my noticing it, or it may even surprise me. Some would probably stick to Davidson's characterisation and say that behaviour in the latter category is not really action at all, since it is not intentional under *any* description. Reflexes etc., are then disqualified by definition. They have nothing in common with intentional actions, so our use of one label for both is misleading and usage is in need of reform.

I find it more proper, though, to say that more than one notion of agency is in use. With Davidson's narrow concept I would be forced to say that it was you who spilled coffee by jiggling my hand. Or, if your jiggling had no intentional aspect either, that no one spilled coffee — but just that coffee was spilled. There is, though, a weaker sense of "action," just as natural, in which I would say that you made *me* spill coffee when you jiggled my hand. I did something, although I did nothing intentionally.

Both notions are useful and there is no reason for rivalry. How, then, is this weaker concept of agency to be distinguished from things that merely happen to me, like when my hair grows or when the earth beneath me moves me away from the sun? We cannot appeal to the empirical fact that it is beyond my power to influence these events, which I take part in.[6] That inability is, namely, assumed to be part of the story about reflex-actions and clumsiness as well. These are acts I cannot help doing.

Like many other verbs, "grow" can be used transitively or intransitively. When used transitively, this term might refer to an activity, which is intentional under some description. A person could perform the action of growing weed. (He could believe that he is growing tomatoes. That is sufficient to see his growing weed as an action in Davidson's sense.) I may grow a beard, and so on. However, that active sense of "grow" does not *entail* the presence of any intention. A plant grows, but it is also capable of growing a new bud. We might even explicitly say about, say, a chemical, that it *acts* upon a certain substance, e.g. by making it corrode or dissolve. These everyday expressions are clearly not metaphorical antropomorphisms. They indicate, therefore, that we are capable of distinguishing a weak notion of acting, which is not intentional under any description.

When Crane exposes his view that dispositions are independent causes, he describes them as causal "agents." (1998 p.220). Although I disagree, as I have made clear, with his view that disposition-talk asserts the existence of independent directed *states*, I do not find it intuitively unnatural to talk about explanatory conditions or events in agency-terms. We point, e.g., to certain conditions as being *responsible* for their effects. The weak notion of acting fits well in with these ways of speaking. The chemical's constitution is causally responsible for the corrosion and the inner features of the plant causes its bud to grow. This way of speaking places the object in a certain

causal role, and refers to an effect for which the object is a condition. “My hair grows” expresses no such claims about a relation between cause and effect.

The notion of action we employ when we talk about reflex-actions and other types of behaviour without intentional aspects, does not only bear a metaphorical resemblance to the weak notion characterised above. When your jiggling made *me* spill coffee, my causal role for that effect is stressed — although no intentions are in play. My spilling coffee is in this case something that I do, albeit unintentionally, so it is an action, but not in a stronger sense than the one outlined. The weak causal assumptions made about me in this case differ widely from the conditional statements that would have to be true in order for me to *intentionally* spill coffee.

Davidson’s solution to cases of misfired attempts — that they are intentional under some description — appeals, although that is not made fully explicit by Davidson, to what has been labelled “the accordion effect” (by Joel Feinberg, referred by Davidson 1970 pp.52-55). This effect, alleged to be typical of agency, lets us “stretch out,” somewhat arbitrarily, someone’s action to include different effects. “In brief, once he has done one thing (move a finger), each consequence presents us with a deed; an agent causes what his actions cause.” (p.53) Once we have a “primitive,” or “simple” intentional action, like someone moving a finger, we can view his flicking the switch, illuminating the room and alerting a prowler as things he *did*. The accordion effect lets unintentional effects of my simple behaviour, like my spilling coffee, when I meant to spill tea, nevertheless be things I do. Davidson suggests that the accordion effect might be “a fairly simple linguistic test that sometimes reveals that we take an event to be an action” (p.54). It does not work in all cases of agency, but only in such cases. “The accordion effect is limited to agents.” (p. 53) Davidson denies explicitly that this effect can be assigned to inanimate objects.

I believe, however, that something similar to the accordion effect really can be applied to inanimate objects — in cases where the weak causal notion of agency is applicable. We could say that once the chemical compounds with the substance, its act can be stretched so that it makes the substance corrode, get fragile, break etc. The chemical does, we may think, all these things to e.g. a piece of metal (or we may for some reason want to squeeze the action to include only the first of these effects). Since stretching is applicable to such cases, it is applicable to human actions in the broad sense as well. (I snored, woke up my wife, scared the dog, initiated a discussion, etc.) Furthermore, this indicates, I think, that the accordion effect exists due to the weak merely causal notion of agency, and that Davidson’s way of accommodating unintentional actions therefore exploits a broader notion of agency than the one he suggests.

Although I accept the idea that actions (in the weak sense) can be delimited somewhat arbitrarily, depending upon which effects we want to include in the action, I should perhaps make clear that I do not think that a person’s intentional acting thereby is stretched in time. An intentional action goes on as long as the agent is causally involved in the tendency towards

some end, and this involvement is upheld by his motivating reasons. It is another matter that some plausible descriptions of an action cannot always be fixed before we know what the behaviour actually causes. It is only in this sense that the accordion effect is applicable to intentional actions.

Actions in the narrow Davidsonian sense, i.e. actions, that under some description are caused by beliefs and desires rationalising these actions, form a subclass of actions in the weak sense. Otherwise, the BD model makes no general allegations about the character of causes of people's actions in this weak sense, positive or negative. Thoughts might, perhaps, cause actions without being the reasons for those actions. A certain belief may make me nervous and get me to talk to myself, although what I then say is in no way rationalised by that belief. Beliefs might also trigger reflexes. If that can be possible, it would not contradict the BD model. What the BD model says is that intentional actions are actions caused by their reasons, and that desires are necessary constituents in motivating reasons.

The question of whether a certain action (in the weak sense) is intended is a matter of degree, rather than all-or-nothing. The reasons rationalising what I do (in the weak sense) may causally contribute to my action to some extent, while other inner states of mine might have a greater causal influence. Most noteworthy is the conceptual possibility that an action to a great extent is caused by beliefs or desires without conceptual connection with that action, i.e. by reasons operating as causes, without being reasons for what they cause. In that case, the action (in the weak sense) is only intentional to some low degree.

---

[1] Wlodek Rabinowicz pointed this out to me.

[2] Another victim of this type of dilemma is Parfit's psychologically transparent and selfish desert traveller who is stranded with a useless car. He cannot promise bypassing strangers a future reward in return for help, since he knows that he will have no reason to give it to them when he gets home. (1983 p.7)

[3] In (1998), Bratman suggests a "no-regret" clause as a measure of the rationality of following through prior plans, rather than revising them. The condition is supposed to underpin rational revisions in some cases where revision is not prompted by new information. It says roughly, that if you anticipate at t3, that at t4, when you think about your choice at t3, you will be glad if you stuck to the plan at t3 and sorry if you did not, then you should follow through with the plan — even when following through violates your ranking at t3. In the toxin case, the condition recommends revision, but in other cases it does not. As Jonas Josefsson plausibly argues, "glad" and "sorry" are simply used as expressions of the ranking at t4, of options at t3, here (1999 p.8). On the reductive account of intention, it seems difficult to picture the agent as anticipating at t3, that he will, at t4, rank drinking poison at t3 lower than not drinking it, while he also, at t3, (one might assume) predicts that he will, at t4, rank intending at t1 to drink poison at t3 over not intending to do so.?????

[4] "Intentions-in-acting," like all intentions on my view, are intentional states present in the action from its start and at the same time sustaining causes of the tendencies triggered. In other respects, I believe that Searle's account differs from the BD-reductive analysis. Searle recognises "prior intentions" which may precede an action and cause it by causing intentions-in-action. I am uncertain about how to understand his prior intentions. It is

possible that this concept could be analysed in line with the BD model's view of the motivating reasons that can precede intention and action. However, I do not think that this issue would affect the points I want to make about intentions in the BD model's sense. (Searle 1984 p.88, see also O'Shaughnessy 1991)

[5] Audi ascribes this view (that intending implies believing that one will try) to Stuart Hampshire. The thesis is criticised by Audi, and he claims that his counterexamples also show that it is false that intentional ϕ-ing implies having tried to ϕ. However, I think that Armstrong's idea to that effect remains unmoved by those examples, unless one takes it for granted, as Audi appears to do, that 'trying to ϕ' must involve the belief that one tries to ϕ.

[6] This is Irving Thalberg's suggestion about how to distinguish action verbs from "bodily process and reaction verbs" (1972 p.62). For criticism of this suggestion, see Persson 1981 p. 15.

6 Hume's model

It is often taken for granted that the first and perhaps most thoroughgoing proponent of the BD model is David Hume. "The Humean Theory of Motivation" is a common label for views constituted by all or many of the BD model's assumptions. This way of speaking is so well established that those opposing this interpretation of Hume thereby become forced to question whether Hume really was a Humean when it came to motivation! All in all, I believe there are strong reasons to suppose that the intuitions about motivation and practical rationality, which Hume's theory of action appeals to, fit well into the BD model.

Subtle arguments have been put forward against this standard "Humean" interpretation of Hume. These arguments show, at least, that the BD model reading of Hume requires important qualifications if it is to preserve the internal consistency of his theory of action. These qualifications are in line with the characterisation of the BD model outlined so far. The following brief exegetical digression is not relevant as an argument for any philosophy of action. It may be an argument, though, for supposing that the model outlined, *with* the qualifications in question, really catches some essential elements in people's psychological thinking.

Hume's theory of motivation and action is most straightforwardly expressed in *Treatise of Human Nature*, especially 2.3.3 and 2.3.4, and the interpreter's dilemma can be illustrated with reference to what Hume explicitly says in the *Treatise*.

6.1 Passion as Desire

The BD model treats desires as causal forces, which are necessary to produce actions and which make us disposed to act. Beliefs about means to ends are also necessary but insufficient. A desire is an intentional state contributing, alongside with some belief, to the rationalisation of the act. It has a content that fits some description of the action it makes the agent disposed for. As I made clear in section 3, the notion of 'content' carries no phenomenal implications; content can be characterised in functional terms. It is essential to the BD model that desires are not (necessarily) phenomenally occurrent, and also (a claim distinct from the first one) that they need not occur "before the mind", in the content of some belief or other intentional state.

In *Treatise* 2:3:3 Hume opposes the ancient view that reason and passion are two similar types of driving forces, struggling for control over action. He presents his famous alternative view that reason alone is incapable of producing action, while the essential role of passion is to motivate. So, are Hume's "passions" nothing but the BD model's desires?

Ingmar Persson finds "the Humean concept of a passion to correspond most closely to that of an emotion". He claims that it is a mistake to make Hume the founder of the belief-desire model. (1997 p.196) Barry Stroud regards the dispositional interpretation of "passions" as "non-Humean", since it clashes with a more phenomenal conception of passions, which he attributes to Hume, on account of Hume's general theory of mind. Hume explicitly lists passions among the impressions. Like all psychological states, they are supposed to be "perceptions before the mind". (Stroud 1977 p.166) This ought to mean that they must be phenomenally present, foregrounded, or both.

Stroud stresses, however, that the view of desires/passions as "certain kinds of causal states, or dispositions – and not particular items felt or inferred to be in the mind" is compatible with "the intuitive idea from which Hume derives his theory of action". (p.168) The intuitive idea assigned to Hume appears to be something like the BD model, then. In *David Hume*, Anthony Flew suggests an interpretation close to the BD model. He claims that Hume's famous idea about Reason's slavery under Passion is founded on a

paradoxically wide use of the term "passion". "For the word is here used to include every inclination which could conceivably constitute a motive for doing or not doing anything." (1986 p. 145-46) J. L. Mackie's *Hume's Moral Theory* (1980) and Jonathan Harrison's *Hume's Moral Epistemology* (1976) favours similar, BD model-like, interpretations.

Hume mentions at least four types of pro-attitudes as possible causes of action (in contrast to reason). "Volition" and sometimes "the will" appear to be used in senses close to the modern notion of intention, as overall terms for the whole decision process, the outcome of which is determined by cognitive as well as non-cognitive elements. "Desires" and "passions" denote non-cognitive elements in motivation. These concepts are not clearly demarcated from each other (Persson 1997 p.197).

Hume's terminology is misleadingly common. He concedes that his use of terms like "reason" and "passion" is more technical than in ordinary usage. "We speak not strictly and philosophically when we talk of the combat of reason and of passion". (2:3:3) And in the same section he emphasises that some calm passions are brought under the same heading as judgements of truth and falsehood, "by all those, who judge from the first view and appearance." It seems fair not to judge Hume's terminology from the first view and appearance either. I.e., it is clear that it need not necessarily be farfetched or anachronistic to interpret the provocative term "passion" in a less dramatic sense, in line with modern philosophical terminology.

As Michael Smith makes clear, much opposition against "the Humean theory of motivating reasons" is based upon a phenomenal conception of desires. (1994 p.125) The word "passion" is probably even more loaded (than "desire") with associations to feelings, emotions and sensations. It might therefore be thought that Hume's choice of label simply excludes the possibility that what he has in mind is some dispositional state without necessary phenomenal presence.

However, even in modern usage, "passion" is employed in both senses. According to the *Oxford English Dictionary*, there are at least two common modern ways of using "passion", which might be relevant here: "1. Strong emotion. 2. Strong enthusiasm (*for* thing, *for* doing)". To have a passion for fishing, e.g., does not necessarily mean that being emotionally aroused by fishing (the first sense). It may simply mean that one is strongly disposed to go fishing whenever there is an opportunity to do so.

The passage which most evidently commits Hume to the dispositional and non-phenomenal conception of passion is in *Treatise* 2:3:3, where he makes clear that passions need not be felt at all, and that their strength lies in their motivational power, rather than in their phenomenal intensity.

> Now it is certain, that there are certain calm desires and tendencies, which, though they be real passions, produce little emotion in the mind, and are more known by their effects than by the immediate feeling or sensation.

Hume begins by noting that the term "reason" in ordinary usage often refers to calm passions. This frequent, though less "strict and philosophical" usage, is according to Hume, explained by the fact that calm passions may be phenomenally indistinguishable from cognitive judgements:

> When any of these passions are calm, and cause no disorder in the soul, they are very readily taken for the determination of reason, and

are supposed to proceed from the same faculty, with that, which judges of truth and falsehood. Their nature and principles have been supposed the same, because their sensations are not evidently different.

Since reason "exerts itself without producing any sensible emotion", and scarcely conveys any pleasure or uneasiness, except in rare cases like "the more sublime disquisitions of philosophy", this must mean that calm passions also may exert themselves without giving rise to any sensible emotion. I.e. Hume makes clear that it is not necessary for a passion to be felt at all, in order for it to be capable of influencing action. So Hume's 'passions', as characterised in section 2:3:3 of the *Treatise*, appear to share one feature with 'desires' in the sense employed by the BD model: It is *not essential* to them that they are felt.

"Calm" and "violent" are opposites on a phenomenal scale, and the violence of a passion is clearly distinguished from the motivational force of it. Hume argues, for instance, that men as a fact "often counteract a violent passion in prosecution of their interests" (2:3:3). On the other hand, he presents no other clue to measurement of *strength*, than influence on behaviour.

It is evident passions influence the will not in proportion to their violence, or the disorder they occasion in the temper; but on the contrary, that when a passion has once become a settled principle of action, and is the predominant inclination of the soul, it commonly produces no longer any sensible agitation. /.../ We must, therefore, distinguish betwixt a calm and a weak passion; betwixt a violent and a strong one. (2:3:4)

It seems clear that Hume finds influence on action to be a distinguishing mark of passion. Like desires in the BD model's sense, passions are here characterised in terms of their function as the base of action-dispositions.

Michael Smith notes, following Stroud (1977, p.166), that Hume's presentation of the idea that the strength (as opposed to violence) of a passion is determined by its effects on behaviour seems to commit him to the view that passions are causal forces. Passions are inferred from behavioural evidence, for which they are necessary causes.

Some might, to begin with, question whether this interpretation is consistent with Hume's scepticism about causation. On the standard textbook reading of Hume, he is claiming that there is no such thing as causal force or causal necessity and that the principle of induction is nonsensical. This is e.g. Saul Kripke's view in *Wittgenstein on Rules and Private Language* (referred by G.F. Strawson 1989 p. vii. See also Beauchamp/Rosenberg 1981 p.31 about how widespread this interpretation is.) Would it not, then, be strange to assume that Hume presupposes real causal connections and relies upon the principle of induction in his theory of motivation? How could he infer internal causal forces from their behavioural effects when he denies the validity of inductive inferences? Similarly, as Ingmar Persson notes, a general negative claim about the causal power of beliefs would seem to contradict the "regularity" view of causation according to which "*a priori*, anything may produce anything" (Persson 1996 p.198)

Such objections against interpreting Hume in line with the BD model are, however, clearly based on an over-stated version of Hume's critique of causality. If Hume really regarded the idea of causal necessity as conceptually incoherent, much of his philosophical work, which draws upon psychological, sociological and anthropological

generalisations, would be disqualified by his own standards. As Stroud remarks, Hume's scientific project is all about seeking causal explanations of human behaviour. (1977 p.53). His dependence upon causal necessity is perhaps most evident in his theories of moral practices and moral judgements, where he often explicitly appeals to principles like "The cause ceases; the effect must cease also" (3:2:9). In *Dialogues Concerning Natural Religion* Hume lets his spokesman Philo agree with the other empiricist, Kleanthes, that inductive arguments must rest on the principle: *"Like effects prove like causes"* and Philo stresses further: "You cannot doubt of the principle; neither ought you to reject its consequences." (1779 section 5 p.37).

Hume characterises the idea of a real connection between cause and effect as, in his own word, "unintelligible". But, as Galen Strawson forcefully argues, Hume uses the word "unintelligible" in its most literal sense. He is not saying that causality is nonsense or a contradiction in terms, he is merely making the epistemological point that we cannot know anything about its nature. The eliminative ontological position - that there cannot be such a thing as causal power and that regularity is all there is – would be a dogmatic metaphysical claim of just the sort he abhors; it is ruled out by his own sceptical principles. (Strawson, 1989) [1]. Hume's references to causal powers in his theories of motivation and virtue are not just temporary concessions to those "natural beliefs" even the toughest sceptic cannot help taking for granted in everyday life. On the contrary, they are in line with principles he elsewhere explicitly stresses that scientific enterprise should be guided by.

The view that passions or desires are causal powers would therefore be no anomaly in Hume; it would be one more reason against ascribing the exaggerated eliminative ontological position concerning causality to him. Even though Hume states that we cannot know *a priori* that anything could not follow anything, (e.g., that we cannot know *a priori* that beliefs could not be the causal antecedents of passions or actions), he does not believe in the metaphysical hypothesis that anything really could be followed by anything. Although his main target is the metaphysical assertion that causal powers exist, over and above regularities, he would regard it as just as dogmatic to deny that such powers exist.

Ingmar Persson notes the distinction between calm/violent and weak/strong, but regards it as "a mystery how, on Humean principles, a violent passion can have the weakness of letting a calm one rule behaviour" (1996 p.198). He assigns two views to Hume, which taken together are supposed to make the distinction between strong and violent mysterious; First, that passions are secondary impressions, and therefore identical with emotions – felt passive states which are caused by other perceptions or by beliefs about them. Furthermore, that the vivacity of an impression corresponds to its causal influence. The point Persson wants to make is that Hume really never sketches the modern "Humean" theory of action. On the contrary, Hume *overlooks* the possibility of distinguishing desires in the technical sense — non-cognitive states partly defined by their functional role as causal initiators of intentional action — from emotions. (1997, p.197).

Persson's second claim about Hume's views is least evident. As far as I have seen, Hume never argues explicitly that an impression's power to cause action is proportionate to its liveliness. Nor does Persson present any quotations to that effect. In the passage adduced by Persson, Hume states that the "force and vivacity" of an impression determines its "influence on the mind" (1:3:10). Apart from the fact that the additional term "force" tends to trivialise the claim about influence, Hume never pairs vivacity with *motivational* strength. Effects on behaviour are not mentioned, just the effects on the mind.

It seems likely that the point Hume wants to make by distinguishing phenomenal calmness from motivational weakness is to repudiate precisely that idea about a link between phenomenal intensity and motivation. There is in that case nothing *internally* mysterious about letting calm passions beat violent ones within Hume's philosophy of action.

However, it cannot be denied that Hume's general theory of the mind implies that passions are somehow necessarily experienced, in that it, as Stroud and Persson

emphasise, categorises *all* acts of mind as perceptions. (Stroud 1977 p.166) On either interpretation of "passion", the dispositional as well as the phenomenal, inconsistencies will therefore be found within the *Treatise*. Hume's theory of the mind is at odds with an element in his philosophy of action. If passions are viewed phenomenally, as his general theory suggests that they should be, his description of calm passions in 2:3:3 does not hold.

There is another obstacle against ascribing the dispositional notion of passions to Hume. He explicitly denies that passions can have contents or point beyond themselves; they are always "original existencies", "complete in themselves" (2:3:3, 3:1:1). This sweeping denial is false on most acceptable readings of the term "passion", (Kenny 1963 pp.23). Unfortunately, it can hardly be disregarded as signalling Hume's views, since it is presented as an important part of his argument against admitting that passions can have anything to do with truth or falsity, other than in some derived sense. Since it is essential to the BD model that desires have content — that is a necessary feature of their rationalising function — it must be admitted that this is one of the passages which cannot be made to fit into the BD model interpretation. As the idea expressed is implausible as well, it is tempting to think that Hume would have said something less categorical about the representative function of passions, had he given the matter further thought. In order for Hume's conclusion to follow, he would not have to contend that passions do not represent or point to anything. He could admit that they are intentional states, but merely refute the possibility that they are assertive of whatever proposition they are about.

To sum up about Hume's notion of the passions: The most important reason for regarding the dispositional interpretation as the most reasonable one is to be found in *Treatise* 2:3:3 and 2:3:4. Even Persson and Stroud, although they both favour another interpretation, admit that Hume's distinction between calmness/weakness and violence/strength will make sense only if strength of passion is (re-)interpreted in action-dispositional terms (Stroud 1977, p.167). The price of choosing the dispositional interpretation is that it fits less well into Hume's general philosophy of mind.

In this choice between inconsistencies, I find it more reasonable to read Hume's philosophy of action as refining and modifying some elements of the general theory, than to interpret his motivational terminology completely in accordance with his general declarations. The action-dispositional interpretation would be more generous, and, I believe, likelier to catch what Hume had in mind. In other words, there are reasons to think that Hume *introduces* a category in *Treatise* 2:3:3 and 2:3:4, and that what he overlooks is not the dispositional notion of desire, but the fact that this notion requires modification of the general theory of mind outlined in the *Treatise*.

6.2 The Potency of Belief

The BD model says that beliefs alone are insufficient to produce intentional actions. Desires, conceived as driving forces, are necessary as well. It is a well-known fact that Hume, in a similar fashion, declares "reason" to be motivationally inert. Reason can not

give rise to actions "of the will". Nor is that faculty capable of affecting actions by "opposing" passions. The "Humean" interpretation of Hume equals these negative views. Beliefs are the manifestations of reason and Hume is read as claiming that such manifestations cannot make intentional behaviour without the aid of passions.

"Reason is the discovery of truth and falsehood" (3:1:1). But in which respect does Hume want to deny that reason could oppose passions? Is it our *capacity* to discover truth, or the *manifestations* of that capacity, i.e. our beliefs, — or is it even *truth* that cannot oppose passion? Like Páll S. Árdal, I find "good reasons for believing that Hume is not talking here of reason as a faculty of judgment." (1972 p.11) To be contrary to reason is, for Hume, to be inconsistent with some truth. Although his expressions sometimes seem to equate reason with truth, it is mostly more natural to read reason as *belief* in a true proposition. He talks, e.g., of passions "yielding to reason" when he describes motivational changes on account of new acquired beliefs about a desired object.

This also shows that he is willing to concede, and even stress, the fact that beliefs may influence and create passions. "Reason and judgement may, indeed, be the mediate cause of an action, by prompting, or by directing a passion" (3:1:1). Ideas of pleasure and pain, especially, produce "the new impressions of desire and aversion, hope and fear", and he makes explicit that the pain thought of in these cases need not be felt at all, in order for it to produce passion. Furthermore, all of Hume's "artificial" moral virtues, which are social *practices*, are partly products of beliefs, although their source lies in natural intrinsic passions and inclinations.

Such admissions do not estrange Hume from the BD model framework. It does not contradict the BD model to assume that beliefs create new desires by appealing to other, more fundamental desires. Pain is a state towards which I have a natural and intrinsic aversion. The fact that my belief about the prospect of pain may create other desires poses no problem for the BD model, as long as my belief does so instrumentally, by appealing to desires already existing.

As far as I can see, Hume's examples of how ideas produce passion all presuppose that another passion is the original source. Furthermore, he only admits that beliefs might cause actions indirectly, by causing passions. Ingmar Persson objects to the latter claim, though, that

> it should be noted that, although Hume sometimes talks as though such an interposition of passion were necessary, he elsewhere takes "the will" to be one of the impressions that are "the immediate effects of pain and pleasure" /…/ This allows that volitions, and thereby actions, may arise from ideas without the mediation of passions. (1997 p.195).

Strictly speaking, Hume says here that the will might be the immediate effect of pain and pleasure, not the immediate effect of the *idea* of these impressions. "Of all the immediate effects of pain and pleasure, there is none more remarkable than the *will"* (2:3:1). So Persson's conclusion does not follow. But assume, anyway, that Hume admits the possibility that beliefs about a specific pleasure or pain could give rise to action, without doing so via an extra passion. He might still hang on to the BD model claim, that a more fundamental or general passion must be the source of this action, and that the role of the belief is that of a tool (or a slave).

I have not seen convincing textual evidence indicating that Hume considers it possible for beliefs to cause passions intrinsically, i.e. without appealing to more fundamental passions. Ingmar Persson argues, however, that Hume, unlike modern "Humeans", really thinks of beliefs as motivationally efficacious, at least in the sense that a belief alone might

create a passion, which in turn may be the source of behaviour. He rests his case on the following assumptions:

a) Hume categorises passions as impressions.

b) Hume presents the distinction between impressions and ideas as a difference merely of *degree of vivacity*, i.e. force and liveliness.

c) The vivacity of an impression is its causally relevant aspect.

