Self-Knowledge, Rationality and Interpretation

by

Brett Caloia

B.A., UCLA, 2002

Submitted to the Graduate Faculty of the Kenneth P.

Dietrich School of Arts and Sciences in partial fulfillment

of the requirements for the degree of

Doctor of Philosophy

University of Pittsburgh

2011

UNIVERSITY OF PITTSBURGH

Dietrich School of Arts and Sciences

This dissertation was presented

by

Brett Caloia

It was defended on

August 22nd, 2011

and approved by

Edouard Machery, Associate Professor, HPS

Karl Schafer, Assistant Professor, Philosophy

Kieran Setiya, Associate Professor, Philosophy

Dissertation Co-Advisor: Michael Thompson, Professor, Philosophy

Dissertation Co-Advisor: Peter Machamer, Professor, HPS

Self-Knowledge, Rationality and Interpretation

Brett Caloia, PhD

University of Pittsburgh, 2011

A central concern in the philosophy of mind for the past half-century has been *interpretation*: what mental states should I attribute to someone else? Quine argued that providing a translation of an alien language required seeing that language as logically structured. Davidson and Lewis took this idea further. They argued that the project of providing a translation was part of a larger project of providing an interpretation of the subject. To interpret was to attribute mental states that made the subject's behavior rational. Thus they replaced the injunction to see the subject's language as conforming to logical laws with a broader principle of charity. The principle of charity constrains the activity of interpretation by the untenable assumption that the subject is rational.

I propose replacing charity's injunction to maximize rationality with a principle that directs an interpreter to minimize inexplicable behavior. The positive argument for this new principle emerges from two sources. The first is *empirical*: there is a great deal of evidence that human beings are simply not all that rational. Moreover, their irrationality is predictable and operates in fairly well understood ways. The second is *first-personal*: each of us is aware of a variety of irrational tendencies in our own thought. These sources can be drawn on to make sense of behavior without offering a rational reconstruction. I understand why my frustrated colleague yells at his computer, in part, because I know what it means to be frustrated.

I argue that taking a first-personal account of the subject seriously will mean seeing that the subject might consciously make transitions in thought that are not beholden to a rational ideal. The interpreter may use his own first-personal experience as a model for understanding the subject. This expands the evidential base beyond the observational. Doing this makes it is possible to recognize something as thought without seeing it as held in place by the rational ideal of the network.

TABLE OF CONTENTS

PREFACE

This book is dedicated first and foremost to my mother, Patricia Caloia. Throughout my life she has always sought to show me the value of education and thoughtful engagement with the world. Her unwavering faith and encouragement has sustained me and it is clear to me that my achievements would not have been possible without her. I hope one day to have made as profound an influence on someone's life as she has on mine.

I also wish to acknowledge my wife and best-friend, Ashley Caloia. She is truly the light of my life and a constant source of joy for me.

My very good friend Jeff Chwieroth is also deserving of thanks. He helped me to recognize and pursue my potential and has always been a true friend.

The members of my committee have helped me immeasurably. I thank all of those – especially Peter Machamer- who diligently and frequently read drafts, supplied comments, and generally supported me through this effort.

Finally, a big thank you to all of my very supportive friends and colleagues from the University of Pittsburgh. People like Tim Willenken, Preston Stovall, Stephen C. Makin, Endre Begby, Holly Andersen, Chris Frey, Tyke Nunez, Dan Addison, Kathryn Lindeman and others make the Pittsburgh Philosophy Department a wonderful place to work and study.

1.0 WHICH QUESTIONS CAN AN INTERPRETATION ANSWER?

Interpretation is a slippery term. It can refer both to the product of a certain kind of inquiry and the activity of inquiring. Moreover, it is connected to other terms which have a similarly slippery character. For instance, it appears uncontroversial that offering an interpretation is to offer an analysis of a subject's meaning, but this can't help unless a determinate understanding of meaning can be supplied and this is quite difficult to do. At the outset then it seems prudent to try and demarcate the subject matter that I intend to treat in this discussion of interpretation. The authors I will engage with treat paradigm cases of interpretation as showing that attributing a reason recognizing mental state to a subject allows the subject to be functionally described. That is, such attributions allow one to see the subject as performing some action or other for the sake of something else. This will be my starting point. However, I want to leverage this understanding to argue that explanations which cite motivation, but which may not be reason recognizing, are also properly interpretive. So while I begin with a fairly narrow understanding of interpretation, it is my hope that this essay will begin to push on the boundaries of this conception so as to enlarge it. But even if I am successful at this, a reader might point out that I stop short of a full account. The area that I mean to enlarge the conception into is vast. There are a great many motivations out there and I don't pretend to have the resources to offer their full taxonomy. If I am able to convince my reader that interpretation in this broader sense is explanatory because it makes sense of an agent's motivations, then this will be enough for my

point. Subsequent chapters will take up some of the more specific details of how motivations operate.