From this Persson concludes that the intrinsic difference between passions and beliefs is merely one of degree of vivacity. How could there, then, be a difference in kind when it comes to causal power, Persson asks.

Suppose Hume really claims that the two larger categories to which beliefs and passions belong — ideas and impressions — are to be characterised by a difference in degree, rather than in kind. It is nevertheless quite consistent to claim that there are other qualitative differences between the two subcategories. (Screams differ merely in degree from whispers — but screams of laughter differ in kind from stage whispers.) In other words, a), b) and c) do not force us to abandon the idea that belief and passion differ in kind.

Furthermore, as indicated above, I find it doubtful whether it can be established that Hume subscribes to c), which is ascribed to him on the basis of this quote:

> The effect, then, of belief is to raise up a simple idea to an equality with our impressions, and bestow on it a like influence on the passions. /…/ Wherever we make an idea approach impressions in force and vivacity, it will likewise imitate them in its influence on the mind. (1:3:10)

Another passage that could be adduced for the same point is this:

> I have already observed, that belief is nothing but a lively idea related to a present impression. This vivacity is a requisite circumstance to the exciting all our passions, the calm as well as the violent. (2:3:6)

Both these statements are, however, compatible with Hume denying that beliefs alone, without appealing to existing passions, might influence passions. Furthermore, Hume speaks only of influence on the passions and on the mind, so the passage does in fact not commit him to the view that equally vivacious "beliefs and impressions should have a similar effect on behaviour" (Persson 1997 p.193).

Since passions influence behaviour, it follows, admittedly, that anything affecting passions may have indirect effects on behaviour. The point is that Hume does not have to admit that the motivational effect of beliefs and impressions is *similar*, on account of what he says in 1:3:10 and 2:3:6. He might still, e.g., insist that while the impulse of passion "had it operated alone, would have been able to produce volition" (2:3:3), a belief could only produce volition via influencing other states of mind, i.e. through mediating passions. And the BD model does not have to involve any commitments about how desires are caused, by excluding the possibility that they can be excited by beliefs.

I have attempted to underpin two objections to the unconventional interpretation that Hume really regards beliefs as motivationally potent. Firstly, it is not clear that Hume would admit beliefs to cause passions without appealing to passions already existing. Secondly, even if he allowed that, nothing indicates that he regards beliefs capable of causing actions directly, without any intermediate passions. Although the passages central to Persson's interpretation state that there is a difference in *degree* between beliefs and passions, when it comes to their vivacity, and that their vivacity is causally relevant, these passages do not support the conclusion that the effects of beliefs and passions therefore must be similar. Beliefs and passions differ in more respects than by vivacity, as Hume makes clear in other passages.

I could end the discussion of Hume's view on the potency of reason here, but I will add one further comment of some length. My motive is not only that it makes the BD

model interpretation more fully covered, but that it also sheds light on an important feature of the BD model.

Hume endeavours to prove "that reason alone can never be a motive to any action *of the will"*(2:3:3 my ital.). Similarly, he stresses that reason cannot oppose passion in *directing the will* (2:3:3). "The will" is the mark of intentional action, i.e. an "internal impression" we have "when we knowingly give rise to any new motion of our body, or new perception of our mind" (2:3:1). This modifier indicates that he has a broader concept of action in mind as well — a notion of action *not* done at will. A risk of trivialising the negative claim might appear. If the will is a criterion of intentional acting, it seems superfluous to prove that reason alone, i.e. without any pro-attitude (like the will), cannot produce those acts which are intentional.

A more substantial interpretation is this: Actions, insofar as they are performed knowingly, are by definition behaviours triggered or at least accompanied by some internal state, "the will", that is separable from merely knowing what one is doing. "Intention" or "decision" might be appropriate terms for that kind of inner state. Furthermore, this kind of triggering state cannot be *directed* by reason. Only passion can direct the will, and reason cannot oppose passion in that respect. If we set aside his claim that we feel and are conscious of "the will", that could be compatible with the BD model interpretation. Intentional actions are actions directed by desires, as in the map-metaphor. Desires set goals and beliefs tell us how to reach them. The conceptual claim put forward by Hume on this reading is inherent in the BD model. It says that in the broader class of actions, intentional actions are defined as those actions, which are directed by desires — that is a necessary condition for their being intentional.

"Directed by", as well as "oppose" in Hume's two negative theses, indicates a relation that is conceptual as well as causal. When A *directs* B, A does not merely push B, but A pushes B towards some predetermined destination. When A *opposes* B in directing C, A draws C away from the pre-envisaged place towards which B otherwise would push C. Such descriptions presume the idea of an imagined state of affairs that is supposed to be realised by the action. In other words, these expressions indicate that his negative theses deny reason the weakly justificatory and rationalising power necessary for intentional actions.

In order for these negative theses to be true, it is therefore not necessary that the manifestations of reason are *causally* inert when it comes to affecting actions and passions. Hume might allow beliefs alone to cause actions (in the broad sense) and passions, and he may *have to* admit that this possibility cannot be excluded a priori, due to his epistemology of causality. Still, he could be right in assuming that the subclass of actions we regard ourselves as authors of, are such that their causes rationalise them — and that in order for these causes to do so, they cannot consist just in beliefs. Some non-cognitive goal-directed state is necessary for that rationalisation to take place.

That would be a less substantial (but apparently not uncontroversial) conceptual hypothesis, compatible with admitting that it is conceptually possible that beliefs *cause* actions (actions in the broad sense, that is). After all, that *conceptual* possibility must be entailed by his sceptical view of causality. Similarly, he could admit that beliefs affect passions causally but deny that they are capable of opposing passions in directing the will, since such contrariety requires that the opposing states both *have* a direction, i.e. that they point to the realisation of some state of affairs.

According to the BD model, the mark of intentional actions is that they are caused by their reasons, consisting of beliefs and desires. Desires are required to make behaviour intentional, not only to make behaviour. When Hume stresses that passions are required to produce intentional behaviour, and that "reason alone can never be a motive to any action of the will", it is reasonable to take him as making the same dual claim, both explanatory and justificatory.

Desires + beliefs are the kinds of causes which may rationalise the behaviour they give rise to. As Donald Davidson has made clear, it is also conceivable, within the BD model's framework, that beliefs and desires operate as causes without being reasons for the beliefs,

desires or behaviour they cause. (Davidson 1982 p. 305.). However, the BD model excludes the possibility that belief and desire produce *intentional* behaviour without being the motivating reason for it. Its being caused by states with a related content, which stands in a reason-relation to it, is what makes it intentional.

Perhaps my point needs some clarification. Hume's negative claim about the power of reason need not be interpreted causally. In order for beliefs, the manifestations of reason, to produce actions of the will, these beliefs must not only be causally related to those actions, but logically connected as well. Suppose Hume is willing to admit that beliefs might cause desires and even intentions (without being reasons for them). He might still insist that reason alone can not produce actions of the will. Beliefs do not make actions intentional simply by causing them. Neither could reason be said to "*oppose* passion in the direction of the will", simply by affecting passions arbitrarily, independently of their content. What Hume needs to deny is just that manifestations of reason somehow could contradict the passions needed for actions "of the will". In other words, Hume could concede the point Davidson makes, and which Persson assigns to him; that beliefs (alone) could affect actions and passions causally without standing in any reason-relation to them. Such a concession would not be at odds with the BD model.

I do not know to which extent the average Hume-student presupposes what Persson calls "the Humean presupposition, HP", according to which "beliefs alone do not give rise to *passions* or the non-cognitive states needed for action". In HP, "do not give rise to" simply means "do not cause" (1997 p.190). Passions are secondary impressions or "impressions of reflexion", i.e. states caused by beliefs (by definition). Insofar as the conventional reader assumes belief to have no *causal* influence on passions and actions whatsoever in Hume's theory, he has probably misunderstood what Hume says.

Although passions are necessary as causal elements in motivation according to the BD model version of Hume, their function is not merely causal. And my point is that Hume's negative thesis, just like the positive, therefore should be read in a sense, which is not purely causal. Such a reading would be in accordance with the BD model's view that beliefs alone do not give rise to passions or intentional actions. However, " do not give rise to" is not, then, understood in HP's purely causal sense. This thicker interpretation is also supported by the fact that while Hume openly insists that reason cannot produce intentional action, "action of the will", his explicit denial concerning the influence of reason on *passions* is restricted to the "opposition" between reason and passion. In that way, he avoids the ambiguity between a purely causal and a causal + justificatory sense of "produce". "Oppose" suggests a logical relation.

Ingmar Persson acknowledges this not-purely-causal reading of the two denials about the power of reason. (1997 pp.198) However, he thinks that this interpretation will make it even more evident that Hume never had the BD model in mind. The BD model interpretation of Hume must say, Persson seems to mean, that Hume held both, that beliefs alone cannot *cause* desires and that beliefs alone cannot *cause* actions. But on Persson's own initial characterisation of the "Humean theory of motivation", which he explicitly identifies with "the belief-desire model of the explanation of action" (p.189), the theory claims that agency must be explained in terms of reasons for acting, and that reasons for acting must comprise desires besides beliefs. In that light, I think it would be reasonable to make a similar assumption about the negative thesis. If the positive theory requires that actions are caused by their reasons, which are beliefs and desires standing in a reason-relation to the action, the negative thesis need only deny that beliefs alone cannot be *reasons* for or against passions or actions. That would suffice to block the possibility that they alone could produce "actions of the will".

6.3 Hume's Moral Internalism

Thus upon the whole, it is impossible, that the distinction betwixt moral good and evil, can be made by reason; since that distinction has an influence upon our actions, of which reason alone is incapable. (*Treatise* 3:1:1)

Hume's main argument against moral rationalism is fairly unambiguously expounded in *Treatise* 3:1:1. It says that "morals", "the sense of virtue" or "the distinction betwixt good and evil", i.e. those acts of mind, in which moral opinions manifest themselves, are intrinsically and necessarily action-guiding, while true or false beliefs — the manifestations of reason — are "utterly impotent in this particular".

The proper interpretation of Hume's "sentimentalism" about value is a subject of even lesser consensus than the question of how to understand his theory of action. This is not surprising, since Hume rarely addresses the semantic matters, which are crucial to modern theory of value. His aim is to explain, scientifically, how moral practices are generated and why we have come to regard these practices as worthy of admiration. The meaning of evaluative words is not in focus. Now and then he seems to air views about the meaning of moral language, but these passages lend themselves, unfortunately, to widely different interpretations.

He seems to subscribe to autobiographical subjectivism when claiming "that when you pronounce any action or character to be vicious, you mean nothing, but that from the constitution of your nature you have a feeling or sentiment of blame from the contemplation of it" (3:1:1). As Jonathan Harrison and J.L. Mackie notes, other text passages describe some form of dispositional descriptivism, in which vice and virtue are identified with the capacity of the contemplated object to call forth sentiments in the spectator (Harrison 1976 p.114, Mackie 1980 ch.V). Several places support the standard textbook interpretation, according to which Hume is a proponent of emotivism. (See e..g. Thomas 1993 sec.15.1.3 or Harman/Thomson 1996 p. 97).

The uneasiness and satisfaction are not only inseparable from vice and virtue, but constitute their very nature and essence. (2.1.7)

To have the sense of virtue, is nothing but to *feel* a satisfaction of a particular kind from the contemplation of a character. The very *feeling* constitutes our praise or admiration. (3.1.2)

Hume's presupposition about the intrinsic action-guidingness of moral distinctions would fit even better with prescriptivism. J.L. Mackie argues, on the other hand, that Hume's argumentation would be most consistent on an interpretation in line with an objectification theory of the kind advocated by Mackie himself.[2]

I do not believe that there is textual evidence, sufficient to pin down Hume in any of the value-semantical categories above. Like Mackie, I believe that it may "be impossible to find *the correct* interpretation of what Hume says" on this matter (1980 p.52).

It is nevertheless evident that Hume's *ontology* of value is anti-realistic. (Some interpreters have seen him as an adherent of intuitionism or cognitivism of the "moral sense" kind, which comes with a value-realism. It is fairly easy, I believe, to disprove that reading of Hume.) It is also undeniable that a strong *internalism* about value is made explicit in many passages. Internalism about values is crucial to his anti-rationalistic point as well. On the basis of these facts, I think it is safe to say that, although Hume never explicitly appeals to the BD model, his famous argument, in *Treatise* 3:1:1, rests firmly upon very similar intuitions about motivation. Only a notion of passion as desire in the action-dispositional sense would permit him to conclude so decisively, from the fact that "morals excite passions, and produce or prevent actions" (3:1:1), that moral practices as well as moral opinions must stem from passions.

Hume's statements about the impotency of belief need not be interpreted causally, I suggested above. It would be compatible with what he says, as well as with the BD model, to interpret him as admitting beliefs alone to influence passions or actions as causes. What

he can not admit is that they could do so by standing in some reason-relation to those acts or passions; they must operate *merely* as causes in those cases. The content of those causally active beliefs is in that case irrelevant to the act or passion it causes.

Such an admission would also be compatible with his argument about the role of reason and passion in morality. The details of *Treatise* 3:1:1 explain why morality, i.e. moral practices as well as moral opinions, cannot originate from reason, on the assumption that morality is intrinsically action-guiding. The explanation consists in a variety of attempts to cover and refute possible ways in which opinions about good and bad could be said to correspond to some fact, *a priori* demonstrable or empirically detectable, or to have anything to do with truth or falsity in any other sense.

Although, as I have stressed, his exact views on the meaning of evaluative language are concealed, many passages indicate strongly that he thinks of moral views as incapable of being true or false, and that our "morals" are not reports of anything at all. It is therefore doubtful whether he believes in moral *judgements* in the strict sense of the term. "Morality" he claims, "is more properly felt than judged of." (3:1:2) (The very expression "moral judgement" is completely absent in the *Treatise*, as far as I can see.) I am prepared to conclude, with Mackie, that even if the most dubious arguments in 3:1:1 are discarded, Hume undoubtedly thinks that moral opinions are not a matter of truth and falsehood. Furthermore, he has a strong case for this denial, at least if the internalistic presupposition is accepted.

Hume's denial of the truth-value of moral opinions is an important premise for his negative conclusion about reasons' ability to oppose or give rise to moral views or moral behaviour. On a purely causal interpretation of this thesis, it would seem irrelevant to prove, at such length, that moral views cannot be true or false.[3] Their truth-value is a matter entirely distinct from their causal force. His point must be that beliefs, the manifestations of reason, which are capable of truth and falsity, cannot alone oppose or affect morality *in virtue of their content*, since they cannot contradict that, which lacks truth-value.

In other words, on a reading which both makes Hume's anti-rationalistic argument valid and compatible with most of what he explicitly says, he presupposes a motivational scheme much like the BD model. He would have no reason to stress the internalistic assumption about values, as a proof that morality stems from passion, unless he thought of passions in terms of dispositions to act. He would not have to show that moral opinions are incapable of being true or false, in order to dismiss the possibility that they can be produced by reason, unless he thought of that possibility in terms of beliefs affecting passions or actions by *being reasons* for or against them.

6.4 Passions as Driving Forces, not Data

Although it may be impossible to know which specific theory of evaluative meaning Hume would have favoured, I am prepared to defend ascribing him a strict internalist view of value. To regard something as good is, necessarily, to have motivation towards getting it. This presupposition plays an important role in Hume's argument for the view that morality must come from the passions. Therefore, it is one of the reasons to think that Hume uses "passion" in a sense close to "desire" in the wide dispositional sense employed by the BD model.

If evaluations express passions, and passions are revealed in action-tendencies, it seems as if people must always act in accordance with their evaluations. However, according to Hume, it is not "contrary to reason to prefer even my own acknowledged lesser good to my greater" (2:3:3). If the goodness of an evaluated object merely reflects my passions towards it, and those passions are bases of action-dispositions, then it seems impossible to choose a lesser good instead of a greater, and therefore pointless to discuss the rationality of this choice. Will not the degree of goodness I bestow upon the chosen object ultimately be determined by my readiness to get it?

The simplest way of avoiding the paradox would be to stress "*my own*" in Hume's exclamation. The model of motivation ascribed to Hume does not restrict the range of possible *objects* of desire. It concerns the structure of motivation, not its objects. So nothing in this model stops me from preferring some other people's good, or some impersonal good, even if this means that I will have to give up a great deal of *my own* good, analysed in terms of personal well-being, for instance. But that solution would not make it possible for me to prefer a lesser good to a greater, *all things considered*. An interpretation according to which that would be possible should preserve more of the provocative tone of Hume's statement.

In *Hume's Moral Theory*, Mackie claims that the possibility of preferring a lesser good to a greater implies that "greatness of the goods cannot be measured by the degree of my preference, but perhaps by the amount of pleasure they will bring."

> Since a desire is an original existence, logically distinct from the expectation of pleasure from the desired object (which, being a belief, has a representative function), it must be logically possible for desires to fail to be correlated with expected pleasures, and then reason cannot require that they should be so correlated. (Mackie 1980 p.46)

It seems clear, from Hume's views on morality, that he regards the amount of pleasure generated by an action, directly or indirectly, as relevant to its being virtuous. Suppose, for the sake of argument, that he really thinks that we measure goodness by expected degree of pleasure, as in Mackie's suggested extension of Hume's thesis. This would make abstention from the greater good logically unproblematic.

But perhaps it is difficult to characterise 'pleasure' without making references to desire or action-tendencies. R. B. Brandt suggests that "an experience is pleasant if and only if it makes its continuation more wanted" (1979, p.40). I.e. if and only if there is a desire for continued experience, and that desire is causally dependent upon the quality of the experience, then it is an experience of pleasure. Would such an analysis undermine Mackie's solution of Hume's apparent paradox?

I do not think so. Brandt's definition implies that degree of pleasantness of a certain experience is tied to degree of being wanted. But his analysis excludes neither the possibility that a potential *future* experience of pleasure is unwanted, nor the possibility to want (perhaps more urgently) *other* things besides pleasant experiences, even if these experiences are present at the moment of wanting. I.e., Brandt's concept of pleasure keeps up our logical ability to chose a lesser good, even if 'good' is understood in terms of expected pleasure.

The trouble with Mackie's suggestion is, instead, that it clashes with Hume's internalism, as well as with his insistence that moral evaluations have nothing to do with truth. Expectations about pain and pleasure are true or false and they are not intrinsically motivating. This is not to deny that forecasted pleasures and pains almost always are instrumentally potent, since beliefs about them will appeal to fears and hopes, which are natural in sentient creatures.

A third possible way of avoiding contradiction is to assume that your "acknowledged" good is separable from your actual good, without denying that the goodness you attach to an outcome is determined by your preferences. Your actual choice may reflect your evaluations, i.e. your motivating desires. Nevertheless, the desires, which are foregrounded in your deliberation, are not necessarily those, which determine your decision. The foreseen good, which you are capable of acknowledging, could be a function of your expectations along with your *views* about your desires (actual and future etc.). The actual good, brought about by your choice (if successful) is a function of the desires that motivate you. My point is not merely that you may be mistaken about the outcome, or

self-deceived about which desires you are driven by. The important thing is that desires might play two different parts within the Humean, or BD model, framework.

An example from Thomas Nagel's *The Possibility of Altruism* might be illuminating. When Nagel attacks the view that all motivation has desire as its source, he argues that the possibility of *prudence* shows that desires merely are necessary in motivation as *data.* I do not believe that he establishes that conclusion. He is clearly right, though, in supposing that Humeans or adherents of the BD model, will have to separate the expected future desires, whose satisfaction the prudential person is anxious about, from the desires out of which the prudential behaviour in question originates.

A well-informed but imprudent person (perhaps an adherent of Parfit's Present-Aim Theory –1984 part II) chooses his own acknowledged lesser desire-satisfaction, even though his choice results from his strongest present desires. It seems to be completely in Hume's spirit to claim that such a choice, insofar as it is well informed, would not be contrary to reason. A Humean norm of practical rationality would not say that it is rationally commended, permitted or prohibited. The choice is *a*rational in the sense that norms of reason are irrelevant to it. That reading of Hume appears to be Nagel's as well, since his anti-Humean argument appeals to the intuition, or prejudice, that time-biases like imprudence *must* be irrational. So, in one more sense, Hume could still regard it impossible to prefer a lesser good to a greater. It follows trivially from the BD model that one cannot but maximise those desires, which are the driving forces of one's choice. Still one might insist that it is not only possible, but also beyond rational criticism, to choose one's *acknowledged* lesser preference-satisfaction before one's greater.

I have presented three ways of reading Hume's assertion about the arationality of preferring a lesser good to a greater. In the first interpretation, Hume simply means that we are capable of refraining from what we believe to be best *for us*. We are not thereby forced to say that it is possible to choose what we regard as a lesser good instead of a greater, *all things considered*. Against the first interpretation speaks its triviality. Few people would deny that there are situations in which it could be rational for someone to give up some of her own wellbeing, future preference-satisfaction or whatever personal thing she would like for herself. This interpretation would bring Hume's famous list of arational choices to an anti-climax. According to that list is not irrational to prefer, *first*, destruction of the world before scratching of my finger, *second*, my total ruin rather than the least uneasiness of an Indian, and third and *finally*, "even my own acknowledged lesser good to my greater".

The second solution is Mackie's suggestion that degree of goodness corresponds to expected pleasure, rather than to strength of desire. That interpretation would make it unproblematic to choose a lesser personal good when a greater could be obtained. However, this reading is hard to combine with Hume's commitments about evaluations. Expectations about pleasure are not intrinsically motivating. Furthermore, they are capable of being true or false. Hume's argument about morals would therefore be severely weakened if he equated opinions about the good with judgements about future pleasure.

My third suggestion is that Hume exploits the distinction between foregrounded desires, desires as data, and backgrounded desires, desires as driving forces. The claim that it is possible, and not necessarily irrational, to prefer less preference-satisfaction does not contradict the view that peoples' strongest motivating preferences are always those revealed in action. That interpretation would maintain the rhetorical value of Hume's list of escalating arationalities. It would also be compatible with his value-theoretical idea that a person who regards something as good necessarily is motivated towards it. Since such an evaluation on this interpretation is an expression of desire, and not of belief, it would also suit Hume's exclusion of evaluations from the domain of what can be true or false.

6.5 Hume's Model — Summary:

The BD model interpretation of Hume reads "passion" as "desire" and "reason" as "belief". So when Hume says that reason is utterly impotent, this means that beliefs alone cannot motivate. He identifies passions via their complementary role as bases of tendencies to act. His claim that passion is reason-proof means that desires are beyond direct rational criticism, in virtue of their non-assertive character.

There are two main difficulties with the BD model's Hume. First, Hume's general theory of the mind excludes the possibility of psychological states, which are not "perceptions before the mind". Thus, passions should not, like BD model desires, be just causal bases of dispositions for behaviour. They ought to be phenomenally present or experienced in some sense. Either Hume could mean that they are emotions, as in Persson's and Stroud's interpretation, or that they are foregrounded as data in our deliberation – or both. The second difficulty, which Persson brings up, is that Hume openly states that beliefs, in virtue of their "force and vivacity", imitate impressions in their influence on the mind and on the passions. So, it is alleged, Hume cannot think that beliefs are motivationally incapacitated, or that they are incapable of affecting passions.

The first of these obstacles to the BD model interpretation shows that there is some inconsistency within Hume's views, even within the *Treatise*. Some internal tension will remain regardless of which of the two interpretations that comes closest to Hume's intentions. If we stick to the view implied by his theory of mind, passions cannot have the essentially action-guiding role in which they are cast by his theory of action.

Hume openly informs us, in section 2.3.3, that the experienced intensity of a passion is irrelevant to its motivational influence. Firm, dominant and motivationally efficient passions are not necessarily felt at all. Since the influence of impressions is related to their vivacity, according to what Hume says, elsewhere in the *Treatise*, his distinction between violence and strength of passions will then be clearly untenable. On the phenomenal reading of "passion", the contradiction is blunt.

It seems more generous and reasonable to assume, at least, that "the intuitive idea from which Hume derives his theory of action" comes closer to the view that passions are "certain kinds of causal states, or dispositions", (Stroud, 1977 p.168). In that case, the remaining inconsistency might merely reflect the fact that this intuitive idea occurred to Hume when he considered motivational mechanisms and that he overlooked its implications for his initial assumptions about the mind.

The second objection was that Hume analyses 'belief' in a manner, which makes these states causally efficient and resembling passions in that respect. Therefore, it is said, he must admit that they can affect passions and perhaps actions directly and non-instrumentally. To begin with, I have tried to show why I find it doubtful whether Hume's text supports this interpretation at all. Hume stresses the motivational importance of beliefs. But he never says, as far as I can find, anything which commits him to the view that beliefs *alone*, without appealing to passions already established in the agent, can influence passions or actions causally. Furthermore, his explicit statements concerning the influence of beliefs are only about their effect *on the mind*. Without added non-trivial premises this passage does not imply anything about the effects of belief on behaviour.

A more important point is this. Hume could and should admit more openly that beliefs can have direct causal effects on passions and actions. They can, alone, cause change in passions, i.e. not just by channelling more basic passions. They can causally affect actions without the aid of intermediating passions. Without inconsistency, Hume could stress this possibility, since his negative thesis about the power of reason is not merely causal. So I believe that Ingmar Persson might be correct, when he claims that Hume never denies that beliefs could *cause* passion or behaviour. The possibility Hume refutes is that beliefs by themselves could *produce* intentional actions or *oppose* passions. They cannot, on their own, be reasons for action, therefore they cannot produce intentional action. And alone they cannot oppose passions, since they can only be reasons against what might be false. The distinction between beliefs/desires operating as reasons, and beliefs/desires operating merely as causes, is compatible with the BD model, and even essential to it.

Although there is no explicit textual evidence for ascribing the foregrounded/backgrounded distinction to Hume, this last BD model-friendly addition to the interpretation would not be *ad hoc*, since there are independent reasons for thinking that this is what he had in mind. One such reason is that it would make sense of his idea that "it is not contrary to reason to prefer even my own acknowledged lesser good to my greater", without trivialising it. Generosity is in favour of this addition as well; Hume's theory of motivation would be squarely against our experiences if he held that passions are necessary in all action, not only as driving forces, but also as objects of attention.

Barry Stroud and Ingmar Persson show convincingly that it is less obvious than many of us — including Mackie and Flew — thought, that Hume's theory of action is identical with the BD model's. Nonetheless, I believe there is enough evidence to make it fair to assume that Hume exploits an idea, which is close to the BD model. Even if we cannot say whether Hume would have accepted the details of the model outlined in the first part of this book, I find no strong objections against calling that model "Humean".

---

[1] Although the regularity interpretation appears to prevail among philosophers, leading historians of philosophy support Strawson's reading. Edward Craig, John P. Wright and Donald P. Livingston all argue that Hume was not a regularity theorist about causation. (Strawson 1989 p.vii). D.W. Hamlyn's *History of Western Philosophy* does not ascribe an eliminative position to Hume, either.

[2] I believe that one of Hume's arguments in Treatise 3:1:1 provides firm evidence against Mackie's suggestion that Hume, had he reflected more closely upon the value-semantical issue, would have accepted a projectivist "error theory" about value, of the kind made famous by Mackie. Hume's objection to the possibility that vice and virtue are objective features of the contemplated object is entirely phenomenological in character. He appeals to the *experience* of evaluating an object as the ultimate proof that there is no moral matter of fact to discover. "The vice entirely escapes you, as long as you consider the object. You never can find it, till you turn your reflection into your own breast" (p.203) etc. So he claims that the phenomenological constitution of evaluations shows that value is no part of the evaluated object. In that light, it seems farfetched to bestow upon him a theory, which presupposes that we, mistakenly, *experience* value to be part of the external evaluated object.

Mackie dismisses the passage quoted here, since he regards it as an implausible description of what the "ordinary person" might mean. "To give Hume a defensible view here, we must read him as intending to say that this is what you ought to mean, because this is all that, on reflection, you could maintain." (1980 p.58) However, since Mackie admits that the "defensible" view must be "given" to Hume by supposing that he intended to say something else than what he actually says, Mackie implicitly concedes that the passage in question supports a view which is less close to Mackie's own.

[3] Ingmar Persson seems to agree that Hume's theory of motivation, presupposed in his argument against moral rationalism, appears to be more "Humean", i.e. in line with the BD model, if one assumes that Hume denies truth-value to moral judgements. But Persson is of the opinion that "there seems to be no evidence that Hume thought that moral judgements, or indeed any judgements, lack truth-value and so would not be exercises of reason". In the proper, useful but admittedly somewhat antiquated, sense, the function of *judge*ments is to state propositions and so they are by definition capable of truth and falsity, of being asserted or denied. Hume can therefore hardly be expected to question the truth-value of any judgement. It is not likely that Hume thought of moral views as judgements in this strict sense. Cf Árdal's comment on Hume's "Moral 'Judgments'" (1972 p.17)

7 Functions of Deliberation

After finishing this paragraph, I am going out to have lunch with a good friend. He is no hypocrite, and a man of firm principles. One of them says 'Do not lend people money.' His norm admits of no exceptions, and he is capable of arguing, quite convincingly, that it is justified prudentially as well as morally. Nevertheless, he is going to lend me lunch money. This is how it works: When he realises that I am unable to pay, he will say that he wants to treat me to lunch. In passing, he adds that our next lunch together then may be on me.

Consider my prediction verified, and let that example paradigmatically define '*internal deviance*'. (Leave it open whether my friend's description really is meant as a euphemism for a loan, or if he views himself as presenting a gift to me.) 'Internal deviance' is a purposively vague, wide and neutral concept, begging no questions of reason or morality. It indicates a gap, a turn or a fork somewhere in the sphere of deliberation, evaluation and intention — but no specific analysis, explanation or evaluation of the deflection is presupposed. I dub the opposite of internal deviance *internal lineality*.

"Accidie," "akrasia," "incontinence", "procrastination," "weakness of will" and the like refer to different prima facie undesirable instances of internal deviance, while "autonomous agency," "continence," "strength of mind" and similar terms can be employed to approve of different kinds of internally lineal behaviour. For some reason, philosophers have tended to ignore reasonable, good or neutral internal deviance, as well as stupid, bad or trivial internal lineality.

Why and how does someone's motivational process lose its internal lineality? Some suggestions:

a) His motivating desires just momentarily change — for reasons of instrumental belief change, or due to direct conative effects of perceptual or internal attention to features of the situation or the imagined options at hand. In Richard Brandt's terms (1979), this means that the internal deviance comes between occurrent desires and normal desires, typically "in the heat of the moment".

b) His behaviour is directly (not via desires) conditioned by his viewing the relevant facts from a perspective, understanding or "seeing as" that is untypical of him.

c) He (untypically) takes a certain fact, which he may be constantly aware of, to be a reason for acting — or alternatively, does (untypically) *not* see a specific fact as a reason.

d) He desires to have other first-order desires than he actually has.

e) He desires to be moved by other desires of his than the ones that actually move him. According to Harry Frankfurt, this means that his second order *volitions*. fail to move him.

f) The desires that actually motivate him happen to depart from his moral, prudential or other desires considered especially important in virtue of their having a certain content. David Pears would describe this as a case where he is motivated by his evaluations "in a weak sense", contrary to his evaluations "in the strong sense". Gary Watson has a similar idea.

g) The desires that actually move him are not the desires he believes that he is moved by.

h) He believes that if he were fully rational, other desires would motivate him. In that case, Michael Smith would say that his desires diverge from his *valuing*.
i) He fails to infer ϕ! from 'realisation of p is desirable' and 'ϕ-ing is (or leads to) the realisation of p.'
j) He regards it as best (full stop) to ϕ while he regards it as best given all available evidence not to ϕ. This discrepancy beteen a relativised and an absolute practical judgement (to be discussed further on) is called "incontinence" by Davidson.

Possibly with the exception of a), these varieties of internal deviance illuminate contrasts between effective motivation and different functions of *deliberation.* To possess deliberative capacities is to be able to actively view an issue from different angles, to distract attention away from unpleasant sides of an instrumentally valuable act, to form practical judgements, to draw practical conclusions or to form opinions — cognitive and conative — about one's own motivation and deliberation. None of these non-overt acts are necessary constituents in intentional acting. Our inclination to deliberate varies with individual personality and circumstances. Some of us are for the most like Parfit's cat (1984 p.ix) or what Frankfurt calls "wantons" (1971 p.11); we seldom affect our decisions by spending time on thinking about whether we *like* our desires to become effective. We do not care which of our inclinations are the strongest. Walter Mischel's extensive empirical research, referred to earlier, seems to corroborate the assumption that there are great individual and situational variations among people's predisposition for exercises of self-control and other self-imposed instruments for motivational stability or change (Mischel 1968).

The most interesting cases of internal deviance are supposed to be *motivated*. They are not just "essentially surd" (Davidson) elements that pops up in overt behaviour for no reason at all. Wishful thinking, for instance, counts as a case of internal deviance precisely because the implausible belief is not just a result of ignorance or logical incompetence, but because the agent is somehow *driven* to the acquisition of it. Twitches or other unintentional activities may occur as strange components in an otherwise intentional action, but many cases of internal deviance are revealed as pieces of intentional behaviour. That is what makes deviance philosophically relevant here. But this is also another point on which psychological research agrees with my conceptual presuppositions. With reference to motivated irrational belief formation, David Pears writes:

> Experiments have established not only that these tendencies exist but also that they are extraordinarily prevalent. Just as Freud had shown that many faults attributed to incompetence or chance are really motivated, so too these experiments have identified a further range of faults that neither belong to the province of chance nor are the results of ordinary incompetence. /…/ Of course, we may, if we like, classify them as a special kind of incompetence, but the important point is that they are not the kind of incompetence that we attribute to a person who finds a task beyond him. (1984 p.45)[1]

"Self-deception" and "fault" implies some kind of flaw, but even in this case of internal deviance, one might in principle distinguish between reasonable and unreasonable results of the deviant motivating process. To put it bluntly, in some cases it may simply be desirable — in the sense of prudent, for instance

— to believe what you want to believe, even though no clear evidence is there to contribute to your doing so.

Deliberative activity occurs at different stages in the formation of intentions. Deliberation may e.g. produce or obliterate desires, as well as contribute to the triggering or inhibition of them. The different varieties of deliberative acts are mostly psychologically intertwined. My way of grouping them together should therefore be taken as just a practical suggestion, indicating no fundamental action-theoretical demarcations. The purpose of this chapter is twofold: I want to describe some typical elements and functions of deliberation and also to make clear how they relate to (and differ from) motivating reasons in the BD model's sense.

Judgement on these activities will, however, mainly be suspended and saved for the next and concluding chapter.

7.1 Perspectives and Understandings

7.1.1 Cognitive and Conative Effects of Attention

Any efficient salesman knows that things, situations, and even arithmetic figures, can be seen from more than one viewpoint. Change of perspective means altered understanding of whatever is the object of his customer's deliberation. The salesman realises that change of perspective affects behaviour, and, most important, he is capable of manipulating the customer's perspective. Such manipulation does not necessarily involve lying or hiding relevant facts. If he is skilful, he can change your mind about his merchandise without adding any novel information or resorting to emotive language, threats or any other kind of emotional pressure; he simply guides your attention in a way which will make you more sympathetic to his goods.

Think, e.g., of someone who tries to sell you an insurance policy. A week before you get his phone call, you have received a leaflet with facts considered relevant by the insurance company. These may include brute statistics about expected length of life, causes of death and the most common injuries for people of your age and profession. Perhaps also predicted raises of fees for medical care, costs of this insurance in comparison with the price you would have to pay for similar policies advertised by rival companies, etc. Facts are fixed; the salesman's mission is merely to make you view these facts in a light advantageous to him and his employer. He will perhaps draw your attention to some of the most disastrous (remotely) possible scenarios, and prevent you from dwelling too long upon your present economic situation. Or he might stress that the "real" cost of buying their product is lower than it appears at first sight, due to present government regulations concerning income tax reductions on account of private insurance costs. There need be nothing immoral about the salesman's attempt to affect your behaviour in this way, like lying or withholding relevant facts would have been, prima facie. Nor is there necessarily anything imprudent or irrational about your change of mind if he succeeds.

Nothing stops an agent from operating with these basic tools of consumer psychology on his own motivation. My friend who was nice to lend me money for lunch, contrary to his resolutions, may have done so. It is quite possible

that he really chose to think of his paying for my lunch as a gift, and my presumed paying for our next lunch as just a conventional way of reciprocating his favour. His presenting the action in those terms to me could have been a euphemism for a loan, as in my initial interpretation, but it could been taken literally as well.

Richard Brandt and Frederic Schick both explicitly argue that representations and understandings require a radical revision of the BD model, since these phenomena, in Schick's terms, "can matter on their own" and "be independent factors, coequal to the agent's beliefs and desires" (Brandt 1979 ch.3 sec. iv, Schick 1991 p.88). Laboratory evidence and many credible imagined examples give understandings and representations an interesting explanatory role, distinct from that of beliefs and desires. Nevertheless, I will argue that representations are not on a par with beliefs and desires. Characterisations of representations can, however, play an important role in helping us realise how beliefs and desires are formed, and especially to understand why certain beliefs and desires (among the beliefs and desires held by the agent at the moment of acting) may suddenly gain strength to become causally efficacious.

Schick ascribes to Aristotle a view similar to the one I want to defend here (Schick explicitly denies this view): That beliefs, desires or intentions spring from understandings, i.e. "that understandings must somehow come first" (1991 p.60). The modifier "somehow" should be taken seriously in my view, though. Direction of attention may *underlie* and *explain* belief, desire and intention formation. However, understandings are not necessarily separate mental events, chronologically taking place prior to the motivating reasons they explain. To understand a situation in a certain way is to be engaged in a motivation-affecting process. Representations, perspectives and understandings are all about focussing attention, deliberately or for some external reason. The conceiving of a present situation or an option in a certain way is a feature of the motivation forming process, not a separate state preceding it.

When Schick (independently of Brandt) and Brandt attempts to establish the distinct independent explanatory role of 'understandings' (Schick 1991) and 'representations' (Brandt 1979) as being on a par with the role of beliefs and desires, their strategies are similar. They both rest their cases on appeal to intuitions on examples, fetched from fiction or real life, as well as from psychological laboratories. These examples concern internally deviant motivational processes which are such that they cannot, it is claimed, be adequately described in terms of a pure BD model.

Brandt's most central evidence for "the Law of Dependence of Action-Tendency on Representation" is the before mentioned series of psychological experiments designed by Walter Mischel, also discussed by Mele (1987 p.88, 1987 and 1995 e.g. pp 46-47.) A brief recapitulation: Mischel measured children's inclination to wait for a certain reward, initially ranked above another item that they can get as soon as they declare that they refrain from the greater reward. The experimenters then manipulated the children's motivation by showing them pictures of the preferred reward, placing the preferred item in front of them on the table, asking them to think of the item in more or less attractive terms and using various other means. It is not inessential to their results, I think, that the items in question were snacks and candy, like marshmallows. This is Brandt's description:

> It has been established that older children and children with a higher IQ, and probably children with a clearer perception of time-intervals (and possibly a wider time-perspective) tend to be more successful in waiting for a larger reward as compared with seizing an immediate gratification /.../ Most of their experiments involved choices between a lesser preferred food now, and a more preferred food for which the children might have to wait twenty minutes. /.../ They had no doubt that the more preferred food would be forthcoming: in some of the experiments it was on the table to be seen. (Thus, evidently, the degree of expectation is not the only thing besides valence that affects action-tendencies) (Brandt 1979 p.62)

Mele regards these experiments as showing how attention may affect the triggering of desire (and argues, also, that they prove the separability of strength and causal efficiency of desire). Brandt's conclusion is different. His initial formalisation of the BD model is $E \times V = T$, which he defends as a well supported, testable and substantial theory of motivation. (*E*: expectancy of outcome, *V*: valence — i.e. the resultant of desires and aversions — of the outcome, *T*: tendency to act.) But Mischel's experiments is evidence for upgrading the theory with the additional assumption that action-tendencies are also proportionate to degree of vividness of representation, says Brandt: $E \times V \times R = T$. The most interesting result of Mischel's test seems, on the face of it at least, to contradict Brandt's interpretation. Contrary to the experimenters' initial hypothesis, several kinds of more detailed attention to features of the preferred food apparently *weakened* the child's inclination to wait for it, while only some ways of making the representation vivid strengthened that tendency.

Brandt stresses that his evidence is inconclusive, and that "we should assert the 'law' of dependence of action-tendencies on the adequacy of representation only with a considerable degree of caution". (1979 p.64) (It is noteworthy that the supplement is crucial to Brandt's optimistic assumption about the powers of cognitive psychotherapy, which in turn is essential for saving his widely spread criterion of practical rationality from being just Humean instrumentalism about practical reason.)

Schick presents a variety of interesting cases concerning perception of probabilities, Jesuit morality and the psychology of genocide administratators, like Adolf Eichmann. They are all adduced to support his addition of a new causal factor to the belief-desire model. The most thoroughly discussed story in *Understanding Action* describes a situation of a kind which seems to be frequent according to soldiers' biographies and other war reports. (See e.g. Walzer's *Just and Unjust Wars* ch.9.) During the Spanish Civil War, Orwell fought against the Fascists. Orwell was convinced that his use of violence was for a worthy cause, and it is indicated in his biographical writings that at several times he killed, apparently without any great moral qualms, Fascist human beings. At one time, however, he got an opportunity to snipe at a Fascist who "was half-dressed and was holding up his trousers with both hands as he ran". Orwell did not "feel like shooting at him" since "a man holding up his trousers isn't a "Fascist", he is visibly a fellow-creature, similar to yourself". (Quote by Schick, 1991 p.1)

Orwell does not explicitly claim that his beliefs and desires are unchanged throughout this event. The following parallel case described by Michael Walzer would serve Schick's cause better. It contains a greater element of self-reflection. During World War I, Lieutenant Emilio Lussu, later a socialist leader and anti-fascist exile, was watching the Austrian trenches:

> A young officer appears and Lussu takes aim at him; then the Austrian lights a cigarette and Lussu pauses. "The cigarette formed an invisible link between us. No sooner did I see its smoke than I wanted a cigarette myself..." Behind perfect cover, he has time to think about his decision. He felt the war justified, "a hard necessity." He realized that he had obligations to the men under his command. "I knew it was my duty to fire." And yet he did not. (Walzer 1992 p.141)

His decision to let the man live was not merely based upon physical repulsion. Unlike the documented behaviour of other officers in similar situations, he did not solve the situation simply by ordering someone else to do what he could not force himself to do. Lussu also considered that opportunity, which was open to him.

Return to Mischel's laboratory for a moment. I believe Mele is quite right in assuming that Mischel's "studies of delay of gratification provide excellent evidence /…/ that representations of a wanted item have two important functional dimensions — an informational and a motivational one" (Mele 1992 p.164). Roughly, the various manners of affecting the children's motivation create two opposing tendencies. Focussing on consummatory "sensuous" properties of the preferred object, like the children do when it is on the table, or when they are told to think of its chewy quality, makes them feel frustrated and less inclined to wait. More neutral direction of their attention, like slide shows and suggestions that they think about the marshmallows as "fluffy clouds", seems "to remind them of what is to be gained by waiting, without frustrating them by focusing attention on consummatory qualities". (Conclusions by Mischel et. al. referred by Mele 1987 p.90.)

In other words, some directions of attention may either confirm or disconfirm the beliefs the initial ranking of alternatives was based upon. Attending to certain other features of the options can amplify or block desires directly (or almost directly, perhaps by stimulating bodily sensations of various kinds). These stories about how beliefs and desires can be affected do not contradict the BD model. Input can simply have more or less cognitive effects, and be more or less directly conative.

Everyday intentional explanation is mostly incomplete. We rarely make explicit all the expectations and wants we nevertheless would admit affect us. Our tendency to drift towards the simplest and most general explanation — to take explanatory shortcuts, to use Dretske's expression — seems defensible from a practical point of view. In some cases we need to be reminded of the complexity of our background assumptions as well as their motivational importance. The enemy's cigarette made no difference or appeal to Lieutenant Lussu's political and military commitments — the values that came to mind when he tried to explain his behaviour. This detail of the situation might be the one that have reinforced other normal human desires of his, too evident to make explicit in this case.

It is also possible that some such small change in perceiving the situation changes the agent's expectancies concerning things he normally has a fixed probabilistic view of. Brandt and Mele both assure us that the children in Mischel's laboratory knew for certain, "had no doubt" (Brandt), that the later reward is as safe as the immediate one. Freud assumed that small children regard adults as omniscient and omnipotent, and the normal thing for children is perhaps also to take it for granted that normal adults are trustworthy. But when people do strange things, a seed of doubt about their reliability may perhaps not be out of place. It *is* strange of a stranger to place a preferred

snack in front of you and tell you that you can have a less preferred snack now, but then you cannot have the candy in front of you. We cannot be sure that the children's view of the probabilities remained unaltered throughout the experiment. What they *said* about this is not a reliable indicator. One should expect that mistrust normally is something the agent keeps to himself. A possible specification of this type of explanation is suggested by Michael Bratman in "Practical Reasoning and Acceptance in a Context" (1992). His crucial distinction is drawn between subjective probability-assignments, given background assumptions taken for granted, and subjective probability-assignment to those background assumptions.

It might be noted that an early theme in Walter Mischel's own work is that personality "traits and states" are the observer's and researcher's constructs to a large extent (Mischel 1968 ch.3). This thesis is proposed with explicit reference to observations of delay experiments of the type referred by Mele and Brandt, as well as to numerous other laboratory observations (Delay-experiments with children are mentioned in 1968 ch.6 and 9.) His reason for being sceptical about traits is that people's dispositions to choose between a fixed set of alternatives appears to be much more shaky and changeable, even due to seemingly irrelevant alterations of features of the choice situation, than standard personality categories would admit.

Mischel's experiments, as referred in *Personality and Assessment* from 1968, disconfirm various hypotheses about behavioural constancy and consistency. A great variety of elements in the choice situation, e.g. such as those mentioned in the referred example, will affect the choice of an agent, even when the possible objects of choice are kept constant. Although regularities in stimulus conditions tend to support predictable behaviour, agents simply do not behave predictably when background conditions are changed. Trait labels and trait ratings are therefore often more accurate as providers of "evidence about the personal constructs, stereotypes, semantics, or subjective "reasons" of the person who makes the statement" than about the person who is being described, according to Mischel. (1968 p.72). "Global traits and states are excessively crude, gross units to encompass adequately the extraordinary complexity and subtleties of the discriminations that people constantly make." (Mischel 1968 p. 301) This does not mean that Mischel thinks of persons as "empty organisms"; he admits that they have "structural counterparts" of behavioural dispositions — p.295. The point is that ongoing cognitive reorganisation continuingly modifies behaviour, and that choices in themselves in turn leads to further cognitive changes.

As Mischel's delay tests show, attention need not be directed towards external objects or circumstances in order to affect motivation. Perspective and understanding may differ simply with respect to how alternatives are imagined. It is compatible with much of what Schick and Brandt say to assume that when an agent's representation of a given situation changes, this means that, although his beliefs and desires concerning the situation may be fixed during the process, what happens is that the subset of beliefs among them which he is 'aware of', 'focussing', or 'attending to' is actively or passively altered. Without extending the BD model of motivation, one may also admit that such channelling of attention is likely to affect his overt behaviour.

Schick's frequent use of nouns like "seeings" and "understandings", along with his insistence that understandings are causal factors, indicates a realistic claim about these entities. Although they are distinguished from

"entertainings" and also from conscious awareness, he also stresses that they are mental states. "Beliefs *cum* desires gain their force from their connection with a third causal factor" (p.84). This is the basic idea: Suppose someone "wants *h* and believes *h-only-if-k* and that he could bring k about. If he sees some option he has as a *k'*ing, he takes that option. His seeing so is part of what moves him. Without that part, he has no full reason, and incomplete reasons have no force." (p.85) Further on, Schick adds the requirement that the agent "saw his objective as a realization of h", (although that requirement appears to depict "seeing as" as a function of goal-directed states, i.e. desires.) So, dominant desires and overall expectancies are not enough to make reasons causally effective: A given outcome can be represented by "coreportive propositions" (like Orwell's "I shoot a Fascist" and "I shoot a fellow-creature"). Something must, in those cases (therefore in all cases), determine which proposition the agent will be governed by, says Schick.

In a comment on *Understanding Action*, Rolf Gottfries has pointed out that the notion of understanding risks being trivialised here: Understandings seem to be characterised as whatever is needed to explain why I am moved by certain beliefs and desires. (Gottfries 1994 p.3) Note also that, unlike "beliefs" and "desires", "understandings" "graspings" and "seeings" are rarely used in daily speech in the substantivised form construed by Schick. An eliminative ontology of understandings, according to which the understanding of a situation is a feature of the deliberative process, rather than separate internal states, would be far less counterintuitive (more in accordance with "folk psychology") than it is when it comes to beliefs and desires.

My picture of what it is to understand an alternative in a certain way would allow Orwell to regard the Fascist as a fellow human being all the time. The man's trousers could have made Orwell focus upon different details of the situation, presumably much more psychologically complex than Orwell's report reveals. Some minute element may simply have tipped the scale against shooting — by appeal to normal background desires or affecting various motivationally relevant subjective probability assignments. Such desires and probabilities need not only concern the course of action most evidently staked out before him. In two papers on Akrasia (1980a, 1980b), Amelie Rorty describes various ways in which an available alternative course of action might catch our attention: Visually, or by giving rise to "imagined intensity or excitement" or by promising pleasure, and so on. This way of focussing will, in turn, affect motivation.

Ingmar Persson presents a similar view in a discussion of weakness of will, although it is limited to *perceptual* attention. Persson contrasts propositions dispositionally stored in the agent's mind with the propositions represented in the agent's episodic thought. He points out that we have a tendency to be mentally preoccupied with what we perceive; "a bias towards the perceived". Furthermore, when someone thinks a lot about a certain event, his desires with respect to that event are likely to be affected. "This starts a process of spiral reinforcement in which this desire further amplifies the tendency to think more about the sensibly present, and what to do about it" and so on. (1992 p.193) So Persson describes a process in which our way of focussing attention causally affects behaviour. It would not be difficult to apply Rorty's or Persson's ideas to Lieutenant Lussu's dilemma, or to the children in Mischel's experiment who were affected by e.g. being shown pictures of the preferred food. The bias towards the perceived has to do with our tendency to be caught

by the perceived in our belief and desire formation. The children's behaviour shows that this bias need not necessarily strengthen previous rankings — it is not a bias in the sense that it makes us think *better* of the perceived. (It is another matter, to be discussed in section 7.2, that Orwell, Lussu and Mischel's children mistakenly but understandably might have *believed* that their desires and beliefs were unaltered throughout the situation.)

To sum up about perspectives so far: The understanding of a situation or an imagined option from a certain perspective is the process of focussing attention, perceptual attention as well as mental. That process may be triggered by properties of the decision context beyond the agent's control, but the agent's motivating reasons can also actively contribute to the process, e.g. as a means of self-control in the light of expected momentary temptations. "Mental" attention to a feature of a situation should be understood as the elevation of certain facts and values — including facts and values of one's own beliefs and desires — to the foreground of deliberation. It is to make them appear in the *content* of the beliefs and desires which may or may not move the agent. Ways of representation have two kinds of impact on motivation, as recognised by Mele: an informational and a direct motivational one. Furthermore, perspectives may give cue to formation of beliefs and desires, as well as to the eventual triggering of dispositions from these bases.

7.1.2 Viewing Things from the Right Perspective

Schick stresses that the question of how "a person *ought* to see things has no possible bearing on this," and that *justification* of understandings is a matter which, as "theorists of action, we can shrug off" (p.164). That is a mistake. Different canons for justification of beliefs and desires partly define their functional roles in specific cases, as the direction-of-fit metaphor indicates. The question of how we justify these motivational elements is not irrelevant to the descriptive project — theory of action.