It is also helpful to note that one of the chief uses of the term 'interpretation' is to mark out a highly subjective activity in which someone offers a personal understanding of the subject. This sort of thing comes in degrees. At one extreme, perhaps, is the rather hapless undergraduate who offers a short story about his car troubles in response to an essay which asks him to 'interpret' *Euthyphro*. At the other is, perhaps, a market analyst drawing on expertise and years of past experience to try and explain the puzzling moves of certain economic indicators for the week. It is this usage which seems to me to have the most breadth and it is this which presents a potential challenge to my starting point- 'just how can we account for that which is common to the undergraduate and the market analyst?'. I want to try and side-step this question by focusing on a fairly determinate conception of what interpretation is. So what I aim to produce is an analysis of a specific kind of explanation which shows that less-noticed but nevertheless kindred explanations have a claim to be treated as a kind of interpretation. Because this analysis is so general it leaves little room for considering the particular and unique contributions that individuals make to questions of interpretation. However, I have striven to note those areas where individual contributions are properly a component of this kind of explanation and I hope that future work can expand on these to provide a fuller account of the role of interpreter in interpretation.

1.1 DENNETT'S VIEW

To offer an interpretation is to offer a specific kind of explanation. It is an explanation that attempts to provide the meaning for the explanans. Understood this way then an interpretation can only apply to those things which are meaningful. Moreover, not just any explanation offered with respect to a meaningful thing can count as an interpretation.

Dan Dennett has argued that offering an interpretation is to provide an answer to a specific form of the question 'why?' which presents the subject as being for the sake of some purpose. (Dennett, Interpretation 7) Further, the subject's coming to serve that purpose must be the result of a recognition that some reason exits for the subject to perform in this way. (Dennett, Interpretation 7) An interpretation then presents the subject as purposeful in response to reasons. This understanding of interpretation drives an important consequence for this theory. It makes clear why the invocation of an intention in explanation is to provide an interpretation. Generally[1] an intention places the current action in the context of a broader functional explanation that aims at a outcome. Thus the current action gets its functional description by being part of some efforts toward some future outcome. Further, recognition of an end as an end[2] means seeing it as in need of justification because to conceive of something as an end is to see it as worthy of pursuit for some reason. This, in turn, requires that the intentional explanation which determines the subject's functional role will be referencing a justification for having that end in the first place. Thus, for Dennett, the question 'why?' is so distinctive and has such a central role in driving

[1] I take this to be a paradigmatic case of intentional explanation. However, as I intend to show, there are kindred forms of motivation which, when cited, can also provide interpretive explanation. It should also be noted that some actions are intentional without there being any *further intention* with which they are done. I take it that these don't present any special problems for Dennett's analysis since this doesn't preclude them from being functional in the way he specifies.

[2] Or, a reason as a reason.

interpretation because answering it presents a functional description that references reasons recognized as reasons for the subject's performance. Indeed, one might treat his distinctive 'why?' question as asking 'what is it for the sake of?'. (Dennett, Interpretation 8)

This analysis has a number of notable features. First, it provides little support for first-person privilege. While, at times, the author of a behavior or artifact may have intended the object to have just the functional role or rational justification that would make for the most plausible interpretation, generally this need not be true and it will be possible to treat the author's own answer to the 'why?' question as providing just one more data-point relevant to an interpretation. (Dennett, Interpretation 3) Second, Dennett does not limit interpretation's relevance to only human action. This is because he thinks that other natural functional systems can contain mechanisms for taking account of reasons. For instance, he takes the description of evolutionary processes as natural *selection* to be aptly named since this process is aiming at promoting survival by recognizing the contribution to reproduction that certain mutations make. (Dennett, Interpretation 9) In other words, it is *selecting* for those features in order to promote survival. The notable feature of his theory that I will be focusing on however concerns the need for idealization to drive inquiries into interpretation. Dennett argues that because interpretation of a subject requires attributing reason recognition to the subject, this must be assumed to be optimal. Dennett claims we use "an assumption of rationality or cognitive/conative optimality to structure our interpretation" and later that "the presumption is that there is a (good) reason.". (Dennett, Interpretation 9) So for Dennett, the interpretation of a thinking subject's behavior will begin with an assumption of rational optimality. In short, the interpretation proceeds by assuming that there is a good reason for all the subject does.