When it comes to understandings, the normative problem is not that, as agents, we have to face the concluding open challenge of Schick's *Understanding Action*, and attempt to find some new solid ground for knowing whether our understandings or seeings are right. (1991 p. 164) We already apply criteria of rightness to understanding and "seeing as". Some understandings of an option or of the choice context can be more farfetched and less close to the truth than others, and some ways of representing an object are far less honourable than others. A well-known example of the latter is e.g. 'to regard the other as a means'. Almost all examples brought forward by Schick and Brandt from the human battlefield are such that we are inclined to think of the behaviour described as irrational, immoral, or at least strongly potentially loaded with such associations.

Commonsensical criteria of correctness or aptness of understanding (like those we readily apply to Eichmann, or to the children who were unable to act so as to get the more preferred food) are stated in terms of the standards we apply to belief and action. Our inability to find a convincing autonomous BD independent norm for understanding or representation gives further evidence that it is reasonable to take understandings and representations as features of the process *governing* belief, desire and intention rather than as explanatory

factors on a par with beliefs and desires. "Seeing things right" is seeing them so as to get the affected beliefs, desires and actions right.

In motivation, according to David Pears, there is a kind of "emotional salience": Salient or urgent desires may "captivate judgment /.../ in the rush of the last moment." (1984 p.175). Furthermore, as Schick says, certain (out of the possible) propositions reporting a situation can be regarded as "standing out" or "leaping out" or be especially "salient" for a person in a dilemma like Orwell's. (1991 p.81) Salient features of a perceived object make important impact on our beliefs about how the real object is constituted. Various popular prints creating optical illusions exemplify that salience also can be misleading.

What makes a feature salient is, in turn, dependent upon the particular conceptual and perceptual organisation patterns of the observer. As Leonard B. Meyer argues in his influential work on the understanding of *music*:

> The work of the Gestalt psychologists has shown beyond doubt that understanding is not a matter of perceiving single stimuli, or simple sound combinations in isolation, but is rather a matter of grouping stimuli into patterns and relating these patterns to one another. And finally, the studies of musicologists, bringing to our attention the music of other cultures, have made us increasingly aware that the particular organization developed in Western music is not universal, natural, or God-given. (1956 p.6)

The basic principle of perceptual organisation is what the Gestalt psychologists called "the law of Prägnanz" which says that organisation will always be as good as the prevailing conditions allow. 'Good' here sums up a set of laws and principles formulated by Gestalt psychologists on the basis of empirical evidence. The 'law of good continuation' is an example: "[To] the factor of good continuation in purely spatial organization there corresponds the factor of the smooth curve of motion and continuous velocity in spatio-temporal organization." (Koffka, *Principles of Gestalt Psychology* 1935, quoted by Meyer 1956 p.92).

An important insight carefully underpinned by Meyer is that even the most basic perceptual organising processes have a non-cognitive emotive side, apart from its function as a tool towards belief formation. We *strive* for order among our data and *feel* some patterns to be incomplete, discontinuous and unstable. Music plays with those kinds of preconceived and expected patterns, according to Meyer's well-argued analysis. When it does so by eliciting our utmost capacities to extract intelligible patterns, it rewards us emotionally. But in many cases it is simply frustrating to experience inability to get things straight; to organise them in accordance with some *Gestalt.* Since pattern organisation thus does not spring from perceived data alone, but requires active imposition of some organisation norm, the salient features of an object should not be expected to always guide us correctly. Things that were not salient might have affected our beliefs, if we had noticed them.

As I noted, Schick claims that the absence of criteria of correctness of understandings "can't count against our theory, a theory that makes how we see things central. Indeed, how could it have counted? Our theory of action is a descriptive theory, a theory of what people are like" (p.164). Again, in the light of that commitment to descriptivity, it is remarkable that nearly all internal deviances adduced as evidence of understandings by Schick (as well as by Brandt) concern behaviour striking us as imprudent (the children got the lesser good outcome in terms of their own preferences), irrational, immoral (Eichmann) [2], or at least heavily morally loaded (Orwell). We do certainly

judge some understandings to be unreasonable or blameworthy. The problem is not that we lack criteria of correctness of perspective, understanding and so on. It is rather that these criteria are always parasitical upon different standards for evaluating beliefs and motivations. We have no BD *independent* norm for understandings.

Schick is not worried about the lack of criteria, mainly because he thinks that we are just as lost when it comes to criteria of justification of beliefs and desires. "There is no solid ground to stand on, no external, objective basis for a judgment on beliefs and desires. /.../Why then should we be worrying here?" (p.164) However, even granted that Schick is right about the absence of universally accepted and objective standards for beliefs and desires, there is still a great difference when it comes to justification of understandings. As a matter of fact about discoursive practice, criteria of correctness for beliefs and desires *exist*, although their epistemological status is philosophically problematic, although they are subject to continuous revision and endless controversies etc. That fact should, I believe, be taken into account even in a purely descriptive attempt to catch the elements of intentional explanation. If the plausibility of representations and understandings in fact is judged via their effect on beliefs and desires, it seems reasonable to suppose that their explanatory role also must be understood *via* beliefs and desires.[3]

7.1.3 Taking as a Reason

A convinces B that if he does not change his diet, he will soon gain quite a few pounds of weight. Will that give B a reason to change his diet? Is it a reason for him even if he does not *agree* that it is a reason? ("Yes, I am fully aware of that fact, but I do not count it as a reason to refrain from popcorn and beer.") If he *does* agree that it is a reason for changing diet, is he then motivated by this reason? These are all open questions.

We might truthfully say about a person that his desires and beliefs concerning p motivate his ϕ-ing, but still find it unlikely that he thinks of p as a reason to ϕ. In that case, p figures in his *motivating* reasons to ϕ, but not in his views *of* his reasons for ϕ-ing. And vice versa, he might regard p as a reason to ϕ, but still not be motivated by his thought of p in his ϕ-ing. Philosophical opinions about what it means to have a reason in this not necessarily motivating sense range from Williams' Humean position that all claims about reasons that do not appeal to any subjective states of an agent are false (Williams 1981 p.113) to Nagel's idea that external facts can be reasons for acting in a strictly objective sense (1970).

In spite of a vast philosophical debate, the concept of a reason is still unclear and deserves much more attention than I will afford here. Motivating BD model reasons may be contrasted with "objective" (Nagel 1970), "external" (Williams 1981), "potential" (Gibbard 1990) or "real" (Persson 1981) reasons. One might also want to sort out the reasons that actually cause an agent's actions from the "normative" (Smith 1994), "justificatory" (Schueler 1995) or "grounding" (Bond 1983) reasons one thinks that she has in the circumstances. A third type of categorisation distinguishes reasons that necessarily are bound up with a specific agent, from reasons that are "agent-neutral" (Parfit 1984 sec.57) or "impersonal" (Nagel 1986). The last type of

distinction is put forward as a claim about conditions for *validity* of reasons, rather than about their actual or potential motivational role.

These distinctions should not be equalled. However, in common for many of the above mentioned concepts of reason for acting is that they, unlike the notion of a motivating reason, can be ascribed to you even when you are unmoved by them. Some fact may be impossible for anyone to detect. It may still be a potential (objective, external and real) reason for you in the sense that it would appeal to your desires and affect your behaviour if you came to know it. Other facts may be such that you know about them, but they do not bother you anyway — i.e. they do not appeal to any of your motivationally relevant states. Still, I might think that you have a (normative, justificatory and grounding) reason to be concerned with those facts; e.g. a moral or a prudential reason. (You may even think so yourself, since views of that fact's relevance to moral and prudential considerations may be present in your deliberation, without being instantiated in the content of your motivational states.)

Let me narrow my ambitions concerning the ravelled bunch of reason concepts and settle with an attempt to fit what I regard as the basic contrast into the BD scheme. Parfit has a pedagogical way of phrasing it:

> [There are] two kinds of reason: *explanatory*, and *good*. If someone acts in a certain way, we may know what his reason was. By describing this reason, we explain why this person acted as he did. But we may believe that this reason was a very bad reason. By "reason" I shall mean "good reason". On this use, we would claim that this person had *no* reason for acting as he did. (1984 p.46)

To regard something as a reason for acting is to regard it as a good reason for acting, on this characterisation. How do good reasons relate to motivating reasons?

Three points of departure: Firstly, talk about good reasons is essentially normative, while we can assign motivating beliefs and desires to a person without expressing any kind of recommendation about what he *should* take into account. Good-reason ascription is in a sense like *advice.* (This assumption forebodes a discussion about norms of reason, to follow in the next and concluding chapter. Here, the ambition is merely to make clear the role of 'taking as a reason' in deliberation.) Secondly, as G. Schueler makes clear, the issue between *internalism* and *externalism* about reasons concerns *good* reasons (Schueler 1995 ch.2). Williams' internalism about reasons, that "all external reason claims are false", would be completely empty unless taken as a claim about good reasons. Thirdly, the difference between explanatory and good reasons is not that explanatory reasons are merely causes, while good reasons are justifications. Motivating reasons justify (thinly) as well, though there is an important difference as to the kind of justification between motivating and good reasons.

Many good-reason claims are clearly meant to be understood in the internal sense: "If you want more sugar, (you have a good reason to) ask the waiter". Given your presumed preference, asking the waiter would be a good thing to do, is what this advice normally says. But note that the instrumental claim in itself can be understood either as conditional or as unconditional on the subject's preferences (as Hare made clear 1968 in "Wanting: Some Pitfalls"). This means that even the instrumental reason-claim can be ambiguous between an externalistic and an internalistic reading. When your notorious brother in

law advises you: "If you want more sugar, steal some" he may be quite aware that *you* would not adopt the means-ends justification he sincerely recommends.

Since we often hear external reason-claims, Williams' position that they are all false is initially counterintuitive. His own favoured example with the fictional character Owen Wingraves, whose relatives thought that there were strong reasons for him to enter a military career, is a good illustration. As Williams describes the case, the family is fully aware that there is nothing in Owen's set of beliefs and desires to support the choice of a military career (1981 p.113). Our sympathies are with Wingrave, and all but the most romantic militarists would probably agree that his family is wrong. But we nevertheless, I think, regard the issue as a question open to meaningful debate. Williams does not explicitly say that the family's view is nonsense, just that it must be false. To begin with, he indicates that their claim would *intuitively* appear false in his view, but they would also have to face a more theoretical challenge:

> *What* is it that a person comes to believe when he comes to believe that there is a reason for him to ϕ, if it is not the proposition, or something that entails the proposition, that if he deliberated rationally, he would be motivated to act appropriately? (1981 p.109)

One suggested answer follows the Kantian path staked out by Nagel in *The Possibility of Altruism* (1970), and states that good reasons essentially are impersonal or universal, i.e. binding for all agents in the relevantly similar conditions. Such an account may admit that many types of reasons are relative to agents, but that fact can be accommodated by allowing the agent's motivational set to enter the relevant conditions. (See e.g. Scanlon 1998 p.74) Nagel explicitly pairs universality of reasons to anti-Humeanism about motivation, while Williams appears to argue (conversely) from Humeanism to the denial of impersonal reasons. T. M. Scanlon, in his recent *What We Owe to Each Other*, discusses the relation between these views, and attempts to narrow the gap about practical reason*ing* between the Humean and the anti-Humean on this matter (1998 Appendix).

But Scanlon's view of what reasons *are* seems anti-Humean: Reasons are propositions, and to be moved by a reason is to be moved by a *belief* with this proposition as its content. To take something as a reason can *not* simply be to have a desire towards it, Scanlon stresses. 'Reason' is a primitive concept for Scanlon, and universality belongs to its formal properties. Setting aside his claim about what kind of thing a reason is, I am prepared to go along quite a bit with Scanlon's description of what we do when we take something as a reason for acting — i.e. when we come to regard a proposition p as a good reason for ϕ-ing.

Scanlon writes:

> an important source of the widespread resistance to Williams' claims [is that] his internalism seems to force on us the conclusion that our own reasons, too, are all contingent on the presence of appropriate elements in our subjective motivational sets. (1998 p.367)

Scanlon's diagnosis of the resistance may well be correct. We do not want to think of our own reasons as being *valid* relative to our motivational set. To use Scanlon's modified version of Williams own example, if the person who has no reason to beat his wife (and strong reasons not to do it) is *me*, I would strongly object to someone saying that if my motivational set happened to be different,

then I *would* have a (good) reason to beat my wife. Grant, also, that in good-reason claims, "the things that are reasons are /…/ the same kinds of things that can be the content of beliefs — propositions". The hat's having a certain colour is what I take as a reason for not buying it, not the fact that I *believe* that it has that colour. But in order for that reason to operate, I must believe in it. I must decide whether it *is* a reason, or whether it is no reason at all. "In addition, I have to take it to be a reason for the attitude in question." (Scanlon 1998 p. 56) Of course, it is also possible that my beliefs figure in the content of my good-reason claims; that I take the fact that I have certain beliefs as a (good) reason for a certain course of action.

The "attitude in question" is desire in the motivating sense, so Scanlon regards reasons as constituents *in* desires. A consideration *seeming* to be a reason is the first central element in what is usually called desire, according to Scanlon. He also takes the initial element of "directed attention" to be a *defining* characteristic of desires. But apart from these constitutive claims, I believe that this could be a truthful picture of how (good) reasons figure in our deliberation —when they do.

To take p as a reason for ϕ-ing is to adopt a norm wherein p figures. To adopt a norm concerning p is to be moved by considerations regarding p. On the BD model, just as in Scanlon's account, to be moved means having desires produced or triggered. From the third person perspective, there is nothing strange about adopting norms about what other people should do, whether there is anything in their own motivational sets to support it or not.

It is quite in line with my view of which role "seeings" play, to note that the process ending in *taking* p as a reason to ϕ often begins with just *seeing* p as a reason to ϕ. To focus attention on features of the imagined option may often be the gateway towards formation of desires and intentions.

If we admit that good-reason claims are normative, there is no reason to regard all statements about external reasons as false or improper. (Their truth-functional status need not bother us here.) That question would depend upon the plausibility of accepting (some) reason claims as universally valid, as Scanlon suggests in his conciliatory appendix on Williams' internalism about reasons. The universality of a reason claim has nothing to do with the question of its claim to *objectivity*. That is a basic distinction manifest e.g. in R M Hare's moral philosophy. Arguments for the universality of such a norm will have to appeal to other normative intuitions. Scanlon is therefore on the right track when he makes clear that the universality of reasons is not an issue that should divide internalists from externalists.(1998 p.74). This way of looking at it would, however, not vindicate the assumption that the universality of (good) reasons must be a *formal* property of them, nor that taking p as a reason can not in the end just *be* a pro-attitude. (I would like to think that value semantics could be left open here).

Some subscribers to the BD model may prefer to preserve the air of pure descriptivity "taking as a reason" appears to have in the ears of many philosophers. In that case, they must accept Williams' necessary reference to the subjective conditions of the person to whom a reason is ascribed. They can still endow someone with a reason for acting when he rejects the reason, but only insofar as his rejection can be explained by deductive flaws or other mistakes in his instrumental reasoning. However, I believe that this move towards neutral instrumentalism would fail to catch the other element in good-reason talk: The advice.

When I point out, as an instrumental piece of *advice*, that "if you want sugar, then you have a reason to ask the waiter", the consequent is, as Hare puts it, detachable (1968). I have committed myself to *recommending* (that is a normative activity) "ask the waiter" on the condition that you want sugar. But should I merely point out that asking the waiter is instrumental to your getting what you want, I am not committed to the detached imperative. In that descriptive instrumental sense, I could sincerely assent to "if you enjoy violence, you have a reason to beat your wife". That would not, then, commit me to the detached consequent "beat your wife" on the condition that you enjoy violence. My expressing a conditional recommendation of the detachable sort is what marks the difference between a normative instrumental-reason claim of the advice kind, and a purely descriptive one. The whole point of talking about (good) reasons is advisory, therefore it is normative.

To conclude this section: We talk about BD reasons in order to explain why people do what they do, but we also think that people may have good reasons for doing and desiring things they actually do not care for. You may be your own observer, and note the same about yourself — then as a form of internal deviance. You can direct your attention to a certain state of affairs p in the context of deliberation about ϕ-ing, and in that sense see p as a reason to ϕ, without believing that p *is* a reason to ϕ. (Perhaps you think of ϕ-ing as instrumental to p without believing it possible to ϕ, for instance.) You may also believe that p *is* a reason to ϕ, without *taking* it as a reason to ϕ. Various convictions about the importance of p to your ϕ-ing may be attended to in the foreground of your deliberation, although you may fail to direct your attention to the features of p that would affect your dispositions or trigger your desires. You may also think that *if* you were able to focus your attention on certain features, then you would have other desires strengthened or triggered.

Scanlon is right in depicting good-reason claims as universal. We regard their validity as conditional on circumstances, but not as relative to agents (though the agent's internal properties may count among the relevant circumstances). A purely descriptive notion of (good) reasons would have difficulties in making clear how we could ascribe a reason to someone on any other ground than his actual motivational set. The universality claim could not be preserved, then. Furthermore, such a purely descriptive account would not catch the advisory function of good-reason talk, even in merely instrumental good-reason talk. The primitive concept in 'taking something to be a (good) reason' may well be '(good)' rather than 'reason'.

I would like to think that this account of good reasons so far is neutral between expressivism and its rivals in the semantics of value — although the position does commit me to internalism about assent to norms. 'Taking as a reason' in the normative sense entails being moved. But even on the supposition that sincere assent to such a justifying consideration just expresses desire, the rest of the account of the process could be preserved.

> Just to assent to such a consideration, whether or not one acts on it — whether or not the desire comes to operate in the background — is by some accounts to give expression to a suitable sort of desire: perhaps a desire for the option, perhaps a desire for the relevant property. This latter will be a disposition, not necessarily to choose the option on offer, but with options between which you are otherwise indifferent to choose an option with the property rather than without. /…/ We are happy to admit, for present purposes, that assent to a justifying consideration may express desire in some such way". (Pettit and Smith 1990 p.567)

7.2 Practical Judgements

In connection with my earlier discussion of Davidson's non-reductionism about intentions, I noted that his use of the term *judgement* in connection with motivation is ambiguous. Before he makes clear (in response to Peacocke's complaint) that to make judgement for him is to have a disposition, the term is easily misunderstood as referring to a phenomenally occurrent mental state. But even after that, it is still not clear which *kind* of disposition he has in mind when he claims that intentions are judgements. As Annette Baier remarks, his view of reasons appears to have undergone a change since "Actions, Reasons and Causes" (1963) — she labels his later position "rationalism" (Baier 1986). It does not seem altogether improper to read his change of terminological preferences (towards a more frequent use of "judgement" in connection with intentions) as indicating a shift in view. To describe intentions as judgements is to stress the inferential character of the motivational process, rather than the causal.

Before attending to the question of how judgements figure in deliberation, let me propose a restriction about the use of "judgement", "sentence" and "proposition", just to get my terminology straight in this section (no metaphysical claims intended):

A proposition is a state of affairs. To make a judgement is to actively assert a proposition, i.e. to hold the proposition for a fact. The *making* of a judgement is an ongoing foregrounded psychological activity — unlike mere believing or coming to believe something. The indicative mood of a sentence can be employed to express judgements, but all such uses do not express judgements; the indicative grammatical mood can be employed for a variety of other purposes, like questioning, commanding, recommending etc. If a sentence is used descriptively, it asserts a proposition, and the sentence is a judgement in the linguistic sense.[4] (If expressivism about value is correct, there are no value judgements.)

On this use of the term, judgements figure in deliberation as foregrounded assertions. BD intentions are not judgements in this sense, nor are desires. I hesitate to say that beliefs are judgements; it seems more plausible to say that beliefs (as well as desires, perhaps) can give rise to judgements. One possible element in the specification of the dispositional role of (some) beliefs might even be the tendency to make judgements in this sense.

Davidson's talk of judgements as identical to intentions or actions has an honourable background in the tradition of Aristotelian practical syllogisms. Such syllogisms can be understood as a metaphorical way of representing motivation:

> But if Aristotle's account were supposed to describe actual mental processes, it would in general be quite absurd. The interest of the account is that it describes an order which is there whenever actions are done with intentions. (Anscombe 1957, p.81)

If the conclusion of a piece of practical reasoning literally is an action, as Aristotle says, then he wants to represent the actual motivational process. The logical terminology must then be taken metaphorically. In that case, the practical inference metaphor would fit well in with the BD model. Practical

syllogisms has a non-cognitive component and an instrumental component, and the logical terminology metaphorically suggests non-contingent relations between these components and the resulting action.

On the other hand, it could also seem reasonable to take his syllogisms as literally describing a common type of deliberation — foregrounded deductions using desires as data, leading to judgements concerning the options at hand. When Anscombe and Pears discusses whether there can be truth-relations between practical conclusions and actions, with reference to practical inferences, that seems to be a more plausible interpretation.

Anscombe discusses the question of whether there could be a truth-relation between an agent's value-judgement and his action in a way analogous to the relation between e.g. beliefs and perceived objects. Can an agent contradict himself in action, just as he may contradict himself in belief or in speech? With support of her interpretation of Aristotle, Anscombe claims that the cases really are analogous. (1985) As David Pears has made clear, there are several difficulties with this analogy (1984 p.157). The most evident difference is perhaps that while beliefs or sentences necessarily possess semantic properties, e.g. reference and meaning, this is true only for a small subclass of actions in general, e.g. some speech-acts.

Pears draws a distinction between questions about the possibility of contradicting oneself in action, and questions about contradictions in descriptions of one's actions. “When the agent is his own spectator, he may face both charges of self-contradiction: he acted in a self-contradictory way and later he described his own action in a self-contradictory way” (p.155). That distinction is also problematic. To say that a person commits a contradiction in describing his actions is to imply that he could not possibly have been acting in the way he claims that he acted. If it is possible to contradict oneself in action, then this must mean something else than that the description that fits this action is self-contradictory. If someone is faced with both charges, his prosecutor contradicts herself.

Pears poses an analogy between the biasing effect of salience on sense-perception and the effect of intensified desires in the last moment before acting. Affective pressure in the last moment before acting may strengthen or weaken a certain desire unproportionally. There is, however, a misleading element in the analogy. Our perceptual apparatus sorts impressions into patterns on which we base our beliefs about the perceived object. Salient features of the real object might misguide that process, as well as our subsequent beliefs. In an analogous way, attempts at introspective analysis of one's own desires may be biased by the intensity and the salience of some desires so that one tends to overrate their importance in motivation. This overrating misleads our judgement about our desires and beliefs. That kind of judgement may figure in our deliberation, and diverge from our motivating reasons. It may nevertheless have an effect on the triggering of desires.

7. 3 To Judge Best, All Things Considered

Davidson expresses the view that “there is no paradox in supposing that a person sometimes holds that all he believes and values supports a certain course of action, while at the same time those same beliefs and values cause

him to reject that course of action." (1970 p.41) How is that to be reconciled with Davidson's commitment to the thesis "that intentional action always is accompanied by an 'all-out' or unconditional judgment that the intended action is better than (or at least as good as) any other alternative believed to be available"? For Davidson, judgements are dispositions, as I have stressed. (Norms of rationality will be discussed in the concluding chapter, and I my aim here is just to characterise a certain type of internal deviance, without discussing whether it is a plausible description of akrasia.)

Davidson suggests an account of practical reasoning, analogous to Hempel's account of probabilistic reasoning. Normally we follow a principle of continence, which is analogous to that rule of inductive reasoning which bids us to make our inductive conclusions on the basis of all available evidence. The two principles could be stated:

Induction: From 'pr (r,x) and r is the total available evidence' infer 'x'

Continence: From 'pf (r, x is better than y) and r is the total available evidence' infer 'x is better than y'

According to one objection, put forward e.g. by Christopher Peacocke and Susan Hurley, the inference in the principle of continence is much closer to a logical entailment than Davidson thinks. Like Michael Smith in *The Moral Problem*, they think therefore that Davidson's analysis of evaluations as desires really commits him to deying the possibility of deviant cases. The core of the objection is that Davidson's pf operator is not analogous with the pr operator in the inductive rule.

Christopher Peacocke states that Davidsons analogy would not hold with a fair notation. While "'pf' functions simply as a formal relativization device", 'pr' "is more than such a relativization device: it incorporates the notion of probability". Davidson seems to take Peacocke as saying that Davidson is wrong about the correct principle of inductive evidence (1985 p.208). Explicitly, Peacocke does not hint at any specific account of epistemic probability as being the correct one in his article. He just claims that Davidson's pf and pr are disanalogous. But his objection does seem to presuppose that we do not settle with principles of induction which counsel "simple acceptance" (as Davidson wants us to).

The idea seems to be that pr should be read as thicker than pf. In order to get from the evidence to the conclusion, we have two steps to pass in the inductive case — roughly: 1. from 'r, x is probable', infer 'x is probable'. 2. from 'x is probable, infer 'x'. In the practical case, the inference is closer to an entailment, Peacocke says, since it only incorporates relativisation. Since it is unclear, at least in "How is Weakness of the Will Possible?", whether Davidson has intended pf or pr to incorporate further notions (besides the one of relativisation) I find it difficult to judge the adequacy of Peacocke's objection. However, three strategies could be used to meet the argument.

To begin with, even if the practical inference is *closer* to an entailment, due to its lack of a counterpart to the epistemic probability reservation, it is still not an entailment. Although it may be questioned whether the analogy holds when it comes to the normative force of the two principles, I find it clear that the relativised judgement in the practical case is just as conceptually distinct from the absolute judgement as in the theoretical case.

The two remaining countermoves both spell out the analogy. We can choose either a pure relativisation operator, like pf, or an operator incorporating further specifications of the nature of the relation, like pr, and then model the analogy on this concept. We need not use a relativisation device incorporating probability on one side of the analogy and a pure relativisation operator on the other.

The simplest way of saving the analogy against Peacocke's objection would be to interpret both pf and pr as pure relativization devices. Davidson's use of pr and pf for these notions might then be seen just as ways of indicating the inconclusive character of the whole operation. An interpretation along those lines is suggested by Davidson's words: "The 'probably' is rather part of the advice to the rational man: if he accepts the premises, he should give some degree of credence to the hypothesis /.../ As such, it does not belong in the 'conclusion'; it is an aspect of the inference." This means that the principle of induction lets us jump from evidence to simple acceptance (of 'x'), just as the principle of continence bridges evidence with simple acceptance (of 'x is better than y').

Modelled on the thicker notion, we would have to assume that there is a practical analogue with the probability that Peacocke assigns to the conclusion in the first step of the inductive inference. "relative to r, x is prima facie better than y" would perhaps be a plausible way of expressing that. The "prima facie" would then be an evaluative reservation about the judgement, when seen as conditional on the evidence.

It is perhaps also possible to spell out the analogy in an even more elaborate way. When it comes to induction, it may be important to distinguish the question of how "sure" we are about a certain outcome from the question of the probability we assign to the outcome (although more complex probabilistic expectations may account for the "sureness"-factor). To borrow an example from Nils-Eric Sahlin: Suppose you come to the conclusion that given all your evidence, there is a 30% risk of there being a transit strike in Verona next month. You have a bet on that possibility, worth 100 dollars. But you would probably trade that bet for a gamble in which you win 100 dollars if you draw a black ball from an urn containing 70 white balls and 30 black balls. Although the probability assignments are the best you can do in both cases, you are simply more confident in the second case (1988 p.111).

The analogy may hold in this respect as well. The degree of value I assign to an alternative can be distinguished both from the question of how probable it is that the alternative has this value, and from the question of how certain I am when it comes to this probability assignment.

Again, since the analogy holds on the simple relativisation interpretation, it is not necessary to complicate matters in this way in order to make Davidson's point. And even if we accept Peacocke's first interpretation on which the principles are disanalogous, this would not force us to regard the bridge from ''r, x is better than y' and 'r is the total available evidence'' to 'x is better than y' as an entailment.

The idea of surveying one's own beliefs and desires, and then attempt to form an opinion about which course of action that would be appropriate in relation to that body of evidence is a characteristic way of deliberating. The resulting judgement may not be identical with the intention, not even if the all-things-considered statement has been affecting the motivating process. There is nothing in that picture to contradict the BD model. It is another matter, to be

discussed in the concluding chapter, whether Davidson's principle of continence also is a reasonable norm.

7.4 Self-Ascription and Folk Functionalism about Desires

You can always recognise philosophers: They are the people who invariably know what our grandmothers think *without ever asking them* (Gurd and Marshall 1993 p.47)

The two neuropsychologists quoted here apparently have a point. As they would have presumed, I started off my project by claiming (with Hare) that if you want to know what people mean, do not just ask them.[5] The empirical part of my assertions about the commitments inherent in the BD model says that the model is a widely tested and approved method for explaining observations and predicting behaviour. Therefore, the philosophy of action inherent in that conceptual scheme is continuously being corroborated. Since the model is far older than a generation, chances are great that today's grannies employ it as well. My claim about the BD commitments inherent in conventional psychological thinking is mainly based on the traditional philosophical method of appealing to linguistic practices, not on empirical surveys.

Few philosophers appear to have taken interest in more experimental empirical methods when examining the question of how people actually psychologise — of what the philosophy of action of the philosophically untutored "folk" really looks like when you ask or test them on the subject.

As I have noted, most specific sequences of behaviour can be given different descriptions. The experimenter's characterisation of the situation may therefore be theory-laden in a way that begs the action-philosophical question. Furthermore, as we have seen from Brandt's, Mele's and Mischel's different ways of using Mischel's laboratory results, even a fixed and neutrally formulated report in statistical terms admits interpretation in accordance with any of the preferred structuring schemes. This means that conceptual analysis will have an important part to play in ranking the plausibility of those descriptions anyway.

Having made those reservations, I will suggest that some experimental evidence nevertheless strengthens some of my BD model assumptions.

In an interesting article from 1993, Alison Gopnik presents a set of conclusions drawn from quite an amount of empirical research on children's conceptions of what goes on in the mind of other people, as well as in their own. The central experiments referred were designed and executed by Gopnik and her colleagues, but several other studies are also brought forward. What her research underpins is a non-behaviourist functionalist "theory-theory" about how we come to understand intentional states. Her view is a functionalist theory-theory in that it assumes that we learn about intentional states by observing patterns of behaviour and constructing theories about the causes of those behavioural patterns. So we do not model other people's intentional states on our own, as directly experienced in our minds; the model we construct (at some time around the age between 3 and 4) is gradually and simultaneously applied to ourselves as well as to others. Gopnik's corroborated hypothesis is also *non-behaviourist* in the sense that it recognises underlying

psychological states and regards the experiences they lead to in ourselves and others as one sort of observed data we as children construct our theory from (1993 p.12). It should be stressed that her conclusions concern knowledge of intentional states — states with a propositional content — not e.g. the existence of qualia or the nature of introspection.

In an article from 1977, Richard Nisbett and Timothy Wilson review extensive experimental evidence suggesting that people "have little or no direct introspective access to higher order cognitive processes" —p.231. Subjects are sometimes unaware of the stimulus that made them respond, unaware that they respond, and unaware that the stimulus affects their response. Reports on our own cognitive processes are therefore based on "a priori, implicit causal theories, or judgments about the extent to which a particular stimulus is a plausible cause of a given response" (p.231). The subject's report of his cognitive process tends to be accurate "when influential stimuli are salient" and also are plausible causes of the responses they produce — in such cases, the psychological reality comes to fit the theory. Gopnik's conclusions would be entailed by theirs, as far as I can see, but her own explicit claims are less far-reaching.

One of Gopnik's critics is Alvin Goldman, whose theory about how we come to master mentalistic words is introspectionist in character and says that such words directly refer to distinct qualitative aspects of inner experience. I have no reason to go into the details of his criticism here, nor of his predictably different interpretation of Gopnik's result that the children simultaneously develop capacities to give adequate reports on their own and other people's mental states.

An assumption common to both participants in this debate (as well as to several other commentators — see Commentaries on Gopnik/Goldman in *Behavioral and Brain Sciences* 1993) is, however, of relevance to a theme in my exposition of the BD model. Like Nisbett & Wilson, Gopnik and Goldman both couple functionalism about psychological states to the denial of immediate accessibility of these states. That is not a conceptual necessity; on a realiser version of functionalism, we could imagine contingent but straight correspondence between certain phenomenally available qualities of the realiser state and its functional role in relation to inputs and outputs. Our reason for not taking that possibility seriously is, one might say, introspective. It simply does not square with the phenomenology of action. Behavioural patterns do not fit any definite phenomenological counterparts. Due to this brute observation, we choose between good introspective access and functionalism about the nature of psychological states.

Functional roles may involve the production of knowledge or self-reports about the state. That can be part of the state's specific function. Although some psychological concepts ('embarrassment', for instance) perhaps may be specified in that way, the state in focus here — desire — does not entail such self-knowledge, as I have argued in chapter 4.

Goldman presents functionalism as the orthodoxy concerning intentional states among theorists of the mind: "Even friends of qualia (e.g. Block 1990) feel committed to functionalism when it comes to desire, belief and so forth" (1993 p.23). Considering the variety of comments to Gopnik's conclusions, it is hard to say where the "orthodox" label is most appropriate. But the view of desires and beliefs as distinct in virtue of their interlocking behavioural

functions is, admittedly, widespread. Stalnaker's dispositional scheme is often quoted with approval:

> To desire that P is to be disposed to act in ways that would tend to bring it about that P in a world in which one's beliefs, whatever they are, were true. To believe that P is to be disposed to act in ways that would tend to satisfy one's desires, whatever they are, in a world in which P (together with one's other beliefs) were true. (1984 p.15)

If the commitment to functionalism is further narrowed, so that we confine it to *desires*, I believe that Goldman's generalisation would be even more reasonable. (That would leave open a possibility to characterise belief in a way that widens Stalnaker's circle.) If the BD model correctly catches people's psychological thinking, that general commitment should be expected. As I have stressed, the essence of the BD model is its dispositional but realistic picture of desires, which is a form of functionalism about these states. That view of desires leaves a functional role for beliefs, but no further claims about the nature of belief must come with the BD model.

Folk functionalism is the psychological assumption that people in general form opinions about psychological states from observations of their functions, via a theory or model that enables them to see patterns. As I noted, Gopnik explicitly restricts her claims about folk functionalism to states with a propositional content. She groups these states into three categories (with reference to Searle's 'directions of fit'): Beliefs and desires of various kinds, and, in the terminology of her research subjects, "silly states" — i.e. images, dreams, pretences etc. For my purposes, it is sufficient that her results support functionalism about *desires*.

With the aid of results from her own research and that of others, Gopnik shows that a representational model of the mind — where the understanding of possible misrepresentation plays an important role — replaces a direct causal model (not unlike the one ascribed to Stoutland in ch.3) between the age of 3 and 4. To simplify somewhat, when the direct model is applied to other persons, it implies that all beliefs are shared and true, since beliefs simply "transfers" or "copies" what is the case.

This feature of simply mirroring the present interaction with the environment goes for desires as well. Just as belief simply is what is the case around me, desire simply is what I do in this situation. Before the theory change, children have difficulties e.g. in understanding that "objects are desired under a description, and that desires may vary as a result of that description" (Gopnik 1993 p.6).

Gopnik's own experiments are designed to measure whether these noted differences, between the three-year-old psychologist and the four-year-old one, can also be detected in their self-reports. She finds that similar differences appear, at the same stages of age. The *three-year-old* is unwilling to ascribe false or different beliefs to others. Similarly the *three-year-old* will not ascribe a false belief to herself at an immediately earlier time. When she discovers that the box contained pencils, not candies, as she had been led to think earlier, she reports (when asked) that she believed that it contained pencils earlier as well.

Desires: The *three-year-old* who has difficulties in regarding other people's rankings as different from his own will also have difficulties in appreciating desire change in himself. When their desires were satiated — hungry children were fed crackers at snack time, for instance — "a sizable minority of 3-year-olds (30%-40%) reported that they had been in their final state all along."

(Gopnik 1993 p.8) The "absolute levels of performance" in reporting immediately past desires "were strikingly similar" in this task, compared with a similar experiment concerning reports on desires of others. (p.8) If simple embarrassment could have accounted for the child's refusal to admit her earlier false belief — a factor intended to be eliminated by the design of some of the experiments though — this could not apply to the desire tasks, as Gopnik makes clear. There is nothing embarrassing about admitting that you were hungry before you had the snack. Furthermore, according to Gopnik, in other tasks, "the children are quite willing to admit their ignorance" (p.9).

Intentions: The child who has difficulties in admitting different *intentions* in other people will also have difficulties in recognising immediately previous intentions in herself:

> Children were given a red crayon and asked to draw a ball; halfway through the experimenter said, "Why, that drawing looks like this big red apple, could you make it a big red apple?" Children complied. Then we asked the children to report their past intention; 50% of the 3-year-old reported that they had originally intended to draw the apple. (Gopnik 1993 p.8).

Several critics point out that the children's *current* reports are accurate, and claim that this supports introspectionism. (e.g. Harris 1993 p.48) That criticism misses the target. The direct causal theory Gopnik ascribes to the children links reports on psychological states directly to the present interaction with the environment. Their contemporary reports about beliefs and desires should therefore be expected to simply fit the situation and what they do in it. Since they are capable of remembering many other things at the age of 3, they should be able to remember an immediately preceding state, if it had been distinctly present in their mind.

What seems to happen between 3 an 4 is that they become able to organise their observations of others and themselves (including, as Gopnik is willing to admit, their experiences of some psychological states) and apply a theory that assigns intentional states to agents — states that must underlie and explain the patterns they can extract.

The important lesson is not that the representational model of desire is more plausible than e.g. a direct causal model. What is interesting here is, firstly, that they both *are* models, and secondly, that the early development of these models show that we do not have any direct knowledge of our desires or intentions. Some types of mistakes about one's own psychological states indicate that they are not directly accessible. To quote Mischel again:

> It has been widely assumed that poor correspondence between self-reports and actual non-test behavior, or poor correspondence between the subject's self-reports and ratings by observers, indicate that persons are either unable or unwilling to describe their behavior accurately. /…/ Equally possible, poor correspondence between self-report and nontest behavior may reflect the fact that most self-reports elicit the subject's global interpretations about his typical behavior and his personal constructions about his psychological attributes or traits. (1968 p.69)

---

[1] Pears refers to Nisbett and Ross, *Human Inference: Strategies and Shortcomings of Social Judgment*, and articles by Tversky and Kahnemann (e.g. in *Psychological Review*, 1973) from the early 1970s as pioneering in giving a firm scientific basis of the hypothesis that irrational belief formation often has a rationalising cause.

[2] One possible explanation of the behaviour of WWII genocide administrators presupposes that they mostly had beliefs and desires of a very ordinary kind, and that their triggered actions therefore must depend on an inexplicably distorted perspective or focus of attention towards the ongoing genocide they were administrating. Hanna Arendt, to whom Schick refers, defends such a view of Eichmann, and Gitta Sereny defends a similar view in *Albert Speer: His Battle with Truth* (1995). That view is a matter of controversy among historians, though. Daniel J Goldhagen in *Hitler's Willing Executioners. Ordinary Germans and the Holocaust* (1995) defends the opposite view: That ideological commitments and deeply rooted prejudices and aversions towards their victims was a standing condition, established long before WW II, which Hitler merely had to put in motion. Both explanations would be compatible with my picture of the function of understandings and representations.

[3] There is a possible response to my claim that the lack of BD independent criteria for understandings indicates that understandings form and trigger beliefs and desires rather than co-operate with them. That is to meet the challenge presented on the last page of Schick's book and formulate independent criteria of correctness for understandings. Let me just indicate, by mentioning some objections to imaginable independent standards of adequacy, why I am just as pessimistic about the success of such an attempt as Schick appears to be. *Coherence:* Even if we, unlike Schick, hold on to the principle of closure, coherence can never be a sufficient criterion of adequacy of representation. The internal consistency among Eichmann's representations of the results of his actions could have been greater than ours — he might have been able to consistently avoid attending to distracting features of options at hand. Nevertheless we would not doubt that our representation of the suffering is more reasonable. *Completeness:* As David Velleman argues against Brandt, one might question if the question of whether a given option has been exposed to "all facts from all the angles and in all lights" is empirically determinable, and also whether it is determinate in principle. (1988 p.369). Issues may oblige us to invent new representational possibilities, and there is simply no way of telling whether an issue is considered from all possible angles. When we say that someone has considered an issue from all angles, we mean that he has considered the issue from all angles that are *illuminating* for this specific issue. The question of what makes a representation illuminating is the one we began with. *Vividness:* Brandt speaks of adequacy of representation in terms of completeness and vividness. Vividness is sometimes characterised (by Brandt) in terms of richness in *detail* – but that interpretation would make it difficult to distinguish adequacy of representation from adequacy of belief. Another possible interpretation of 'vividness' might be put in terms of phenomenal *intensity*. However, the insight expressed in Schick's and Brandt's points about *salience* is precisely that the intensity might be misguiding. *Psychological stability:* Brandt's rationality test hangs on contingent facts. Attitudes are rational if they would survive cognitive psychotherapy. (Repeated vivid representation of causally relevant facts at appropriate moments.) Perhaps the rationality of representations in themselves might be (hypothetically) tested in this way: 'A certain representation of a fixed body of facts is rational if it would survive repeated confrontation with other possible representations of those facts.' It does not seem unreasonable to think that if a new perspective on a given fact makes me unable to go back to my former way of seeing that fact, this indicates that there was something wrong with the former perspective. But can this criterion be applied independently of our standards for beliefs and desires?

Suppose I am a libertarian of a non-sophisticated kind and that I am trying to bring up my children in line with this ideology. My son is, by nature, soft hearted and full of empathy. In many situations, he is prepared, and even finds it morally required, to give up some his own goods for the benefit of someone who is less fortunate. We both agree upon the relevant facts: These other persons are suffering, we would suffer as much as they do if we were in their shoes, and we are able to help them without substantial sacrifices. By consistently drawing my son's attention to the trivial fact that their suffering is theirs, that he does not feel their pain, and that persons in that sense are *separate*, I cure him, eventually, of his altruism. His initial representation of the situation has not, then, survived repeated representation from another point of view. Most of us would hesitate to take that as a sign of the incorrectness of his old way of representing the situation. Our view of its correctness will depend upon our (political, moral, scientific etc.) opinions concerning the beliefs and motivations it gives rise to.

[4] As Lars Fröström has made clear, there is an important distinction to be drawn between the weak sense of 'assertion', in which the descriptive sentence asserts a proposition, and the strong sense of 'assertion' that figures in the act of asserting. (1983 ch.2)

[5] Gurd's and Marshall's own amusing description of empirical research on a small sample of grannies illustrates, apparently contrary to the authors' intentions, the danger of believing that you get an unbiased picture of people's metaphysical outlooks by asking them. Grannies are "all unreconstructed Cartesians", claim Gurd and Marshall. Granny says, namely, things like "Of course it hurts if you put the hand in the fire" (thereby proposing interactionism according to the authors). Grannies are also said to believe in qualia, since they can be quoted as saying: "If I say it looks green to me, young man, then it looks green to me". (1993 p.47) Furthermore, Gurd and Marshall agree with Goldman's conclusion (being a philosopher, Goldman draws it from a thought experiment grandmother) that such statements also indicate that Granny is *right* about philosophy of mind.

8 Three Norms of Practical Reason Rejected

8.1 The Value of Deliberation

The discussion of the preceding chapter showed that the BD model allows motivating reasons to diverge from the reasons that are present in an agent's deliberation. Her practical judgement, understood as a foregrounded assertion, might recommend another course of action than the one she is about to enter. That sort of judgement might reflect justifying considerations that normally *would* move her, and also considerations that, at the moment of acting, she *desires* to be moved by. One of the important functions of forming a practical judgement of that sort might be precisely to affect one's own dispositions.

My ambition in ch.7 was to characterise some forms of deliberation and examine their role in relation to motivating BD reasons. As far as possible I tried to avoid presumptions about the negative value of that possible gap between motivation and deliberation. That is not the traditional philosophical approach. Philosophers have tended to concentrate on various forms of prima facie irrational forms of internal deviance, or on prima facie reasonable forms of internal lineality.

The normative issue of which elements you should be moved by in your deliberation can be framed in other ways than in terms of normative reasons. It is also central to at least three other debates in practical philosophy: The essence of *evaluations*, the nature of *autonomous agency*, and the analysis of *akratic behaviour*. I am not claiming that if you take a stand on one of these issues, then you automatically have a solution to the others as well. But in all these discourses, a crucial problem is to identify certain elements in motivation as having priority over motivating desires in general. When some components of motivation are elevated to the status of 'evaluations' for instance, this is normally assumed to imply that there is something especially irrational about failing to put those components into action. Also, if a person's motivating desires do not conform to his evaluations, doubts may be raised as to whether his resulting actions really should be regarded as autonomous. It is not surprising to find that similar

solutions have been suggested concerning how to analyse autonomy, evaluation, and akrasia.

A good illustration of the interdependency of these three notions is the simultaneous suggestions about use of bi-level theories of desiring for the purpose of analysing autonomy, evaluations and akrasia, respectively. In the 70s, Gerald Dworkin developed an account of autonomy in terms of an agent's capacity to reflect upon his first order desires, identify with them and have the ability to change them. Under the influence of Frankfurt's widespread ÒFreedom of the Will and the Concept of a PersonÓ (1971), Dworkin developed and modified this bi-level theory of autonomy in *The Theory and Practice of Autonomy* from 1988. Frankfurt employs the bi-level view of desiring in analysing cases of addiction and weak-willed behaviour, and regards the capacity for affecting the efficiency of one's first order desires as being among the criteria for being a person. David Lewis suggests that desires about one's first order desires are identical with *evaluations* (1989) and this is how Michael Smith understands Frankfurt's proposal as well (1994 p.142).

Judging from these philosophers' different approaches, it appears plausible to adopt a terminology such that, very roughly, an autonomous intention is the result of desires that are in line with the agent's evaluations. At least in the sense that the agent *could* have made the triggering of his first order desires conform to his evaluations if he had chosen to. Conversely, it seems reasonable to think that weakness of the will is a failure consisting, roughly, in the agent's inability to form intentions that are in accordance with his valuing. If self-control is the opposite of akrasia, as Mele suggests in *Autonomous Agents,* autonomy is closely connected to the capacity of exercising self-control. (Mele 1995). But just to assume a loose interdependence in this way does not provide us with a substantial understanding of these notions.

It is not my ambition to develop analyses of autonomy, akrasia or evaluations here, but to make some points concerning the restrictions the BD model sets — or have been thought to set — on these concepts.

On one understanding of 'evaluation' — as equivalent with 'desire'—, the BD model excludes the conceptual possibility of acting against one's

strongest evaluations. This is not necessarily a *reductio* of that sort of analysis. We need an account of evaluations that explains our observation that we sometimes appear to act against better judgement. But the account should also explain why there is something paradoxical or unintelligible about people who convince us that their evaluations do not support their actions.

ÒValueÓ, ÒautonomyÓ and related terms are positively loaded. Examples like ÒnaturalÓ or ÒdemocracyÓ show that attractive terms always run a risk of overexploitation.[1] If my ambition in this chapter were merely to depict common usage, a long catalogue of different lexical uses would be necessary. I see no reason to believe that people use these words similarly. Some, like Socrates, apparently think of evaluations in a way that would make akrasia genuinely paradoxical. Others, like Donald Davidson, think that what makes Socrates' view paradoxical is that it Òit denies what we all believe, that there are akratic acts.Ó Their disagreement is apparently not over what we can observe in people's behaviour, but over how these observations should be described.

My ecumenical admission concerning actual usage does not imply that my view of the different uses of 'autonomy', 'evaluation' and the corresponding 'weakness of will' is just as permissive when it comes to how we *should* use these terms. To adopt a certain usage in this context is also to accept a norm concerning respect for others. I.e. to regard another's decision as autonomous, or as rooted in his evaluations, is to mark that there are certain prima facie reasons to respect that decision. The substantive principles of practical reason these different uses of the terms reveal must therefore be critically examined.

Before turning to specific suggested analyses of evaluations, I will state three restrictions I think that any notion of autonomy and evaluation should fulfil.

Firstly, the analysis must be *content neutral*. To avoid paternalism, we have to allow people's values to depart from our own. The elevation of a certain motivational state to the status of evaluation must not depend on the observer's evaluative commitments. Disgraceful or ridiculous evaluations are not disqualified by definition. (That does not contradict the admission

in chapter 1, that charity is an unavoidable component in picking out the rationalising causes of other people's behaviour. The question here concerns our classification of these causes, on the assumption that we already know which desires and beliefs have triggered the subject's actions.)

Secondly, the analysis must nevertheless do justice to the prima facie value of respecting autonomy and the prima facie unreasonableness of acting against one's evaluations. To be more specific, the account should make sense of the view that *capacity* for self-control (understood as the ability to bring one's motivation in line with one's evaluations) is valuable. This in turn presupposes that evaluations have a more important role to play in our lives than motivational states in general.

Thirdly, the analysis must explain how evaluations can have a part to play in motivation, within the BD model. This restriction is conceptual, not merely psychological. In other words, I presuppose value-internalism: To adopt a value is to be motivated. Within the BD framework, this means that *some* kind of link between evaluating and desiring must be present. Hare's argument for this supposition holds, I think: If talk about values merely expressed beliefs, holding a value-judgement and failing to act upon it would be as unproblematic as Òthinking a stone is the roundest in the vicinity and not picking it up, but picking up some other stone insteadÓ (Hare 1963 p.69). Since acting against one's better judgement *is* problematic, i.e. it is an apparently paradoxical behavioural phenomenon requiring additional explanation, there are reasons to think that expressions like "better judgement" necessarily indicate motivation.

Externalists might argue that Hare's example is unfairly rigged, since it refers to a property, roundness, which commonly not even in a contingent way is bound up with motivation.[2] The odd appearance of deviant cases could be explained, they might say, by the contingent fact that (perhaps for cultural or biological reasons) people actually are disposed to act upon their value judgements. To hold a value judgement and fail to act upon it should not be compared with failure to pick up the roundest stone (it is hard to see that as a failure at all), but perhaps with failure to avoid pain, or failure to

laugh at *A Day at the Races*. That would turn deviance into a biological or psychological problem, rather than a philosophical one.

Though that kind of comparison would be more fair than Hare's, I believe this attempted externalist strategy really would serve the internalist's point even better. When we compare our intuitions on cases of deviance from the connection between valuing and acting with our intuitions on cases of deviance from tight, but apparently contingent, connections between judgement of fact and acting, the difference is brought out more properly: Failure to avoid pain, or failure to appreciate Groucho Marx, are examples of odd *behaviour*. But, unlike our picture of acting against better judgement, our attempted *descriptions* of these deviant acts do not typically have a paradoxical air.

The preceding chapters have been intended to be constructive in their conclusions. Rival accounts of motivating reasons and intentions have been brought up and rejected insofar as these rejections have been supposed to shed light upon the BD model's entailments and limitations. That was my intention, anyhow. This last and concluding chapter differs in that respect. Three main suggestions about internal criteria for practical reason will be examined and eventually turned down. All three suggestions have been explicitly proposed as compatible with Humean BD motivation theory. The first is Davidson's Principle of Continence, suggested as the weak rationality criterion akratics breach. The second suggestion is the bi-level criterion of evaluations or autonomous desires, mentioned above. The third BD compatible theory of practical reason that I argue against is of the type Richard Brandt and Michael Smith have formulated in slightly differing versions. It says that you should adjust your desires to the desires you believe you would have if you were rational.

I cannot rule out the possibility that other plausible norms of reason can be stated and defended with greater success — or that these three norms can be refined and modified to meet my criticisms. This book ends inconclusively, in that respect. Nevertheless, I think that if these promising attempts to raise BD criteria of practical reason above mere instrumentalism of the Humean sort do not succeed, there are reasons to be pessimistic about other attempts of this kind as well. That does not entail a

general pessimism about the possibility to critically examine the goals we strive for, and affect our intrinsic desires in the right direction. It just indicates that principles of individual practical rationality may have less to offer in that project than we hoped for and that the nature of those problems is social, rather than individual. I.e. it may turn out that in judging the worth of people's goals, we cannot avoid considering their social roles, and also appeal to our own values and concern for them.