On this analysis it might be thought that Dennett neatly divides the world into two. One the one hand there are things that are uninterpretable: namely those things that are not for the sake of anything else. On the other are those objects that do admit of interpretation and so any explanation of them must be structured by the optimality assumptions built into the attribution of rationality.

However, there is reason to think that Dennett's first cut cannot be made so cleanly. Anscombe's analysis of intention begins with an analysis of action. She notes that while the paradigm case of action explanation is provided by citing an intention, the distinctive tone of her 'why?' question is not rejected by a person claiming that there was no particular reason or further intention that lay behind the action, though this is not to say that action was unintentional.(Anscombe 25) In the same way that answering 'none' when queried 'how much money do you have in your pockets?' is not to reject the question, answering 'no particular reason' is not to reject the applicability of the question 'why?'.(Anscombe 25) Further, given that both Anscombe and Dennett are after an intention in asking the questions that animate their inquiry, it seems safe to say that one way of understanding the distinctive register of Anscombe's 'why?' question is to hear it as expressing the same query as Dennett's: an inquiry that asks 'what is it for the sake of?'. So if both Anscombe and Dennett can be understood to be making the same form of inquiry, it is prudent to inquire into the nature of intentions since these will help illuminate the activity of interpretation. I will argue that using Anscombe's analysis of intention as a starting point will lead the way to seeing interpretation as an activity that is larger than supplying only intentions as explanations of meaning.

In Section 1.2 I will first use the work of Dennett and Anscombe to show what is involved in providing an interpretation in the clearest case. It will be seen that certain questions

become pressing in offering an interpretation. These help to demarcate interpretation from other forms of explanation as well as show the possibility of non-paradigmatic of interpretation. Examination of these non-paradigmatic cases illustrates the explanatory force of offering accounts of a subject's motivations as interpretations. In Section 1.3 I will argue that explanations of meaning are provided by pointing to motivating personal level mental states that do not qualify as intentions. It is on this basis that these other states can be seen as forms of interpretation. In Section 1.4 I take up a challenge provided by the possibility of certain 'merely causal' explanations of action. I argue that these do not have a claim to be interpretive because they do not provide an explanation of meaning.

The divisions of this paper owe something to the taxonomy of possible answers to the 'why?' question that Anscombe develops in *Intention*. (Anscombe 25, 38) I will first deal with the clearest case of intentional explanation: namely those cases in which an intention toward a further end is to be offered as an answer to the 'why?' question. Then I will move on to a set of cases which I argue is still interpretive. These will be cases in which the 'why?' question is answered in the negative without rejecting the applicability of the question. Finally I will turn to answers to the 'why?' question which reject the applicability of the why question. I argue that this is to offer a form of explanation which is distinct from interpretive explanation. However this raises a potential problem because it might be thought that this distinct style of explanation is capable of superseding the need for interpretive explanation at all. I show that this concern is unfounded and so conclude that interpretive explanation is a response to a distinct form of inquiry. The form that this inquiry must take to supply its distinctive form of explanation is given through an examination of how it proceeds in each case.

1.2 INTERPRETATION AS ANSWERING THE QUESTION "WHAT IS IT FOR THE SAKE OF?"

In asking the question 'what it for the sake of ?' there is an implicit assumption that, whatever the subject is, it can do the job that it is intended for. This will hold true of artifacts, texts and people – to borrow Dennett's categorization of those things which interpretations are offered for. (Dennett, Interpretation 1) So making the judgment that the subject is for the sake of something else will naturally lead one to ask what that something else is that it is for the sake of. Because it is reasonable to assume that the subject is suited for the task, one way that the question will be answered is by canvassing the things that the subject might be suited to and from there trying to narrow down this list to get a sense of the purpose that the subject is, in fact, serving.

If the assumption driving 'for the sake of' questions is that the suitability of the subject can be read off of close observation of the subject, then the fact that many subjects could suit more than one purpose will complicate efforts to figure out which purposes are the correct ones to attribute. Indeed given that it will be possible to imagine, for any given subject, a very wide variety of purposes that the subject could be lent to, it will be imperative that some principles are invoked to narrow the list. And, in fact, in common sense interpretation, it is clear that a higher standard than mere suitability is used to explain the purpose behind a subject. It is common to restrict the range of considerations to just those that the subject is well-suited for or particularly good at serving. Thus a judgment that this would make a good x, is one step toward describing it as an x at all. This shows that there is some tendency toward idealization in interpretive inquiry. However, it will become clear that the principles favoring idealization will also favor tempering that ideal in many circumstances.

It seems that it is a certain metaphysical view of artifacts and agency that is driving this style of inquiry. Constructed artifacts and behavior are produced by agents. Overwhelmingly, this is done to serve some particular purpose or other. This much appears uncontroversial and already seems to be registered by the fact that 'artifice' is etymologically related to 'artifact' in obvious ways. Filling this in, however, requires an understanding of agents since it is they who produce the subjects of interpretation. The tendency toward idealization is part of interpretation because of the role that agents play. By being recognizers of and responders to reasons the inclusion of the agent's state of mind is relevant to providing the interpretation of some subject. And since one must assume the subject rational to see her as having the right kind of commerce with reasons, this assumption will be biased toward idealization by attributing good reasons and sound reasoning to the agent. But it is here where I believe the room needs to be made for features of human agency which temper that idealization. Human beings recognize reasons in distinctive ways. We entertain ideas consciously within a stream of consciousness that is only semi-voluntary at times. Further, we are subject to certain tendencies within this activity. Associations readily present themselves, past experience and habits influences thought, and emotion seems able to color a great deal of what we think about. I argue that once it is understood that our actions are borne in an active consciousness of this type, it will be incumbent on us to recognize the distinctive role of this kind of consciousness in producing thought and behavior. Just as idealized adaptionist explanation in biology must be tempered by an understanding of the specific mechanisms involved in natural selection, so too must the idealized rational interpretation be tempered by an understanding of human thought.

Consider an example. If a paleoanthropologist is excavating a midden containing animal bones and flakes of stone, she might consider this spot a good place to recover

arrowheads and spear-points. How should she go about this task? It is obvious that her initial collection is going to be focused on recovering stones of the right shape and size to do the job that arrowheads or spear points are designed for. But, of course, not every triangular piece of stone of the right size will be what she is looking for. She can expect to find, intermingled with her cache, stones which acquired the right shape and size through natural means as well as waste flakes that were the byproduct of the production of stone tools. To sort through this she will have to further discriminate. She will look for marks on the stones that indicate that they were bound to a wooden dowel or she will try to find evidence that certain stones went through a process of additional sharpening. In other words she will make these discriminations by pointing to additional features that make this stone or that particularly suited to the job of serving as a point. She judges these stones to be arrowheads and spear-points by pointing to evidence that reveal them to be good at doing the task for which they were crafted.

So there will be pressure, in interpretation, to invoke an ideal and then to let that ideal guide the activity of interpretation. But it is important to note that not any ideal will do. Indeed the most idealized version of the ideal can be a poor guide to the activity of interpretation. In the example above our anthropologist invokes a tempered ideal to guide her work. Consider that the tools used to take game have undergone a great deal of technological progress since the Paleolithic era. Now, obviously it would be absurd to suggest that the anthropologist should hunt in the midden for pointed blades of stainless-steel and carbon fiber since these are the materials that the best archery equipment is composed of. But the absurdity here stems from the fact that the two notions of good are so disparate. A ‘good’ arrowhead by contemporary standards is vastly different from a ‘good’ arrowhead by the standards of the Paleolithic. In invoking the idealized standard it will be important to index the standard to the situation in a way that is

appropriate. So, at least one source of pressure which militates against an overly idealized constraint in interpretation will be found in noting the historical context and the need for idealization that is appropriate there.

In fact, the use of a tempered ideal for guidance is commonplace in the interpretation of artifacts. The suggestion above only seemed absurd because of the disparity between the best standards of production for archery equipment known and the impoverished resources of our Paleolithic cousins. Perhaps if the interpretation to be done was dealing with present day behaviors and materials it would be appropriate to invoke the most idealized standard as a guide? This thought however seems to fall short of providing reasonable guidance to interpretation when one recognizes that it will be safe to assume that the subject matter being dealt with will probably be of human authorship. That is, in interpretation the goal is to attribute a mental state (typically an intention) which makes sense of the subject. This is built into the understanding of the question given at the beginning of this section. Since one will be hard pressed to literally attribute a rational enough mental state to anything but a human being[3], it must be the case that a safe assumption for guiding the activity of interpretation will be that the subject was created by a human being. As will be shown, this too mitigates how idealized the guiding principles can be. Indeed it appears to me that recognizing this is enough to suggest three aspects of human thought and action which can provide a countervailing influence to the tendency to idealize interpretation.