8.2 The Principle of Continence

The internalist account of evaluations that would fit most easily into the BD model simply identifies valuing with desiring. Since intentional actions are desire-based, that would also imply a ÒbackwardÓ connection: Doing implies valuing. Among ordinary language users, I believe that such a link often is presupposed. Tom Sawyer skilfully exploits that conceptual entailment in the episode where he gets his friends to pay him for letting them do the painting he has been ordered to do. His trick simply consists in pretending that he paints the fence for no further reason at all (not for fear of punishment, nor out of hope of reward). Tom's friends conclude that he must see some kind of value in the act of painting in itself. His behaviour would otherwise be incomprehensible. If acting did not imply valuing, then performing an action and failing to see any value in it would be unproblematic. In that case, Tom's strategy would not have worked.

The account of evaluations that emerges from Donald Davidson's early articles on philosophy of action, ÒActions, Reasons and CausesÓ from 1963, and ÒHow is Weakness of the Will PossibleÓ from 1970, equals evaluations with desires. It is therefore remarkable that, unlike the Socratic view of internal deviance, Davidson's is an attempt to do full justice to the paradoxical character of akratic actions. He does not want to solve the problem by giving up, modifying or stating more precisely any of the three fundamental assumptions generating the paradox. He asserts: "I am convinced that no amount of tinkering with P1-P3 will eliminate the underlying problem: the problem will survive new wording, refinement, and elimination of ambiguity." (1980 p.24)

P1. If an agent wants to do x more than he wants to do y and he believes himself free to do either x or y, then he will intentionally do x if he does either x or y intentionally.

P2. If an agent judges that it would be better to do x than to do y then he wants to do x more than he wants to do y.

P3. There are incontinent actions.

D. In doing x an agent acts incontinently if and only if:
(a) the agent does x intentionally;
(b) the agent believes there is an alternative action y open to him; and
(c) the agent judges that, all things considered, it would be better to do y than to do x. (1970, p.94)

Davidson wants to save the paradoxical ring, but of course not at the cost of contradiction. To show how this is possible is the main aim of his first essay on the subject: "How is Weakness of the Will Possible". His solution depends on a distinction between the two senses (presented in section 7.3) in which the agent can judge it better to x than to do y. An intention is identical to an “all-out” judgement about what is best, while the better judgement that an akratic intention breaches is of the relativised all-things-considered sort. ‘All things considered’ means, here, just as in the principle of induction ‘all available evidence considered’, although the nature of the evidence might be different in the practical case. In two later articles, "Paradoxes of Irrationality" and "Deception and Division", he introduces "mental compartmentalisation" to explain why the phenomenon occurs.

I have already declared that I am prepared to accept the conceptual possibility of separating, in one’s deliberation, conceptions of what is best to do, all things considered, from conceptions of what is best to do (full stop). I do not think that Peacocke succeeds in showing that Davidson’s distinction in practical reasoning could not be analogous with the distinction between evidence-relative all-things-considered probabilistic assertions and unconditional assertions accepted on the basis of them. Peacocke is right in pointing out that it is unclear whether Davidson intends to distinguish the normative or epistemic prima facie reservation from the

pure relativisation component in his connectives. But the analogy can be preserved even if that important distinction is spelled out, I think. In both cases, we can distinguish between the reason giving force as being conditional on available evidence, and the degree of credence we give to the hypothesis/practical judgement on those conditions. A relativised judgement about which course of action that would be most desirable, given available evidence, may be formed as a justifying consideration that moves an agent, and still fail to be decisive of his choices.

Hurley and Peacocke both claim that all-things-considered practical judgements are more closely tied to all-out judgements than Davidson thinks. If they were right, there could not be akratic acts compatible with Davidson's P1-P3. My own doubts about Davidson's principle of continence in practical reasoning are of a different kind. Although I find it possible, within the BD framework, that the kind of discrepancy Davidson describes may occur, I am not sure that this is the kind of internal deviance that must be prima facie unreasonable.

As Davidson says "there is no paradox in supposing that a person sometimes *holds* that all he believes and values supports a certain course of action, while at the same time those same beliefs and values cause him to reject that course of action." (1980 p.41 my emph.) Often we act without making any judgement on our desires and beliefs before action takes place. Even if we engage in such acts of deliberation, they must not be identified with the network of beliefs and desires which actually cause us to act — the motivational states that are tokened in the intentional background.

To display *continence* in acting is to form intentions that are in accordance with the following principle (Davidson's original formulation was quoted in section 7.3):

From '(relative to r, x is better than y) and (r is the total available evidence)' infer 'x is better than y'

The imperative form of the principle is significant. ÒInferÓ should not be taken literally, but as a substantial recommendation. The akrates, according to Davidson, is a person who fails to make his unconditional all-out

practical judgement in accordance with that command. In this way Davidson fulfils his ambition to avoid representing the akrates as committing a simple logical blunder. So Davidson's analysis of akrasia is not as radical as the view Annette C Baier compares it to: "It is as if Achilles tells the tortoise, 'if you don't accept modus ponens you needn't conclude "q" from "if p then q and p"'" (1986 p.119).

Is Davidson’s principle of continence the norm of practical reason we are looking for? To begin with, it is clearly a content-neutral rule, and Davidson’s analogy with the principle of induction appears to give it a certain normative force as well. That analogy with relativised and unconditional probability-assignments illustrates a parallel contrast between two features of practical reasoning — the conditional judgement and the ÒjudgementÓ that issues in action.

At a closer look, I do not think that the principle of continence could be given a similar status as a practical piece of advice. The person who continuously chooses the paths that are least likely to lead him to his destination, given the map he has got, will soon get lost, or at least be late. But the nature of the evidence is rather different in the case of continent acting. In this case, the pathfinder’s own ends and values are among the mapped data. His conditional judgement asserts what he *holds* that he believes and values, and relativises a practical conclusion to that assertion. It is not evident that someone who forms judgements about his own motivational apparatus will be better off in daily life than someone who does not. It is not certain that the course of action an agent’s beliefs and values *cause*, is worse, in terms of her own values, than the course of action she *holds* that these beliefs and values support.

To begin with, we have no direct access to our desires. As Mischel (like Gopnik and Nisbett & Wilson) noted, we have a tendency to endow ourselves with the motivational states that are judged as plausible causes of our behaviour, or as reasonable explanations in the light of our own interpretative theory. Furthermore, the ascription of desires to a person consists to a large extent in making assumptions about what can be conditionally predicted about him. The capacity to make such predictions is limited from a first person perspective. So there are epistemological

limitations to self-knowledge of motivation that simply make our estimations of our own beliefs and desires a shaky ground for assumptions about the best course of action.

Furthermore, a more formal reason for discrediting the principle of continence might perhaps be advanced here. If the 'all things considered' judgement that an agent fails to act upon is thought of (by the agent when he forms the judgement) as being valid relative to all relevant beliefs and desires in the agent's mind, then, one might argue, it cannot be conditioned on *all* relevant beliefs and desires in his mind. (Some underpinning of this claim follows below.) Hence it cannot be expected to correspond to an unconditional all-out judgement which results from all relevant beliefs and desires in the agent's mind. Then one could hardly blame someone because of his failure to make his unconditional all-out judgement correspond to a judgement seen as conditioned by all relevant beliefs and desires in his mind.

If I want to predict tomorrows weather, I can consider all available facts about the weather situation as 'the total available evidence' and think of my predictive judgement as being relative to this evidence. But if I want to predict my intentional behaviour, I cannot, in principle, base my prediction on every relevant belief and desire in my mind. One of the relevant elements (relevant, that is, in the sense that it may affect the final decision) is how I think my desires and beliefs relate to the course of action. Though an omniscient spectator might think of my value-judgement as being conditioned by all beliefs and desires present in my mind, this way of thinking is not open to me when I form the judgement.

A possible countermove is to assume that the belief about relativisation can be self-reflexive; ‘r’ includes ‘my judgement is relative to ‘r’’. But it may be doubted whether that sort of self-reflexivity could be allowed without getting us into genuine paradoxes. In other words, it is not clear that Davidson can allow 'r' in the principle of continence to be identical with everything I believe to be true and relevant when I make my judgement. Then I could not think of my judgement as being acceptable relative to r.

Among the things included in a judgement made on broadest possible base must be considerations about the limitations of my knowledge. Otherwise it would anyway be inappropriate to think of the 'all things considered' judgement as being a particularly good action-guide. There is nothing incontinent in failing to act upon '(relative to r, x is the best thing to do) and (r is the total available evidence)' if one can add the premise 'r is insufficient for a wise decision'.

If I think of my value-judgement as being held relative to a certain body of evidence, this evidence cannot be thought of as including all relevant considerations in my mind. One consideration is left out: The judgement is held relative to a certain body of evidence. Implicitly, this consideration carries with it the idea that the available relevant evidence can be incomplete, that is, that all real evidence possibly might alter the judgement. That is a self-evident *possibility* in the case of induction, and it is even imaginable that a person after having experienced a series of erroneous inductive inferences forms a habit of taking it for granted that the expectation he regards as plausible, given all evidence available to him, is implausible. (It may require some elaboration to make the rational force of Hempel’s induction rule withstand that possibility.) But when it comes to the principle of continence, I am inclined to think that this kind of self-mistrust not only is possible, but that it should be expected. The fact that I may have left out some of my own goals in the survey on which I base my conclusion about how to act, is a thing that would make me withdraw my judgement, if I realised it.

Davidson admits that there are difficulties in making clear the character of the conditional 'all things considered' judgement which, in the case of continent acting, is supposed to correspond to the all-out and unconditional judgement adjoining intentional actions. He does not proceed to develop an account of what it is to judge something, all things considered. Instead he attempts to avoid the problem by modifying his definition of incontinence ("D") so that it becomes independent of the idea of an agent's total evidence (1980 p.40). An incontinent agent, in this modified sense, "does x for a reason r, but has a reason r' that includes r and more, on the basis of which he judges y to be better than x". This definition allows, Davidson thinks,

that there are incontinent actions even when no judgement is made in the light of all reasons.

As Davidson says, "it might also have been incontinent of him to have done y, since he may have had a better reason still for performing some third action z". More embarrassing to this definition is the possibility that, on the basis of r", which includes r' and more, the agent judges, after all, x to be better than y. The doing of x would then fulfil Davidson's definition without being incontinent in any intuitively plausible sense of the word. (I suppose here that he does not have a further reason which includes r" and on the basis of which he judges y to be better than x etc.). To exclude this possibility, one would have to add a clause like “and there is no reason r" that includes r' and more, on the basis of which he judges x to be better than y.” But what would that require from the agent? The possibility that there is such a further reason that could include r', and be relevant to his ranking of x and y, can only be excluded if there is nothing in his beliefs and desires that would change his ranking, if he came to think about it. We could only assume that about him if it is true that he would rank x above y when he takes all his beliefs and desires into account, including the belief that his judgement is conditional on r'. This appears to get us back where we started: Including the idea of relativisation in the evidence on which the akrates bases his better judgement requires him to form a self-reflexive belief of a problematic kind.

Summary of 8.2 Donald Davidson’s principle of continence says that the non-akratic agent infers ‘x is better than y’ from ‘x is better than y, given all available evidence’. The normative force of this assumption lends support from the analogy with a principle of induction, which says that a rational person infers ‘x’ from ‘x is probable, relative to all available evidence’ In one respect, the analogy holds; neither of the two inferences are deductively valid. (Some have thought so about the practical case.) However, the different nature of the evidence makes it less clear that it is irrational to breach the continence rule, than it is when it comes to violation of the rule of induction.

Among the most important things to consider in the practical case are the agent's own motivational states — his aims and goals. For reasons mentioned earlier, there are formal as well as empirical reasons to be pessimistic about our ability to get a truthful picture of our own motivation. General self-mistrust about the reliability of available evidence would therefore be less perverse in the practical case, than it would in the theoretical one.

A special difficulty for the principle of continence may arise due to its reference to *all* evidence. In the theoretical case, it is reasonable to think that the fact that 'x' is held relative to the subjectively available evidence normally would be of little relevance to the subject's belief. I.e. if that fact was included in her survey of the evidence, it would not change her belief that x. In the motivational case, the relation between the body of evidence and the inferred recommendation is more crucial. Among the data she should observe when deliberating about what is best, this relation should count. The conceptual possibility of including the very relativisation in her all-things-considered judgement presupposes that she can include a meaningful self-reflexive belief of the right kind in her judgement. That possibility might be questioned.

8.3 Two Notes on the Psychology of Internal Deviance

8.3.1 Breakdown of Reason Relations

Davidson explicitly intends his account to be applicable to cases in which the agent is "aware that he is not acting in accord with his own best judgment" (1970 p.40). But how can an agent who normally acts in accordance with the principle of continence suddenly fail to apply it? What is the agent's reason for doing one thing when he believes it would be better, all things considered, to do another thing? For this, "the agent has no reason", according to Davidson's initial characterisation.

> We perceive a creature as rational in so far as we are able to view his movements as part of a rational pattern comprising also thoughts, desires, emotions and volitions. /.../ But in the case of continence, the attempt to read reason into behaviour is necessarily subject to a degree of frustration. (1970 p.42)

It may look as if Davidson ends up just restating the paradox without answering the question of why it happens. But he gives one substantial answer: there is no reason explanation to be found. This answer risks drawing the notion of incontinent action closer to unintentional behaviour than Davidson explicitly intends. In "Paradoxes of Irrationality" and to some extent in "Deception and Division", Davidson develops his explanatory account. The incontinent agent, as well as the self-deceiver is, to use Pears' expression, divided against himself. (1986 p.131).

Davidson starts out by distinguishing between mental causes operating as reasons and those operating merely as causes. A mental state may cause another without being a reason for it. The relation between them need not be logical. Recognising a tune may cause me to remember a name; a young man may think he has a well-turned calf because this thought is pleasurable to him, etc. (1982 p.305 & 298).

It must be admitted that this kind of non-inferential mental causation is common. At the same time it is not apparent how this idea should be regarded in relation to the interpretative view of intentional explanation in general. Which criteria do we use for supposing that two mental states are causally related? It seems to me that only if I can reconstruct some intelligible chain of association will I be inclined to suppose that one particular thought or desire of mine is the cause of another particular thought or desire of mine. I may recognise a tune and remember a name simultaneously; I may also recognise the smell of coffee and remember some of Donald Davidson's views on mental compartmentalisation simultaneously. But I will hardly be inclined to think of these thoughts as causally related unless it makes sense against the background of my further beliefs and desires. (Perhaps I remember something connected both with the tune and the name, for instance the lyrics or the place where I first heard it.)

So even in the case where the relation between causally linked mental events is non-inferential, we seem to rationalise them by placing them within a pattern of desires and beliefs. We ascribe someone wishful thoughts, for instance, when we see some understandable relation between the objects of his desires and the thoughts we suppose result from these desires. The mere existence of non-inferential mental causation does not support the idea of a divided mind.

"It is far more plausible", as David Pears says, "to restrict the scope of this kind of theory to cases where a mental cause operates as a reason but produces its effect irrationally" (1986 p.136).

Davidson's examples of non-rational mental causation are, however, less problematic than the type of examples his theory of non-rational mental causation is supposed to account for: Cases in which the causally related thoughts or desires are held to be internally inconsistent by the agent who holds them. If we want to use the notion of non-rational mental causation to account for internally inconsistent beliefs and desires, mental compartmentalisation is likely to be the next step. Incontinent action occurs, in Davidson's view, when a person

> holds that all he believes and desires supports a certain course of action, while at the same time those same beliefs cause him to reject that course of action. If r is someone's reason for holding that p, then his holding that r must be, I think, a cause of his holding that p. But, and this is what is crucial here, his holding that r may cause his holding that p without r being his reason; indeed, the agent may even think that r is a reason to reject p. (1970 p.41)

When the agent judges r to be a reason to reject p, which he is caused by r to hold, then he must, according to what it means to have a reason (in the Davidsonian sense) view r as being related to a network of beliefs and desires in which his holding that p is excluded. Mental states and events are, as Davidson puts it, "constituted the states and events they are by their location in a logical space" (1982 p.304). If my belief in r causes me to hold p, which I hold to be incompatible with r, then r, viewed as a reason for rejecting p, must belong to a different logical network than my judgement that p. In the defining case of incontinence, my unconditional

judgement that x is better than y is caused, across the boundary of mental subdivision, by r, included in r', on the basis of which I consider, within another logical network of my mind, y to be better than x.

Davidson's use of terms like "quasi-autonomous" and "semi-independent" makes it difficult to pin down his theory in a position where it can be critically examined. How strong is the autonomy? The concluding remarks in "How is Weakness of the Will Possible?" lead the thoughts to the far end of the spectrum: "What is special in incontinence is that the actor cannot understand himself: he recognizes, in his own intentional behaviour, something essentially surd." The experience of recognising a genuinely surd but in a sense intentional element in one's own behaviour is typical of the most clear and dramatic example of mental compartmentalisation: Split-brain patients.

In this case the division is so strong that most people hesitate to describe the split-brain patient as one person. The person with whom one can communicate verbally (usually the one tied to the left hemisphere) does, for instance, frequently express surprise or aversion towards behaviour rooted in the other mental compartment within the body. If incontinent or self-deceptive agents were divided as strongly as this, the problem of inconsistency would be as unproblematic as inconsistency in opinions among different persons within a group.

It is clear that Davidson wants to see the divided agent as one person, responsible for his action. "The analogy does not have to be carried so far as to demand that we speak of parts in the mind as independent agents /.../ The breakdown of reason-relations define the boundary of subdivision" (1982 p.304). As we have seen, breakdown of reason-relations must mean something stronger than mere non-inferential mental causation if this breakdown shall be able to account for incontinence. The required mental causation must not only be non-logical, but illogical. How is the unity of the agent to be upheld, then? Our criteria for supposing that mental states are causally related seem to be dependent upon the possibility of tracing some reason-relations between them.

It would not help to add the qualification that there are non-inferential causal relations between the subsystems. Because, as Davidson makes clear

in his analogy, mental states may causally influence mental states in other persons as well. "What I have tried to show" Davidson says, "is that the very general features of psychoanalytic theory that I listed as having puzzled philosophers and others are, if I am right, features that will be found in any theory that sets itself to explain irrationality." But it seems to me that it is precisely our way of viewing an agent as a consistent intentional system that allows us to postulate beliefs and desires even when they are unknown to the agent himself. What this shows, I think, is that there is a tension between Davidson's idea of intentional explanation as an active imposition of a rationalising interpretative framework, and his theory of akrasia as requiring mental compartmentalisation.

That conclusion could be also turned into a more constructive move about the analysis of akrasia, along the lines suggested e.g. by Olav Gjelsvik. That is to remove akratic behaviour from the sphere of full-blown intentionality and assume that this phenomenon requires us to apply other explanatory schemes than the ones employed to frame the original paradox. Gjelsvik describes the clash as one between a ÒnaturalistÓ view of agency, and a traditional BD story (Gjelsvik 2000a). The BD model would not be challenged by that solution, which is a way out of the paradox, compatible with Socrates' renunciation of intentions executed against better judgement. Justin Gosling says in *Weakness of the Will: "*"What Socrates has to do, to win conviction, is show how apparent cases of people being overcome by fear, pleasure or the like, so as to act against their better judgement, are really cases of people doing what they think best." (1990 p.17). But another possibility for him would be to show that these are really cases of people acting unintentionally. The typical philosopher (who wants to have her paradox and solve it too) would react to this move with another challenge: If akratic actions are unintentional, how come that these behaviours have such a paradoxical appearance — unlike reflexes and other unintended acts that may be unwanted? The answer to that challenge is to understand akratic acts within a scheme sufficiently Òclose to the system in which intention has its original home.Ó (Gjelsvik 2000 p.124). As I argued towards the end of ch.5, when we talk about actions in the wide causal sense — actions that are not necessarily intentional under any

description — it is quite possible that actions (in the broad sense) are intentional to a certain *degree*.

8.3.2 Proximity and the Individuation of Options

The naturalised explanation Gjelsvik appeals to is George Ainslie's descriptions of instability of preferences over time (Gjelsvik 2000a, 2000b). Just as Amelie Rorty's suggestions about how social explanations of incoherence in an agent's evaluations may help us understand how one type of apparent akrasia can occur, Ainslie's psychological explanation avoids the conceptual problem about akrasia. (Rorty 1997) Like other mentioned observations of instability and incoherence in motivation, these types of accounts show that a common reaction to Socrates' view — that it simply denies what we all believe — is oversimplified[3].

As I noted in ch.5 on the nature of intentions, there is a certain ambiguity in common between Brandt's and Mele's picture of the choice situation in Mischel's much discussed delay experiments. It is unclear whether the children's noted rankings were between options, where the expected length of delay was included as part of the option (as it should, then) or whether the noted change in rankings concerned the item that was part of the option. Brandt and Mele both regard the younger children's change of ranking as a form of akrasia. But (as Wlodek Rabinowicz made me note) this would not be a case of strict akrasia if the children first were told to rank items, and the change in rankings then were measured via their choice between options. One might add that even if constantly presented as a choice between the option of having the greater reward at t2, or getting the lesser reward at t1, the gradual difference in proximity between the two rewards, as time passes, might in itself be seen as part of the two options, as seen from a specific point in time.

Beside the experiments by Mischel, Mele also appeals to Ainslie's theories. The most relevant contribution of Ainslie's to the present issue is his theory of how reward value is discounted. It is based on observations of animal behaviour and confirmed in psychological experiments on children

and adults (I take the liberty of relying on Mele's and Gjelsvik's own descriptions here —1987 and 2000). The starting point is that the value of rewards appears to be discounted naturally with delay, as one might suspect. I.e. motivation towards a certain reward increases as the time for the reward approaches. Goods previously ranked low will be ranked above distal alternatives, previously ranked higher.

The most interesting thing, though, is that in humans and even e.g. in pigeons (when trained) the discount factor in itself then varies with the time distance to the alternatives. A Ònumber of experiments have shown a preference for a small earlier reward when the delay is short, and a preference for a larger but later reward when the delay is longÓ (There is some evidence for regarding the discounting as hyperbolic — Gjelsvik 2000 p.116). If someone believes that his preference for the better is apt to change, he can exercise self-control. ÒHe can bind himself, for example, as did Odysseus, or employ other techniques that increase the motivational force of the preferred alternative.Ó (Mele 1987 p. 85) Mele makes clear that Ainslie does not suppose that this deliberate controlling device, when understood in the psychological internal sense, is a *better judgement,* in the sense required to regard failure to execute this kind of self-control as a case of strict akrasia. However, that interpretation Òcan certainly be derived from Ainslie's workÓ, says Mele (1987 p.85).

That is a mistake, I believe. Mele describes the situation so that when strict akrasia occurs, the agent is in a manner overcome by the motivational efficiency of proximity, which makes him act contrary to his rankings. Mele distinguishes four elements in this explanation of akrasia: 1) The agent's level of motivation to do the prospective continent act, 2) the agent's earlier level of motivation to do the akratic action, 3) the agent's failure to make effective use of self-control and 4) proximity. (1987 p.85) As with Mischel's experiment, we need here to identify the options in order to get a correct description of the action. Mele's picture of it as akratic seems to presuppose that options are kept invariant in spite of change in proximity. In turn, that seems to picture the Òbetter judgementÓ as a device identifying options from an intertemporally neutral point of view.

But the lesson to be learnt from Ainslie's experiments seems to be quite the contrary. What they illustrate is that we have no such gifts — we are stuck with our temporal preferences. It is another matter that present among these temporal motivational sets are sometimes (only in animals capable of learning) liabilities to direct attention, form habits, follow rituals and engage in other motivation affecting procedures — capacities dependent on our ability to view ourselves as temporally extended agents. When expected delay is sufficiently long with respect to two alternatives, we have the capacity to evaluate them without being biased by their difference in proximity — which will become proportionally greater as the alternatives come closer to us. If I am unable or unwilling to affect my motivation with methods of the kind described, then proximity of rewards will play a greater role for my evaluative assessment of the options at a time closer to the reward. On the other hand, if I succeed in affecting my motivation, this simply means that my ranking of the options wherein the rewards figure (at a time closer to the rewards) will be less affected by proximity.

A comparison with a simple external pre-commitment device might be illuminating: In a calm moment, you tell your family that if they catch you smoking any day, they have the right to make you do the dishes that day. The device can be reliably executed but nevertheless fail to be effective. When reward becomes sufficiently proximal, you value the option 'smoke + dishes' higher than 'no smoke + no dishes'. In order to depict your act of smoking as akratic then, we can hardly appeal to your act of initiating the device — not without simply presupposing an overall (time neutral) measure of rationality. In the moment of acting, you act out your evaluations of the option as seen with the specific proximity to the alternatives of that moment.

Summary of 8.3: Davidson's description of akratic actions as occurring due to a breakdown in reason relations could be combined with a less strict view of internal deviance than the one proposed in Davidson's definition of akrasia. I.e. this sort of breakdown may be seen as an element making the akratic act at least partly unintentional. Mele may be quite right in assuming that Amelie Rorty's social explanation and Ainslie's (along with

Mischel's and others) well testified theories about instability in motivation over time Òjointly provide a basis for a very plausible explanatory hypothesisÓ of akratic action. (Mele 1987 p.84) This basis leads to a less strict view of akrasia, though — and an elimination of the paradox, rather than a solution. Like Gjelsvik's views on the Ainslie type of approach, my picture of this type of explanation differs, in this respect, distinctly from Mele's.

Weakness of the will is still a genuine problem; it ruins people's lives sometimes. But it is a problem about how to live a good life, and how to behave towards the other — an ethical problem, in a wide sense of the word.

8.4 The Authority of second Order Desires

It should be clear from what I say above and elsewhere in this book, that I am inclined to believe that natural Ònon-paradoxically ringingÓ analyses may account for many varieties of internal deviance, including the undesirable case of akrasia.

There are still reasons, though, to pursue the question as to whether practical rationality norms in any substantial way can be used to elevate some elements in our motivation to being worthy of a special kind of respect, over and above the rest of the beliefs and desires that move us. Is there, e.g., anything about the evaluative outlook of a Jehovah's Witness, an outlook that few of us share, that ought to make us respect her attitudes against being saved by blood transfusion? Compared, for instance, to the destructive resolution of a teenager to starve herself, because she thinks it makes her prettier? Both attitudes are hard to understand, but is there an explanation in terms of personal values or individual rationality that would make the commonsensical distinction between them justified? Common sense, I take it, says that the Jehovah's Witness, unlike the teenager, has a certain prima facie right to have her desire respected. The teenager's anorectic behaviour need not be akratic, but it is seen as somehow less autonomous than the religious refusal. Is it possible to do justice to this

intuition within the BD model without giving up the commitment to content neutrality?