First, some aspects of human thought are the product of an active consciousness with a variety of influences exerting effects at any given time. Thought can take on the character

[3] Dennett disagrees since he believes many natural systems can register reasons. However, this doesn't impact my general position since the mechanisms which allow for that recognition in the human case will still be relevant to tempering idealizations in interpretation.

of certain attitudes – being variously optimistic, morose, self-congratulatory as well as a variety of others. These influences all carry certain motivational potential and while there are too many to consider here, I take it that most of us are quite good at figuring out which attitudes a person might be in the grips of and how that influence is borne out in thought and action.

In addition, human beings are prone to error. Moreover they are prone to errors of specific types and these errors can often be anticipated. Thus it would be far too stringent to demand that any subject conform to the most idealized standard for an x for it to be understood as an x, since doing so just wouldn't take into account these propensities to error. Many artifacts are not all that well-designed and so it will be evident that some aspect or other of an artifact might have been better able to do the job were it designed differently. Accounting for such defects is also part of the activity of interpretation.

Finally, human agency is constrained by the world itself. Ignoring this is to imagine that effort, resource allocation and design constraints would do not enter into interpretive explanation. But, even well-designed artifacts show evidence of their manufacture that doesn't contribute to their function. Further, these are features that will be relevant to the fullest possible account of some subject.

These considerations point toward invoking a tempered ideal to guide the activity of interpretation. It will be desirable to have an ideal which is indexed to the appropriate standards for the subject. This ideal will also have to be tempered by knowledge of human thought, foibles, and the constraints that the world imposes. So in invoking the ideal that will be used to guide interpretation it will be necessary to place that ideal in the proper context. The goodness of x will not be understood as inherent in the nature of x but will be contextualized to its environment and authorship.

Once the proper contextual constraints on the idealization have been understood it may seem as though the activity of interpretation would be quite straightforward. It would seem that an interpreter would just need to ask 'what could this object or behavior be put into the service of doing?'. From that question a variety of purposes and corresponding contextual idealizations of tools would be made apparent. Then, there would just be the issue of deciding which of these tempered idealizations was the best fit for characterizing the subject of the interpretation, and so, ascribing a purpose and identity on that basis. However, this is overly simplistic. There is one more potential source of confusion that must be resolved before the question is answered. The issue is concealed in the question above: how would one go about deciding which of the many idealizations possible for the object will be the one which determines the purpose and the identity of the subject. What does it mean to decide which idealization 'fits best'? Thinking the best interpretation is just a matter of choosing the purpose which the subject would best serve is to invite a form of bias into the method. Simply put, that method threatens to misdescribe actions and artifacts that emerge from indecision, ineptitude, confusion and lassitude as aiming at exactly what they accomplish.

One problem here is that it is often ambiguous as to whether an action was mistaken or intended. Sometimes one aims at some outcome and, in going wrong, brings about a behavior or object which could be understood under a different aim than that which guided the person. Consider a case of distracted driving. Mary may intend to travel home by the most direct route possible. However, when the time comes to make her turn she misses it and so ends up traveling on a longer and more scenic route. If this kind of behavior is habitual enough it will be difficult to detect as a failure to fulfill her intention. The attribution of error in repetitive activities is most plausible when some instance or other is aberrant in comparison to the other

cases –"This widget has the brand name stamped on twice, when all the rest only have a single stamp". It is also plausible when the agent seems to be trying to return to a previous step in some process. These are signs that the performance has gone wrong. But not all errors present aberrations of this sort. This is obvious enough from Mary's case. An interpreter sometimes will find it possible to offer two quite different interpretations of this subject matter. Mary could be seen as aiming at the most direct route and failing each time. Alternately it will be possible to see her as aiming at taking the scenic route home.

This problem stems from the fact that each subject will be inherently ambiguous between different reasons and the idealization will always seek to describe her as motivated by her best view of reasons. Some subjects will simply be very good candidates for a variety of purposes and in this cases it will be very difficult to tell which of the candidate purposes is the right one. Similarly, some subjects will be ambiguous between aiming at a purpose well or aiming at a different purpose badly. Some of these ambiguities may not be possible to resolve given limited information. It is here where it becomes clear that narrowly focusing on idealized interpretations is bound to introduce bias.

So intentions can be understood as states which motivate by way of the desirability or justification of the end they contain. Indeed the motivational force these carry is essential to the explanation. Many things can be justified in the abstract. Any person will probably see a great many things as desirable, worthwhile or justifiable pursuits. But the mention of any of these becomes explanatory in that context when they are part of what motivated some particular action. It seems then that giving rise to the motivational force is part of what makes them explanatory in interpretation. Habit, emotion, and vivid imagination, for instance, also have motivational force, or the potential for it, but none of these is obviously end-directed in the way

that an intention is. Thus, if there is an explanatory role that citing motivational force brings to interpretation, then there is no reason to restrict that inquiry to only intentions. Motivations alone may be explanatory without citing intentions.

To conclude this section it will be helpful to review the headway which has been made. It has been shown that offering an interpretation is an activity which requires idealization. However that idealization must be tempered by knowledge of context if it is to serve as useful guide. Further, the environment and human foible are two places where the need for such tempered idealization is particularly apparent. Finally, there is an inherent ambiguity in the observation of an action. This ambiguity is between seeing it as a successful accomplishment and an act which managed to accomplish something. This means that even the appearance of success is a potentially misleading guide to intention ascription. Further, while intentions provide the kind of explanation demanded by the interpreter, this is due to their motivational force. As such, other states with this feature can also be expected to provide that form of explanation.

1.3 'NO PARTICULAR REASON' AND NEGATIVE ANSWERS TO THE 'WHAT IS IT FOR THE SAKE OF?' QUESTION.

So it is clear that expecting an affirmative answer to the 'why?'/'what is it for the sake of?" question can generate a range of further questions that must be dealt with in further specifying the answer to that question. However, this does not exhaust the interpretive work that is connected to the 'why?' question.

Anscombe argues that one way the 'why?' question is refused application is when there is reason to think that the person does not know that they are performing the action under

that description. (Anscombe 24, 38) She takes it as obvious that if one does not know they are doing some x under the description of it as an x, then there is no possibility of that person giving an account of why she is doing it as an x. (Of course she may be able to supply an intentional account under some other description.) Because the question is refused application under these rather narrow circumstances, the negative, but not question rejecting, answer to the 'why?' question must also be treated as in need of an interpretive response.

Now in the previous affirmative answer to the question the interpretation was provided by supplying an intentional end which motivates. But here, the negative answer to the question is speaking to the non-existence of such an end. Thus there will be no end in view which motivates the action. At the same time, however, the question marks the existence of a motive to perform the action. The person recognizes the action or artifact as intentional; as springing from her motives and conscious action. What she denies is that she has a further sake or purpose which explains it. Thus in order to make sense of such cases it seems that a certain kind of explanatory mechanism will have to be invoked. The possibility of a motive which is not an aim or end must be examined.

Cases like this are very difficult to describe. There is a tradition in the literature of treating the basic form of human action as goal oriented. Human psychology is treated as composed of a few trunks that represent our deepest and most basic interests. Thus concern for, say, family, self, and community might be thought to compose the most basic fonts of some individual's desire and then each new action is explained by showing how it emerges from those basic concerns. For instance, by showing how the small and completeable task that is being performed contributes to the overall flourishing of that concern. But this picture strikes me as overly simplistic. There seem to me to be a great many motivational states that emerge

unconnected with one's 'deep' concerns. Indeed, a picture of human agency which denied this would seem to present an agent who is earnest to the point of being super-human.

This 'Humean' picture of practical-reason is sometimes given an 'internal' analysis. Doing so elevates satisfaction as though this were the endpoint of all desire; as though each desire existed in order to bring about some later good called 'satisfaction'. But this ignores the fact that many of our desires aim at some state of the world or other. To place some subjective proxy in the way and describe it as the true goal is to suggest that all actions are in the service of ourselves, when in fact many things we do serve others, serve the world or aim at something other than satisfaction.

So how can sense be made of these kinds of emotional and motivational states if the language that is employed to discuss action is so saturated with goal-oriented prose that it becomes difficult to state the possibility of non-end directed motives without inviting a reformulation in terms of ends?

One way to begin is to focus on doings that are the least like performances or executions. Beginning with a domain of things that are done in response to some circumstance but not for the sake of some future state of affairs would seem to provide a particularly clear case of this kind of motivation. I propose to begin with an analysis of this kind and see how far it will take the inquiry.

Sometimes a person will adopt an attitude in response to some state of the world. For instance, as the weather warms up after a blustery Pittsburgh winter I may find myself to be more cheerful. This good cheer can prompt a variety of conscious action which is done because the warm weather has acted as inspiration. I might elect to smile and make small talk with my fellow bus riders or to put on a nicely pressed shirt and pants. These are attitudes or behaviors

that I adopt in response to the warming weather and they are under my control. Indeed the evidence for their being under control can be seen in the extent that the agent might choose to dial such behaviors up or down in response to the situation. Can such things be made sense of?