An approach developed in many different forms, as I made clear in the introduction to this chapter, is that *higher order desires* have a certain authoritative status, and that ability to enforce one's second order desires is constitutive of autonomous agency. David Lewis, as I remarked earlier, identifies them with evaluations, while Gerald Dworkin utilises second order desires in his analysis of autonomy. Dworkin suggested in an early paper that autonomous first order desires must *actually* have been made effective via enforcement of second order desires. That would disqualify most of our effective desires, though, and in (1988) he suggests that autonomous agency follows from the *capacity* to form authentic desires. Authentic first order desires are in line with our second order desires. Though we may not always reflect upon them and endorse them as reasons for actions, *had* we done so, then we would have made them effective. The capacity to reflect upon and thereby affect first order desires is a condition for autonomous agency.

Dworkin's approach is quite in line with Frankfurt's view of the authoritative role of second order volitions, and the points I will make apply equally to Lewis', Frankfurt's and Dworkin's suggestions.

To begin with, some might think that the BD model's view of desires as identified via behaviour makes second order desires explanatorily redundant. On a crude behaviouristic approach, this worry might have been justified. If preferences did not reflect underlying desires, but were nothing but relations between options, as revealed in overt choice, then it might have been difficult to distinguish the desire for a desire to ϕ from a desire to ϕ. However, on the realistic dispositional BD model view there are, as we have seen, a variety of overt and internal behavioural signs typical of desires to affect one's own motivation. Focussing attention, forming foregrounded practical judgements, and creating internal and external pre-commitment devices of various sorts are the typical examples.

As Michael Smith makes clear, those who analyse evaluations in terms of higher-order desires "face a formidable objection", originally formulated by Gary Watson (1975 pp.107-109) against Harry Frankfurt's view in

"Freedom of the Will and the Concept of a Person". The objection is that no reason is (or can be) given for giving authority to any particular level of desiring. [4]

Why identify valuing with second-order desiring? Why not third order, or fourth order, or...?

The implication of the question is, of course, that each identification is as plausible as any other. But if each is as plausible as any other, then *all* such identifications are equally implausible." (Smith 1992 p.342)

Note that the objection is not that this way of reasoning gets us into an infinite regress. That objection is forestalled by Dworkin in (1988) who simply assumes that authorising power is given to the level which *actually* is highest — normally the second. It is tempting to pursue the regress objection and ask why the lack of second order authorisation in that case should undermine the authenticity of first order desires when the ground level actually is the highest one. If a second order is required then, why not require a third order to authorise the second, and so on. Smith's and Watson's objection is more direct, but to the same effect: The elevation of second order desires is just arbitrary.

There is no reason to presume that second order desires are more well founded than our first order desires, for instance. Our hopes and ambitions concerning our own motivation may be products of vanity, conceit, worship of authorities and other irrational conditions. In an earlier paper, Dworkin adds to his requirement of second order identification with first order desires the claim that Òidentification is not in itself influenced in ways which make the process of identification in some way alien to the individual.Ó (1981 p.61) The issue becomes, then, a question about when influences are *alien.* My strong suspicion is that content neutrality cannot be upheld in working out such criteria. And if the correct aetiology of influences is the vital point, why not go for that directly, without worrying about which level these influences operate on?

Smith and Watson have a forceful point. It also underpins my assumption about the importance of separating the descriptive enterprise of analysing our linguistic practices concerning autonomy and evaluations,

and the partly reformatory project I am presently dealing with. I am not only interested in how we talk about autonomy and evaluations, but also in which authority certain motivational states *should* have. The objection gives us strong reasons to reject any special authorisation of second order desires.

But why should that admission force us to abandon the view that, as a neutral fact about how words are used, people actually and without inconsistency refer to their higher-order desires with the term 'valuing'? Could not people use terms like 'evaluation' or 'better judgement' to stand for *any* level of higher order desiring which happens to be in conflict with the sublevel it is directed towards? Someone might e.g. think that her own ambition to become even more dedicated to her work really reflects an unsound competitive instinct; her third order desires are in conflict with her second order ones, and she might describe her third order desires as her evaluations. If they could, this means that the arbitrariness Smith appeals to is compatible with a conceptual connection between valuing and desiring to desire: 'Valuing' could refer to desiring at a higher but otherwise unspecified level. The arbitrariness of a preference might be a good reason for denying that it *ought* to have normative force. As far as I can see, that would not necessarily threaten the claim that common language-users actually express such preferences in evaluative terminology.

Smith himself stresses, e.g. in ÓInternal ReasonsÓ, that his theory is *conceptual* and that its elements Óshould manifest themselves in the way we talkÓ (1995, p.121). He argues that it is a contingent fact that someone who values ϕ-ing (usually) also desires to desire to ϕ. Frankfurt and David Lewis are diagnosed as conflating this fact with a conceptual truth (p.343). But within that descriptive framework, I do not think that the objection is open for Smith to use against Lewis and Frankfurt. The force of it derives from assuming that ‘evaluation’ and ‘autonomy’ are normative notions to some extent.

This underpins my initial assumption that the elevation of certain motivational elements to *evaluations*, or motivational structures to *autonomous agency* really reveals *norms* of reason.

Otherwise his argument is strong and sound. It is implausible to claim that people *ought* to adjust all their desires to, for instance, their second-order desires. We must be given reasons for assuming that second-order desires have priority over, let's say, first-order desires, or third-order desires. Frankfurt and Lewis provide no such reasons.

Summary of 8.4: Second order desires have been identified with evaluations, and the ability to enforce such desires has been suggested as essential to autonomous agency. There seems to be many situations in which people refer to their second order desires in these terms. However, second order desires do not deserve the special respect we appear to show them by giving them this role for evaluations and autonomy. The second level of desiring has no a special authority, and the assumption that a rational agent gives priority to her second order desires is just arbitrary.

8.5 If You Were Rational, What Would You Do?

Like R.B. Brandt in *A Theory of the Good and the Right*, Michael Smith thinks it is a good thing if people in evaluative dilemmas ask themselves what they would do, were they fully rational. It is even a platitude, Smith says, that "what it is desirable that we do—that is, what we have [normative] reason to do—is what we would desire if we were rational." This theme has been developed in a series of articles, as well as in Smith's *The Moral Problem.*

Brandt defends the idea that questions about the good, or the best thing to do, *ought* to be replaced with questions about what I actually would do if I was fully rational. He even proposes a linguistic reform, so that our value-terms come to express such beliefs, instead of the vague attitudes they now actually express. Our normative language would gain precision and our questions about "the good" would become possible to settle empirically. Nevertheless, he claims, opinions about "the good" in the new stipulated sense will still be as action-guiding as in the original sense (1979 ch.I & VII).

Smith's view differs fundamentally from Brandt's in that Smith also puts forward the *descriptive* claim that the Platitude – 'the desirable thing to do is what we would desire if we were rational' – "does give the content of our evaluative thought" (1992 p.348). As I noted above he claims that such platitudes Óshould manifest themselves in the way we talkÓ (1995, p.121). My own linguistic intuitions on the actual use of celebrated terms like ÒevaluationÓ and ÒautonomyÓ is more ecumenical, as I have made clear. In some cases, it does seem quite plausible to think that people express beliefs about what they would choose if they were rational in terms of what they Òvalue.Ó[5]

My interest here is confined to the *normative* plausibility of this claim. To paraphrase Smith's alleged platitude: Is it desirable that we do — i.e. do we necessarily have normative reason to do — what we *believe* that we would do if we were fully rational?

Smith proposes a quite moderate internalistic claim he thinks an analysis of 'valuing' should meet:

> If an agent judges that it is right for her to ϕ in circumstances C, then either she is motivated to ϕ in C or she is practically irrational" (1994 p.61)

R.B. Brandt gives two reasons for thinking that beliefs about what I would desire if rational should motivate me: First, irrational desires and aversions deprive me of well-being, which is something I desire. Second, it is an empirical fact that most of us desire to be rational. Brandt points out "that the foregoing recommendation of rational desires depends on the prevalence of other desires". (1979 p.159)

Beliefs alone are impotent within the BD framework which Smith, on the whole, seems to accept. Therefore, one might think, there must be some kind of desire my valuing (in Smith's sense) can appeal to if I am to be motivated. In a comment on Smith's book, Ingmar Persson has pointed out that Smith's analysis can meet his moderate internalistic claim if practical rationality is supposed to *imply* having a desire to have rational desires (1994).

Although some formulations suggest that Smith has such a stipulation in mind[6], there are two main problems with this suggestion[7]: It would be *ad hoc,* since accounts of practical rationality usually have no implications concerning specific meta-desires. What practical rationality in the conventional sense requires is merely the right kind of relations between agents' values, desires and actions. Smith gives no reason for supposing that his view of practical rationality differs from the conventional in this sense. Furthermore, the additional clause would trivialise Smith's claim that his account of valuing is internalistic. His view appears, rather, to be that the tendency to desire what you believe you would desire if you were fully rational is a disposition compatible with the epistemic disposition to believe q, if you believe that p, and that p implies q.

Smith's internalism could probably be upheld without the stipulation of an additional desire to be rational. In ÓInternal ReasonsÓ, Smith puts forward the following proposal:

> We can ask ourselves whether we wouldn't get a more systematically justifiable set of desires by adding to this whole host of specific and general desires another general desire, or a more general desire still, a desire that, in turn, justifies and explains the more specific desires that we have. (p.114)
> /.../
> If we do come to believe that our more specific desires are better justified, and so explained, in this way, then note that that belief may itself cause us to have a new underived desire for that more general thing. (p.115)

Partially constitutive of having a systematically justified set of desires is, according to Smith, the set's *coherence* and *unity*. At risk of oversimplification, one might perhaps express his view like this: Suppose you come to believe that acquiring a new desire for ϕ-ing will make your set of desires more coherent. On account of this belief, you will then, automatically come to desire ϕ-ing. The idea is not incompatible with the Humean impotence thesis, since the imagined increase in coherence (i.e. elimination of incoherence) must presuppose that your belief about ϕ-ing appeals to your initial set of desires. It does not seem implausible, I think, to assume (as a conceptual thesis) that practically rational persons will

come to desire things when doing so eliminates internal inconsistencies, since such a mechanism affects the attainability of their initial goals.

Smith offers no example of the procedure, but perhaps this would be one: Suppose a doctor often has to struggle with two desires which cannot always be mutually satisfied: Her desire to preserve human life, and her desire to relieve pain. After reading Peter Singer, she comes to think of a new, more general possible aim – the avoidance of frustration of preferences. Her realisation that this aim justifies, explains and resolves the conflict between both her initial desires is enough to make her desire that new goal. The situation is not that she, as it were, looks at a possible desire Ófrom the outsideÓ and asks herself whether the desire would make her set of desires more coherent. She is, in her deliberative thinking, concerned with the value of a certain possible object of desiring, rather than with the value of her potential desire for that object. Her realisation that avoidance of frustration might be valuable appeals, nevertheless, to her initial desires. I am therefore inclined to agree with Smith that there is one sense, compatible with the BD model, in which the realisation that a certain desire would be rational to have will actually cause motivation in a practically rational person.

Smith states that coherence and unity are *partially* constitutive of having a Ósystematically justified, and so rationally preferable, set of desiresÓ (1995, p.115). As far as I can see, that admission leaves room for another possible objection. If coherence and unity merely partially determine the rationality of an agent’s set of desires, the possibility of mutually incompatible but internally coherent sets of desires must be considered when judging the rationality of a certain desire. As David Velleman plausibly has argued against Brandt, people often consider motivational changes "even though they strike us as feasible only through non-cognitive means" (1988 p.357). Since "the possibility of non-cognitive therapy puts many sets of well-informed desires within our reach", we might have to consider several incompatible sets of well informed desires". (p.363) In that case, a practically rational person might have to consider several optional values which could be combined with distinct but reachable and coherent sets of desires. In other words, the argument above might have to be

restricted: Practically rational persons will come to desire ϕ-ing if they imagine that a desire for ϕ-ing will make their initial set of desires more coherent, *provided* that they do not consider alternative coherent sets of desires within reach (through non-cognitive means). If they consider such alternative sets, non-cognitive factors will probably determine whether they choose the most rational (in the sense only partially determined by coherence) option. In other words, "partially" opens Smith's rationality criterion to relativism, unless it is supplemented in a proper way.

This is Smith's Platitude: *What it is desirable that we do is what we would desire if we were rational.* Granted that it is a platitude that the desirable thing to do is the thing we would do if we were practically rational, it is nevertheless an open question whether it is desirable that we do what we *believe* that we would do if we were rational. It depends upon the reliability of that kind of beliefs.

How can we form beliefs about what we would do if we were rational? As Brandt emphasises,

> the best, the rational, the fully criticized choice is necessarily one which aims at realizing some valenced goals somehow. A person's wants and aversions (possibly altered from what they now are, in a way to be discussed at a later stage) are necessarily relevant to what is the best or rational thing to do. (1979 p.67)

Since Brandt specifies the way in which wants and aversions possibly are altered in order to be rational, the rational thing to do for me is a detectable function of my actual desires. The function consists in a substantive criterion of rationality. So the starting-point for a judgement about what I would desire if I were rational is what I actually desire. And where I end up depends on the nature of that starting-point.

It does not seem to worry Brandt that this way of reasoning might lead to a mild form of relativism. On the contrary, he makes clear that questions about, e.g., the rationality of egoism, or benevolence, will necessarily be dependent upon which fundamental desires the agent in question has to begin with, and *how* these desires are acquired (Are they conditioned or native? If conditioned, is it by a process liable to be influenced by irrational elements? Etc.) Like Hume, Brandt denies that reason alone is capable of

answering normative questions. The difference between Brandt's position and Hume's is merely that Brandt's concept of rational criticism enables him to consider the irrationality of intrinsic desires; to Hume, only instrumental desires seems to be open to (indirect) rational criticism.

Smith does not question (at least not explicitly) the assumption "that the desires an agent would have if she were fully rational are themselves simply functions from her actual desires" (1994 p.165). His point is rather that it is a mistake to conclude from that assumption that reasons are relative.

Smith seems to regard 'relativity' as implying that normative reasons *actually*, and not only possibly, are different for different people. Therefore he regards the truth of relativism dependent upon whether "rational agents would actually end up converging on a single set of values." Since "there is no proof that they would not", we cannot rule out the possibility that normative reasons are non-relative in advance (1992 p.355). I am not, here, questioning that part of Smith's argument, since I believe that the possibility of congruence to which he appeals is compatible with my point about the starting-point of forming of beliefs about one’s own hypothetically rational choices.

Valuing in Smith's sense does not turn value judgements into introspective claims — an accusation effectively forestalled by Smith (1992 p.348) — but his notion of valuing unavoidably makes our valuing *dependent upon* claims about our own desires. Smith might argue that my reading is based upon the following mistaken presupposition about his concept of valuing: Implicit in the claim that the platitude gives the content of our thoughts about desirability lies the assumption that the platitude can be turned into a *reductive analysis* of our concept of desirability, or of a normative reason. However, according to Smith,

> contrary to the objection, the platitude does not even entail that evaluative thoughts are thoughts about our own hypothetical desires. (1992 p.349)

No one would claim that the platitude taken out of its context implies that — but how can Smith avoid that entailment when he also claims that the platitude *does give the content of our evaluative thought*? His answer is

that turning the platitude into a reductive analysis would require "a substantive account of what is required in order to be 'rational'" (1992 p.349)[8]. However, a substantive account is not within reach, he argues, and exemplifies with Williams' criterion:

A has a reason to ϕ in circumstances *C* if and only if *A* would desire that he ϕ's, in circumstances *C*, if: (1) *A* had no false beliefs, (2) *A* had all relevant true beliefs, (3) *A* deliberated correctly

Smith plausibly argues that Williams' analysis is unable to handle cases like that of the woman who thinks of drowning her baby, or the kleptomaniac. To brand them irrational, one would have to add clauses like "(4) A is in a normal emotional state." and "(5) A is in a normal physical state" (1992 p.351-2). The need for extra conditions of that kind "signals the end of a search for a reductive analysis" since any attempt to make clear when a physical or emotional state is "normal" can be guided by nothing but "our conception of what is to count as a good reason or an excuse". And this implies that the starting-point is not "the agent's actual desires, but the value-judgements he actually believes. (p 354)" So, his aims are not reductive: "What we have is, if you like, a non-reductive 'explication' of our concept of a reason." (p 352)

His point here is, I believe, crucial as well as puzzling: He seems to say that the element of valuing which his analysis does not aim to catch, the element which necessarily escapes reduction, is an irreducibly *normative* element, "our conception of what is to count as a good reason". But if our conception of a good reason necessarily is built into our notion of rationality, then how can 'what we would do if we were rational' explicate our concept of a good reason?

Smith anticipates this kind of objection in chapter 2 of *The Moral Problem*, where he characterises the distinction between reductive conceptual analysis and the kind of non-reductive explication he is executing. He argues that such non-reductive analysis might make explicit the knowledge that consitutes understanding of a concept, while, nevertheless, the analysis must make use of the concept analyzed. Granted that such explication often is a valuable tool for gaining knowledge about

our concepts, it nevertheless seems crucial to the value of the explication *which* concepts it must reintroduce. Smith's explication of 'valuing' provides us with the substantial information that judgments about value express beliefs about *objective* matters of fact (1994 p.126). In this particular case, the explicatory value of this information seems, at least, to be diminished if 'value' is reintroduced in the account of the facts in question.

Another possible interpretation of Smith's foreswearing the claim that the platitude should "entail that evaluative thoughts are thoughts about our own hypothetical desires" is this: The unavoidable gap in any analysis of rationality is a contingent fact about where we *actually* will end up if we are fully rational. This contingent fact must, then, be supposed to be impossible to detect in advance: In principle, there is no available substantive analysis, which in combination with knowledge of our present desires will allow us to deduce an answer about our (hypothetical) rational desires. This interpretation of Smith might seem farfetched, but it is supported by his mentioned defence against relativism:

> there is certainly no proof that rational agents would actually end up converging on a single set of values. But, equally, there is no proof that they would not. There is simply no way of telling in advance. We must give the justifications and see where the arguments lead. (p.355)

The first interpretation would trivialise Smith's position. The second would, I believe, merely enforce my point: It would not avoid making the desirable a function of the desired — an indetectable or unanalysable function. Thoughts about the desirable would still be thoughts about hypothetical desires under epistemically idealised conditions.

To conclude this argument: Smith's denies that the Platitude entails that evaluative thought is thought about our own hypothetical desires. It seems difficult to uphold that denial in the light of his assertion that the Platitude describes our concepts of evaluation. There is also evidence for a more normative interpretation of Smith's theory about valuing. His arguments against Frankfurt and Lewis, referred earlier, presuppose such a reading. Anyhow, it is the plausibility of the normative claim that is of interest to

me here. *Should* we attempt to form our desires after what we believe that we would desire if we were fully rational? Brandt explicitly promotes that project, and I am prepared to say that Smith implicitly recommends it.

My objection to this recommendation should be predictable by now: Forming of beliefs about what I would do if rational must be based upon knowledge of what I actually desire. Within the BD framework, there are conceptual reasons to be careful with our first person assertions about our desires. Apart from that, there is extensive empirical evidence for mistrusting them.

Let me briefly recapitulate the arguments. On the dispositional notion of desire outlined, we do not identify desires phenomenally, but via their functional roles in relation to beliefs and behaviour. Knowledge of our desires is therefore related to the possibility of making conditional predictions about behaviour. For formal reasons, we cannot strictly predict our own choices. To judge which actions follow from a certain set of beliefs, from the agent's point of view, is to *make* the decisions. Full foreknowledge disqualifies the alternative paths from being genuine options. The agent is therefore worse off than his other spectators are, when it comes to applying predictive patterns to his different behaviours.

As Schick notes, the experience of a third person privilege when it comes to foreknowledge of choices is not unfamiliar (1999 p.11). People close to us may often know how we will choose, even when we do not know that ourselves. Laboratory evidence gives us further reasons for assuming that our direct knowledge of our motivational states is very limited. Furthermore, that kind of evidence underpins the connection between functionalism about desires and the denial of direct knowledge of them. Since the BD model makes desires the primary candidates for functionalist analysis among our psychological states, there is also a strong empirical case for the BD model's implications concerning absence of self-knowledge about desires. So, do not trust your beliefs about your desires. Therefore, place only moderate confidence in your beliefs about what you would want if you were rational. Those beliefs are no authoritative beacons to steer by.

Summary of 8.5: Smith's analysis of evaluations as beliefs about hypothetically rational desires catches and illuminates one commonsensical way of thinking about valuing. As a piece of empirical semantics, it appears to be true of *some* uses.

If the rationality of acquiring or abandoning a desire is understood in terms of the coherence and unity of the available alternative sets of desires, then it seems plausible to suppose that there is a way in which evaluations in Smith's sense motivate us. Beliefs about the coherence-affecting impact of a desire appeal, necessarily, to my initial set of desires. However, if the rationality of adding or abandoning a desire is determined by other factors than coherence and unity of the available sets of desires, then the connection (between *believing* the objective fact that I would desire ϕ-ing if I was rational, and *desiring* to ϕ) becomes more obscure.

One might question, though, whether it is a wise strategy to let our beliefs about our hypothetically rational acting weigh heavily in our considerations about what to do: Our beliefs about what we would do if we were rational are necessarily dependent upon our beliefs about what we actually desire. There are formal as well as empirical reasons to mistrust the latter.

8.5 Acceptable Ends

I never make up my mind about anything at all, until it is over and done with. (Orson Welles as Michael O'Hara, *The Lady from Shanghai* 1948)

Chapter 7 acknowledged the existence and BD compatibility of several functions of deliberation. It is a fact that people sometimes desire to have certain desires blocked or triggered, and that they affect their own motivation by directing their attention and making foregrounded practical judgements. It is also a fact that motivation is not always lineal. The considerations regarded as justifying a certain action by my deliberative capacities need not be the ones that move me to intentional action. The upshot of the present and concluding chapter is a recommended precaution against overrating these deliberative practices.

Davidson's principle of continence, the authority of second order desires, and the weak forms of practical rationalism suggested by Smith and Brandt have all been rejected. Other views of internal structural rationality might be considered, and it is possible that the three criteria discussed here could be modified to meet the BD model's requirements. But tentatively, simple Humean instrumentalism about practical reason seems to be a good point of departure.

Would that mean that we are obliged to pay all of a person's aims and goals the same amount of respect? Without invoking explicit ethical considerations here, that is not what common sense says on this matter. We appreciate differences in terms of respect when it comes to other people's desires, even if none of these desires are such that we sympathise with them. The aversion to blood transfusion of a Jehovah's Witness is normally respected, for instance, while the self-starving teenager's firm resolution is not. Consider another real life case:

> A man who generally exhibits normal behavior patterns is involuntarily committed to a mental institution as the result of bizarre self-destructive behavior (pulling out an eye and cutting off a hand), which is influenced by his unorthodox religious beliefs. He is judged incompetent, despite his generally competent behavior and despite the fact that his peculiar actions follow ÒreasonablyÓ from his religious beliefs. (Beauchamp & Childress 1994 p.137)

The man's actions, self-destructive, bizarre and peculiar as they were, exhibited perfect internal lineality and self-control, on the evidence we have. There are no apparent structural features of his motivation that would explain our intuitive disrespect for his decisions. Furthermore, the damage he did to himself was less than the typical harm caused by refusing to receive blood. Yet, there is something about him that we cannot accept. When confronted with the ultimate ends another person strives for, our norms of reason appear to leave us without distinct guidelines.

The vague principles we nevertheless have in those cases will inevitably depend on our own aims and goals. The Jehovah's Witness and the self-mutilator both approach the limit for intelligibility from a normal western 2first century observer's point of view. As it happens, legal reactions

indicate that consensus at present sets that limit right between the two. Our concern for their welfare and considerations about their social roles necessarily enter our understanding of what they do.[9]

So even if it *is* fair to say that individual reason has no saying when it comes to choice of individual ends, intelligibility and the limits of interpreting the behaviour of others will necessarily be linked to our own values. In approaching questions of understanding others, we have already transcended the discourse of individual rationality, and gone into the social sphere of morality and politics. Having gone that far, we might as well admit that norms of individual practical rationality have a quite limited application as tools for improvement of human conditions.

---

[1] A good example of overexploitation in this sense is, in my view, the wide use of the term ÒautonomyÓ in health care. The present harmony of opinion about the importance of patient autonomy appears to be so great that people are unwilling to describe any justified action against a patient's will as a breach of his or hers autonomy. To avoid this, they adopt a view of what it is to respect autonomy that accommodates a variety of actions against people's expressed desires – say, via notions like Òsurrogate autonomyÓ. This view is sometimes upheld by linguistic manoeuvres that threaten to turn the principle of autonomy into a moral truism. In *Principles of Biomedical Ethics,* Beauchamp and Childress regard it as a misguided criticism of the principle of autonomy to suppose that Òemphasis on autonomy displaces or distorts other valuesÓ and subverts the authority of medicine. (1994 p.128) Although Beauchamp and Childress also acknowledge the risk of overextending the principle, any non-empty norm of autonomy in health care has to admit that patients' autonomy sometimes will clash with the authority of medical practice.

[2]Wlodek Rabinowicz made this point to me.