I know of no way to answer that question except to try and offer an explanation for it and see if it satisfies the curiosity which sparked the inquiry. But in the case at hand I think there should be no difficulty making sense of the person (though giving an account might prove more difficult). If we imagine that in my friendly state someone says to me 'My you are being particularly friendly today, what are you up to?' there seems to be a straightforward sense in which 'I'm not up to anything, it's just such a beautiful day today how could you fail to be in a good mood?' answers the question makes-sense of my mental states, so discharging the interpretive impetus of my interlocutor.

So, it does seem possible to make sense of a prompted motivation that is responsive to the environment without being aimed at some further end down the road. However, this won't satisfy the theorist who is committed to analyzing all action producing mental states in terms of belief and desire. For such a critic it will be possible to reframe the explanation just provided in the language of belief and desire. As I argue later, such moves are problematic in their own right because they get part of their plausibility by seeming to offer theoretical parsimony. I claim this parsimony is a false economy because the qualifications to the ideas of beliefs and desires which must be introduced multiples entities in an unsatisfying way. The immediate effect of this move, on the issue under discussion, will be to exclude negative but non-question-rejecting answers to the 'why?' question[4]. They will be excluded because they will,

[4] This might be thought to not be a negative answer to the 'why?' questions since it cites a further motivation. However, it does begin with a denial that there is any further intent. As such it appears to be a way of saying 'no

in effect, be translated to the positive case previously discussed. Now perhaps it will be true that these negative answers can't be understood except as special cases of the positive answers. However, such an assumption should not be built into the terminology used to dissect the problem. Moreover, Anscombe, at least, is prepared to treat the negative answers as genuine and distinct from the positive answers. Thus the issue needs to be investigated with an eye to seeing how plausible the belief/desire theorist's attempt at redefinition is and if it is possible to make these negative answers truly distinctive. After dealing with that issue I will turn to an analysis of the questions that will need to be answered in offering negative answers to the 'why?' question as real interpretations of subject.

Two strategies suggest themselves for reframing these negative answers as crypto-positive answers. First, one could argue that these 'inspired-by' actions are ends in themselves. Alternately it could be argued that there is some hidden end that is being aimed at and for whatever reason the negative answer can be understood and can be explanatory without explicit reference to that end. I will deal with these in order.

The case of weather-induced friendly feeling just discussed would appear to be an example most amenable to someone pursuing the first strategy. What makes it possible is that it would be reasonable to suggest that each person has a kind of standing interest in making friends, appearing sociable and, in general, experiencing the joys of camaraderie. Achieving this requires cooperation from the world, but it is often sufficient to simply make oneself open to such goods for them to be realized. Thus it might be said that friendliness is an end in itself because it is part and parcel of such standing ends that one has. In the same way that exercise and

---

particular reason' which Anscombe takes to be a negative but non-rejecting response to her question. (Anscombe 26)

eating well are part and parcel of maintaining one's health, and so in need of no particular justification, being friendly is just part of being social and in good cheer.

So it is the assumed desirability of friendship that can work to make those attitudes which are partly constitutive of friendship appear desirable for their own sake. However, as a general analysis of these negative answers which cite attitudes, it will only work if all such attitudes can be associated with the relevant desiderata. This association is by no means innocent. In fact it structures the way that the attributed mental state is understood in the interpretation. In order to be seen as an end-directed understanding of the personal mental state of the subject it will need to be the case that the goal of friendship is the driving force behind the adoption of the attitude. But some reflection suggests that this way of understanding the attitude is quite inflexible. It can, and often does, happen that one is simply moved to take on some attitude or other. A simple felt urge can alter a person's disposition. Now, it is also true that some dispositions serve further ends. I adopt a wary attitude at the poker table because I am guarding against being out-gamed. But this is not a necessary feature of wariness. Being wary can just strike one as apt when there is no further goal in mind. If a man walks at night and notices that he begins to pass the graveyard at exactly midnight, it may strike his fancy to whistle – but we need not ascribe a fear of the boogeyman on that basis. And, a glorious sunset can arouse a quasi-religious kind of gratitude, even in the atheist.

In fact this failure of an associated end to be necessary to the explanatory force of the interpretation also threatens the other strategy of the belief desire theorist. This problem is easier to show and follows from the last conclusion. Recall that the other strategy of the theorist was to posit a hidden end that was being aimed at. But just as the last case showed that there was no needed end to make sense of prompted behavior, there is no need to posit a hidden end to explain

it either. Thus it seems safe to assume that the strategy of redefinition can be set aside as offering a serious threat to the possibility of interpretations of this type.

Let us now turn to the issues involved in offering an interpretation when the answer to the 'why?' question is negative without rejecting the basis of the question.