[3]That the brute existence of strict akrasia must be the starting point for any discussion on the matter is Davidson's view (quoted before) and e.g. E J Lemmon's (1962). Michael Smith takes a similar standpoint with respect to the Socratic escape from the akratic paradox (1994). In response to an objection of the Socratic kind (Human Action and Causation Conference in Utrecht, 1996) Smith remarked that this way out simply reflects philosophical laziness; unwillingness to perform the conceptual tinkering necessary to accommodate intentional acting against better judgement. This may be so, but a possible diagnosis of the opposite tendency could be philosophical imperialism — unwillingness to leave interesting areas of human activity for the social sciences to

analyse. (The paper Smith gave at the conference was ÒThe Possibility of Philosophy of ActionÓ, published 1998)

[4] Watson's own suggestion about the use of 'evaluation' is that the term refers to Òprinciples and ends which [the agent] — in a cool and non-deceptive moment — articulates as definitive of the good, fulfilling, and defensible life.Ó (Watson 1975) D.F. Pears supposes, similarly, that terms like 'evaluation' and 'value-judgement' in ordinary language often reveal an implicit ranking of the importance of different kinds of desires. Sometimes value-judgements are thought of as expressing "a special kind of preference, based on one's long term interests or perhaps other people's interests" (1984 ch.2), as opposed to *any* kind of preference. I see no reason to deny that such divisions are common in ordinary non-technical English. The distinction can, however, be interpreted in two directions; from the agent's or the observer's perspective. The first interpretation will collapse into a second order account. If read as a norm, it will therefore run into the same difficulties as Frankfurt's, Lewis' and Dworkin's theories. The second way of understanding it will be to give up content neutrality .

[5] A somewhat lengthy example to underpin the linguistic intuition that Smith's ÒdesiringÓ as Òbelieving what one would desire if rationalÓ really catches *one* usage: Two of our daughters, when at the age of twelve and thirteen, sometimes discuss music. The older one spends most of her allowance on CD:s, she reads every review of popular music in our daily paper, she sometimes buy rock magazines and she listens to records a lot. The younger one has other interests and she spends neither much money nor time on music. When she makes choices about what to listen to, she usually displays a more conventional taste than her older sister. Despite their different choice behavior, they seem to agree verbally on how to rank different rock and pop groups. As far as I can judge, their common ranking answers to the choices actually made by the older one, but not to the choices made by the younger one. There is no doubt that our thirteen year old daughter always would choose a record with *Q* before one with *P* while our younger daughter might put on a *P* record but hardly one with *Q*. Nevertheless she does not question the older one's opinion that *Q* is far better than *P*.

I rule out the possibility that she is hypocritical — then, what does she mean when she agrees about the best group? Does she act against her important (stable, reflected etc.) desires? ('Valuing as a mode of desiring'.) Such an interpretation would neither be inconsistent, nor incompatible with a common way of speaking. But is it probable that this is what she had in mind? I do not think so. Her expressed opinion is neither short-term, unreflected or formed under some kind of emotional pressure. Since she explicitly insists that her positive valuing of *Q* comes without any desire for *Q*, an interpretation in terms of valuing as *a way of* desiring would in this case be just as

ungenerous as an interpretation in terms of valuing as desiring. Does she act upon a desire she desires not to have? (‘Valuing as desiring to desire’.) Probably not. She does not seem to desire that she should have a desire to listen to the best kind of music. In fact, she does not bother about her desires when it comes to music; she simply acts upon them. Does she act upon a desire she believes that she would not have if she was fully rational? (‘Valuing as believing’ in Smith’s sense.) It seems to me that the most intuitively reasonable interpretation of her statements and her behavior is that, though she actually desires *P* before *Q*, she is also convinced that if she had as much confrontation with *P* and *Q* (and other relevant facts) as her older sister, then she would prefer *Q* to *P*. So, I believe that Smith’s analysis (like Frankfurt’s and Davidson’s) truthfully catches *one* common way of using value-terms. In that matter I can, however, only appeal to linguistical intuitions.

[6] ÓGiven the *goal* of having a systematically justifiable set of desires, it may well turn out that, as the attempt at systematic justification proceeds, certain desires that seemed otherwise unattainable have to be given up.Ó (1995 p.115, my emph.)

[7]Persson points to other difficulties that interpretation would get Smith into.

[8] In e-mail conversation (1996), Smith made clear that he would no longer say what he said in (1992) — that evaluative thoughts “are not about our hypothetical desires.” They *are* about hypothetical desires. The point he would want to make is that the judgements and inferences we make (which show that valuing is a matter of what would be desired under certain conditions) are such that we cannot hope to spell them out fully without reintroducing concepts of reason or of value. In other words, that explication must be non-reductive in this case.

[9] A popular doubt about the possibility of altruism within the BD model should be met here. It may be expressed like this: Every intentional act is the result of the agent's strongest desires (wants, preferences, evaluations etc.) at the moment of acting. In that sense, people always necessarily attempt to maximise their own expected utility. A genuinely altruistic action, resulting from benevolent motivation, would be to be good to others without any expected personal gains, or even at foreseen losses. Therefore altruism is impossible.

In some versions the argument is reversed. For instance, Alasdair MacIntyre finds it necessary to deny that "if I do something, it is thereby true that I want to do it" in order to establish the possibility of altruism. ("Altruism and Egoism" in the Encyclopaedia of Philosophy)

A standard answer to both versions is that they are founded on a too narrow conception of altruism. An altruistic action, it is said, aims at the good of other people, for instance by satisfying their preferences or making them happy. Even if we necessarily always do what we want most when we do it, there is no logical oddity in wanting other people's preferences to be satisfied or preferring them to be happy. Thus, the altruistic action maximises the agent's expected utility *by* satisfying the agent's benevolent desires. Genuine benevolence is no anomaly within the BD model.

Let me add another small digression concerning benevolent BD motivation here. Rabinowicz and Österberg present an objection to the satisfaction-interpretation of utilitarianism which might be relevant for benevolent motivation in general. As an extension of an argument by Butler, they note that "preference-satisfaction is only possible as a secondary aim; if we would have, as our only preference, the desire to have our preferences satisfied, we would have no preferences to satisfy".

Butler's original argument is directed against the motivational possibility of egoistic hedonism. If someone bothered about nothing but *his own* satisfaction, his motivation would, in Butler's words, have "absolutely nothing to employ itself about" (1996 p.213) Rabinowicz and Österberg claim that it would work against universalistic versions as well.

Butler's argument does not, however, thereby exclude the actual possibility of a well-informed beneevolent agent who is motivated so that he regards desire-satisfaction as the sole intrinsic value. He is not *particularly* concerned with his own desire-satisfaction, although it may count. He simply finds it valuable that everyone get what they want — that is the ultimate goal for all his acting. When they do, he gets what he wants as well.

This satisfaction-oriented agent would only get into Butler's difficulties if everyone else was motivated like him. His attitude will be collectively self-defeating. Actual cases of attempted mutual benevolence sometimes displays that trait. (E.g. when two people tries to pick a restaurant for their lunch and both of them are more anxious about satisfying the other's preferences than about getting to a place that suits their personal taste.) The view that there are satisfaction-oriented benevolent people has, then, some empirical support.

In their discussion of Butler, Rabinowicz and Österberg also suggest an argument that, as I understand it, is slightly different:

It "seems that an axiological view according to which nothing but preference satisfaction has intrinsic value deprives our intrinsic preferences of their authority by denying their objects intrinsic value, while at the same time it feeds on these denigrated preferences".(p.7)

But even if y is of no value and x is desirable on its own account, why should the fact that y is necessarily tied to x relieve x of value? If this fact really empties x of value, the direction of the argument seems arbitrary: Will it not, then, be just as implausible to see objects as intrinsically valuable *because* they happen to be desired by someone, and at the same time be indifferent to satisfaction of desires? The crucial term in this objection is “because”. If this means that the value of the object is *constituted* by its being desired by someone — rather than e.g. *supervening* upon that desire — this might be a less forceful objection.

The following argument against desire satisfaction as the probable intrinsic value of actual well-informed benevolent agents' axiologies is more convincing, I think: In connection with their requirement of "motive internalism", Rabinowicz and Österberg point out that we need not have second order desires for the satisfaction of our desires. According to the BD model, our desires need not figure in the content of our deliberation at all. We do not necessarily or even normally value satisfaction of our own desires independently of their content.

Now, if we normally do not value our own desire satisfaction for its own sake, how do we become motivated to achieve desire satisfaction (independently of content and further effects) for others? One improbable but perhaps possible psychological explanation: Imitation of (or purposive conditioning performed by) preference-utilitarian parents, teachers or other authorities. Besides being improbable, any desire created by such a process (from the attitudes of other people rather than from a natural connection between the desired object and the desire) would be a prime candidate for extinction by what Brandt calls "cognitive psychoterapy" (1979 p.117).

Prima facie, other people’s happiness or some other preference independent goods appear to be more likely as objects of benevolent agents’ motivation, in the light of the BD model.

Adams, R M
"Involuntary Sins", The Philosophical Review XCIV, No 1 1985*
Anscombe, G E M
Intention, Oxford 1957
"Thought and Action in Plato and Aristotle" in *New Essays on Plato and Aristotle*, ed. R Bambrough London 1965
Alston, W P
"Searle on Illocutionary Acts" in LePore/Van Gulick 1991
"Conceptual Prolegomena to a Psychological Theory of Intention" in Brown (ed.) *Philosophy of Psychology*, Barnes & Noble, New York 1974
Árdal, P S
Introduction to Hume's Treatise (Hume 1739) Fontana, London 1972
Arendt, H XXX
Armstrong, D M
A Materialist Theory of the Mind, Routledge & Kegan Paul, London 1968
"Dispositions as Categorical States" in Crane 1996
"Place's and Armstrong's Views Compared and Contrasted" in Crane 1996
Audi, R
"The Concept of Wanting", Philosophical Studies 24 1973
"Deliberative Intentions and Willingness to Act: A Reply to Professor Mele" *Philosophia*, 18, pp.243-245, 1988
Action, Intention and Reason, Cornell U.P., Ithaca N.Y. 1993
Aulisio, M P
"In Defense of the Intention/Foresight Distinction" *American Philosophical Quarterly*, Vol.32, No 4, Oct 1995
Baier, A C
"Rhyme and Reason: Reflections on Davidson's Version of Having Reasons" in Lepore/McLaughlin 1986
Beauchamp, T L, & Childress, J F
Principles of Biomedical Ethics, 4th ed. Oxford U P, Oxford and N.Y. 1994
Beauchamp, T L & Rosenberg, A
Hume and the Problem of Causation Oxford U P New York 1981
Bigelow, J, Dodds, S M & Pargetter, R
"Temptation and the Will", *American Philosophical Quarterly*, Vol.27, No.1, January 1990
Björnsson, G
Moral Internalism., An Essay in Moral Psychology, Stockholm University 1998
Blackburn, S

”Losing Your Mind: Psychology, Physics and Folk Burglar Prevention” in Greenwood, J, (ed.) *The Future of Folk Psychology,* Cambridge UP, Cambridge 1991

Bond, E J

Reason and Value, Cambridge U P, Cambridge U K 1983

Borges, J L

“The Library of Babel” in *The Garden of Forking Paths,* transl. Anthony Kerrigan (in Borges, *Fictions* ed. A Kerrigan, John Calder, London 1985) 1941

Brandt, R

A Theory of the Good and the Right Oxford U P, Oxford 1979

Bransen, J & Cuypers, S.E. (eds.)

Human Action, Deliberation, and Causation. Kluwer, Dordrecht, Boston 1998.

Bratman, M

Intention, Plans and Practical Reason, Harvard U P, Cambridge Mass. 1987

“Practical Reasoning and Acceptance in a Context“ XXX 1992

“Toxin, Temptation and the Stability of Intention” in Coleman/Morris, *Rational Commitment and Social Justice. Essays for Gregory Kavka,* Cambridge U P 1998

Brink, D O

“XXXX” Ethics 108:1 Oct.(1997)

Cambell, J & Pargetter, R

“Goodness and Fragility”, American Philosophical Quarterly, 23:2, April 1986

Champlin, T S

“Tendencies” XXXXX 1991

Champlin T S, & Walker A D M

“Tendencies, Frequencies and Classical Utilitarianism“ *Analysis*, vol 35, no 1, 1974

Charles, D

Aristotle's Philosophy of Action, London 1983

"Rationality and Irrationality", Proceedings from the Aristotelian Society 1983

Charlton, W

Weakness of Will, Blackwell, Oxford 1988*

Chisholm, R.M

"Freedom and Action" in *Freedom and Determinism*, ed. Keith Lehrer, New York 1966

Crane, T

(ed.) *Dispositions, A Debate,* Routledge, London & N.Y. 1996

“The Efficacy of Content: a Functionalist Theory”, in Bransen & Cuypers 1998

Davidson, D

“Agency” (in Davidson 1980) 1971

“First Person Authority”.*Dialectica* 38, 101-112. (1984)

”Actions, Reasons, and Causes’ rep. in *Essays on Actions and Events*
Essays on Actions and Events, Clarendon, Oxford, and Oxford UP, New York 1980 London 1980
"Deception and Division" (in Lepore/McLaughlin 1986)
"How is Weakness of the Will Possible" (in Davidson 1980) 1970
"Intending" (in Davidson 1980) 1963
"Paradoxes of Irrationality" (in Wollheim/Hopkins (eds.) *Philosophical Essays on Freud*, Cambridge 1982)
"Mental Events" (in Davidson 1980) 1970
Replies to Grice & Baker, Peacocke and Pears in Vermazen/Hintikka 1985
"Thought and Talk" in Guttenplan 1975*
Dewey Lectures Journal of Philosophy XXX1990

Dennett, Daniel C.
Brainstorms, Harvester Press, Hassocks, Sussex 1979
Consciousness Explained, Allen Lane, The Penguin Press 1991a
“Real Patterns” *The Journal of Philosophy*, XXX 1991b
“How to Study Human Consciousness Empirically, or Nothing Comes to Mind”, *Synthese*, 59 pp. 159-180, 1982

Dretske, Fred
“Mind, Machines and Money: What Really Explains Behaviour” in Bransen & Cuypers 1998
Naturalizing the Mind, MIT Press, Cambridge, Mass. & London, U K 1975
Explaining Behaviour, MIT Press, Cambridge, Mass. & London, U K 1984

Dworkin, G
The Theory and Practice of Autonomy, Cambridge U P, Cambridge U K 1988
“The Concept of Autonomy”, in Christman, J (ed.) *The Inner Citadel: Essays on Individual Autonomy* Oxford U P (1989) 1981*Uppsats XXXX*

Egonsson, Dan,
Interests, Utilitarianism and Moral Standing, Lund University Press, Lund 1990

Elster, J,
Ulysses and the Sirens, rev.ed. Cambridge 1984

Elster, J & Hylland, A eds.
Foundations of Social Choice Theory, Cambridge U P, Cambridge & Oslo 1986

Flew, A
David Hume, Philosopher of Moral Science, Blackwell, Oxford 1986

Fodor, J A
Representations, Philosophical Essays on the Foundations of Cognitive Science, Harvester Press, Sussex U K, 1981

Foot, P
Virtues and Vices, Oxford 1979
Frankfurt, H G,
"Freedom of the Will and the Concept of a Person", The Journal of Philosophy, volume LXVIII, No 1 1971
Frey, R.G.
Interests and Rights: The Case Against Animals, Clarendon Press, Oxford 1980
Fröström, L
Omdöme och proposition, Studentlitteratur, Lund 1983
Goldhagen, D
Goldman, A,
A Theory of Human Action, Princeton U.P., New Jersey & Surrey 1970
Gosling, J
Weakness of the Will, Routledge, London 1990
Gowans, Christopher W (ed.),
Moral Dilemmas, Oxford 1987*
Gibbard, A
Wise Choices, Apt Feelings, Clarendon Press, Oxford 1990
Gjelsvik, O
"The Epistemology of Decision-Making 'Naturalised'", in Orenstein & Potatko (eds.) *Knowledge, Language and Logic,* pp.109-129 Kluwer, U K 2000
"Freedom of the Will and Addiction", in XXXX 2000
Goldman, A
"The Psychology of Folk Psychology" *Behavioral and Brain Sciences* 16 1993
A Theory of Human Action, Princeton U P, Princeton N.J. 1970
Gopnik, A
"How we know our minds: The Illusion of First-person Knowledge of Intentionality", *Behavioral and Brain Sciences* 16 1993
Grice/Baker,
"Davidson on 'Weakness of the Will'" (in Vermazen/Hintikka 1985)
Guttenplan, Samuel
(ed.) *Mind and Language,* Clarendon Press, Oxford 1975*
(ed.) *A Companion to the Philosophy of Mind,* Blackwell, Oxford U.K., Cambridge Mass. (1994)
Gärdenfors/Sahlin (eds.)
Decision, Probability and Utility, Cambridge U P, Cambridge 1988
Hamlyn, D W
A History of Western Philosophy, Penguin, London 1987

Hampshire, S,
Freedom of the Individual, exp. ed. N J 1975
Hansson, B,
"Risk Aversion as a Problem of Conjoint Measurement" in Gärdenfors/Sahlin (eds.) *Decision, Probability and Utility* 1988
Hausman, D
"Revealed Preference, Belief and Game Theory" unpublished draft 1999
Hare, R M.
Moral Thinking Clarendon Press, Oxford 1981
Freedom and Reason, Oxford UP, Oxford 1963
"Wanting: Some Pitfalls" in *Practical Inferences*, Macmillan, London (1971) 1968
Harman, G, & Thomson, J J
MoralRelativism and Moral Objectivity, Blackwell, Cambridge, Mass. & Oxford, G.B. 1996
Harris, P L
"First-person current" (Gopnik/Goldman commentary) *Behavioral and Brain Sciences* 16 1993
Heil, J & Mele, A (eds.)
Mental Causation, Clarendon Press, Oxford 1995
Harrison, J
Hume's Moral Epistemology, Oxford, Clarendon Press 1976
Hornsby, J
"Agency and Causal Explanation" in Heil & Mele 1995
Hume, D.
A Treatise of Human Nature, (Fontana, London 1982) 1739
Dialogues Concerning Natural Religion, (Hafner Press, Macmillan, New York 1948) 1779
Hurley, Paul
"How Weakness of the Will is Possible", *Mind*, Vol.101, January 1982
Jackson, F & Pettit, P
"Some Content is Narrow", in Heil & Mele 1995
Jiborn, M
"Utility functions in Moral Reasoning: the Problem of Interpersonal Comparison", in W. Rabinowicz (ed.) *Preference and Value* 1996
Josefsson, J
"No regrets: A note on a no-regret condition" unpublished draft, Dep. of philosophy, Lund 1999.
Kenny, A

Action, Emotion and Will, Cambridge U.P. 1963*

Lemmon, E J,

"Moral dilemmas" *The Philosophical Review* 1962

LePore/VanGulick (eds.)

John Searle and his Critics, Blackwell, Oxford U.K., Cambridge, Mass. 1991

Lepore/McLaughlin (eds.),

Actions and Events: Perspectives on the Philosophy of Donald Davidson, XXXOxford 1986

Lewis, D

"Radical Interpretation", in *Philosophical Papers*, vol 1, Oxford U P, Oxford 1983

"Dispositional Theories of Value" *Proceedings of the Aristotelian Society*, suppl. vol. 1989

MacIntyre, A

"Altruism and Egoism" (*The Encyclopaedia of Philosophy*) XXX*

Mackie, J L,

Hume's Moral Theory Routledge & Kegan Paul, London 1980

Melden, A.I

Free Action, Routledge & Kegan Paul 1961

Mele, A R,

Autonomous Agents; From Self-Control to Autonomy, Oxford U P, N.Y., Oxford 1995

"Against a Belief/Desire Analysis of Intention" *Philosophia*, 18, 1988

Springs of Action, Understanding Intentional Behaviour, Oxford U.P., Oxford, New York 1992

Irrationality, An essay on Akrasia, Self-Deception, and Self-Control, Oxford UP Oxford, N.Y., 1987

Mellor, H

"In Defence of Dispositions" (in *Matters of Metaphysics* Cambridge UP 1991) 1974

Causation XXX

Meyer, L B

Emotion and Meaning in Music, University of Chicago Press, Chicago 1956

Mischel, W

Personality and Assessment, John Wiley & Sons, New York 1968

Nagel, E

The Structure of Science: problems in the logic of scientific explanation, Routledge & Kegan Paul London 1961

Nagel, T

The Possibility of Altruism, Clarendon Press Oxford 1970

The View from Nowhere, Oxford U.P. 1986

Nisbett, R E & Wilson, T DeCamp
"Telling More Than We Can Know: Verbal Reports on Mental Processes" *Psychological Review*Vol.84, 1977
O'Shaughnessy, B
"Searle's Theory of Action" in LePore/VanGulick 1991
Parfit, D
Reasons and Persons, Oxford U.P. 1983
Peacocke, C
"Intention and Akrasia" in Vermazen/Hintikka 1985
Pears, D
Motivated Irrationality, Oxford 1984
"A Sketch for a Causal Theory of Wanting and Doing" in *Questions in the Philosophy of Mind*, London 1975
XXX in Lepore/Mc Laughlin
Persson, I
Reasons and Reason-Governed Actions, Lund 1980
"Are Moral Requirements Categorical Requirements on Rationality?" *Theoria* (forthcoming issue) 1995
"Hume--Not a "Humean" about Motivation", *Hist Phil Quart* 14(2), 189-206 April 1997
The Retreat of Reason — A Dilemma in the Philosophy of Life, Draft, Lund University 1992
Pettit, P & Smith M,
"Backgrounding Desire", *The Philosophical Review*, Vol.XCIX, No.4 October 1990
Place, U T,
"Dispositions as Intentional States" in Crane 1996
"A Conceptualist Ontology" in Crane 1996
Platts, M
Ways of Meaning, Routledge and Kegan Paul XXX 1979
Prior, E W, Pargetter, R, & Jackson, F
"Three Theses about Dispositions" American Philosophical Quarterly vol.19, No 3, 1982
Quine, V W
Word and Object, MIT Press, Cambridge, Massachusetts 1960
Pursuit of Truth, Harvard U.P., 1992
Rabinowicz, W (ed.)
Preference and Value; Preferentialism in Ethics, *Studies in PhilosophyXXX*, Lund University 1996

Rabinowicz, W, & Österberg, J
"On Two Interpretations of Preference Utilitarianism" (Draft 1993)XXX
Regan, Tom,
The Case for Animal Rights, Routledge & Kegan Paul, London 1984
"Frey on Why Animals Cannot Have Simple Desires" *Mind*, Vol.XCI 1982
Rorty, Amélie Oksenberg,
"The Social and Political Sources of Akrasia" *Ethics* 107 (July) 1997
"Akrasia and Conflict" *Inquiry* 22, p.193-212, 1980a
"Self-Deception, Akrasia and Irrationality", *Social Science Information* 19, 1980b
Ryle, G,
The Concept of Mind, Hutchinson, London 1949
Sacks, O W
The Man Who Mistook his Wife for a Hat and other Clinical Tales
Simon & Schuster New York 1998
Sahlin, Nils-Eric
"Baconian Inductivism in Research on Human Decision-making", *Theory & Psychology*, Sage Vol. 1(4) 1991
"The significance of empirical evidence for developments in the foundations of decision theory" in *Theory and Experiment*, ed. by Batens, D. & van Bendegem, J. P., D. pp. 103-121 Reidel, Dordrecht, 1988
Sartre, J-P,
Being and Nothingness, (*L'Etre et le Néant* 1943), transl. Hazel E. Barnes, Methuen & Co. London 1958
Scanlon, T M,
What We Owe to Each Other, Belknap, Harvard U.P. Cambridge, Mass. & London, U.K. 1998
Schick, F
Understanding Action, An Essay on Reasons, Cambridge U P, Cambridge U K, New York 1991
"Surprise, Self-Knowledge and Commonality" i PG:s festskrift XXX 1999
Schueler, G F
"Pro-attitudes and Directions of Fit" *Mind* vol.100, 398, April 1991
Desire: its role in practical reason and the explanation of action, MIT Press, "A Bradford book" Cambridge Mass., London U.K. 1995
Searle, J,
"Meaning, Intentionality, and Speech Acts" in LePore & VanGulick 1991
"Response" (to O'Shaughnessy) in LePore & VanGulick 1991
Intentionality, Cambridge U. P. 1984

”A Taxonomy of Illocutionary Acts”, in *Language, Mind and Knowledge,* Univ. of Minnesota Press, Minneapolis 1975.

Expression and Meaning, Cambridge U P, Cambridge U K 1979

Sen, A

“Behaviour and The Concept of a Preference” *Economica,* August 1973

Sereny, G

Albert Speer: His Battle with Truth, Alfred Knopf inc. (Swedish transl. Bonniers 1997) 1995

Shaffer, Jerome A

"Sexual Desire", *The Journal of Philosophy,* Vol.LXXV, No.4 April 1978

Smith, Michael,

"The Humean Theory of Motivation", *Mind*, Vol.96, January 1987

"Reason and Desire", *Proceedings from the Aristotelian Society* 1988

“Valuing: Desiring or Believing” in Charles & Lennon, eds. *Reduction, Explanation, Realism,* Oxford U.P.1992

The Moral Problem Basil Blackwell, Oxford 1994

“Internal Reasons“ *Philosophy and Phenomenological Research* Vol.LV No 1, March 1995

“The Possibility of Philosophy of Action” in Bransen & Cuypers 1998

Sosa, E

“Abilities, Concepts and Externalism” in Heil & Mele 1995

Stalnaker, R C

Inquiry, MIT Press, Cambridge, Mass. & London, England, 1984

Stampe, D W

“Desire” in Guttenplan 1994

Strawson, G

The Secret Connexion: Causation, Realism, and David Hume. Clarendon Press Oxford & New York 1989

Stroud, B

Hume, Routledge & Kegan Paul, Suffolk 1977

Swinburne, R,

The Evolution of the Soul, Oxford 1986

"Desire" in XXX, XXX 1985

Taylor, R

Good and Evil, New York 1970

Thalberg, I

Enigmas of Agency, Allen & Unwin 1972

Thomas, G

An Introduction to Ethics, *Duckworth, London 1993*

Van Roojen, M

"Humean Motivation and Humean Rationality", *Philosophical Studies,* 79: 37-57, 1995

Velleman, J D

"Brandt's Definition of Good" *The Philosophical Review* Vol.XCVII, No 3, July 1988

Vermazen/Hintikka (eds.),

Essays on Davidson: Actions and Events XXX1985

Watson, G

"Free Agency", *Journal of Philosophy* Vol.LXXII, No 8, April 1975

Williams, B

Moral Luck, Cambridge 1981

Wilson, J R S

Emotion and Object, Cambridge UP 1972