The method of interpretation will be complex and more open-ended than in the first case. It will be more complex because while I have used an emotional attitude being adopted with no further purpose as a particularly vivid example of one such negative answer, there is no guarantee that all such negative answers will be modeled on that one.

This is not to say that there will not be a unity among such cases. In fact there is already a unity binding them in that they can all be characterized as non-earnest action – there being no end which animates the behavior.

Still the realm of the non-earnest is very broad and while it is clear that it will include things like emotional behavior, sympathetic behavior and behavior motivated out of habit, it is difficult to spell out exactly where the boundaries lie.

To overcome this issue I suggest that it is possible to use a common-sense understanding of mind to guide the interpretation.[5] The way I want to make use of this idea is as follows. I take it that one of the most common types of interpretation one engages in is self-interpretation. This activity is often immediate and for this reason may not seem like interpretation at all. But this should not be understood as a liability. Freed from the constraints of any explicit theory, immediate self-interpretation is often quick to make use of a vernacular vocabulary for describing mental states. Further because first-personal access presents information that may not be available to others, a range of associations, urges, vivid pictures and emotions will enter into

[5] This suggestion will be spelled out in more detail in subsequent chapters.

these explanations. This access will be especially good at pointing to the motivational impetus which prompted the action. While a great deal of behavior is ambiguous enough to be interpreted earnestly, it should be expected that only a subset of this will be treated as such by the subject herself.

Thus it would seem that first-personal interpretation will be less theoretically constrained and less biased toward the earnest than third-personal interpretation. Further it has some claim to be 'better' because first-personal access to one's mind is generally taken to be more penetrating than third-personal access. Noticing this suggests that leveraging these features of self-interpretation into an interpretive method would have the potential to offer interpretations which were less prone to the bias toward the earnest. But how could this be done, given that each of us only has first-personal access to ourselves? If first-personal interpretation is to be the model for general interpretation, will it be able to offer accounts of the behavior of others?

If the benefits of first-personal interpretation are to be extended more generally it will have to be because our natural sympathy for other people can help put us in contact with their mental states in a way which is not theoretically mediated. So, I suggest that, in those instances where behavior or artifacts are observed which might plausibly be seen as ambiguous between being earnest or being prompted by non-earnest motivators, the method of interpretation should seek to employ sympathy in adjudicating the case. This might be accomplished by simply asking oneself, while employing the resources of vivid imagination 'what would motivate me to do as she did, if I found myself in such circumstances?" Drawing on the kind of sympathetic understanding of the other person would then allow for the resources of self-interpretation unconstrained by theory to provide interpretations of others that were less biased in favor of the earnest.

The most obvious way for such interpretation to proceed is by beginning with a negative answer to the question 'why?' which does not reject the question entirely. Anscombe offers the answer such as 'no particular reason' as an example of this kind of answer and a little imagination should reveal that there are other answers in this neighborhood that would serve equally well. This case serves as a best case scenario because it begins with a personal account of the motivations which prompted the action. If it is granted that generally one's access to her own mind is better, and so carries more weight, than another's interpretation of behavior, this case then begins with a particularly strong form of presumptive evidence. Further the denial that the action was motivated by reason means that what will be offered in the explanation of the behavior will cite motivating features of her mental state. As I have suggested above, it seems that the only natural unity these admit of is not being reason motivated. However, this is just to say that it is difficult to give a principled account here. Instead it seems that it will be most fruitful to illustrate this class by saying what features of one's mental life can be explanatory in this inquiry.

It is easy to imagine scenarios and so fill in the kinds of things that would come after an answer such as 'no particular reason'. A person might go on to describe some strong mental imagery that drove the behavior. "I didn't lock the deadbolt for any particular reason, it is just that it occurred to me to do it after remembering that horror movie we saw earlier in the evening'

I don't mean to suggest that simply answering this question would be sufficient for providing an interpretation. One will have to resolve a series of questions after providing an initial answer to better fix the attribution. For instance, almost all emotional behavior will be ambiguous in the sense that it will be possible to characterize it as earnest. As was previously seen, being friendly is consistent with interpreting the subject as holding the goal of seeking

friendship though it is also possible that such behavior be prompted by mood or be part of one's personality. As such it will be necessary to ask if there are not further markers present in the vicinity of the behavior which would sway the attribution toward the earnest or away from it. So, in answering the question in this way what is being acknowledged by the agent is that the behavior is prompted by a form of motivation that is not being driven by reference to some end. This need not remove the behavior from the domain of reason but it is to treat the reason, whatever it may be, as not gaining its force from the desirability of some end being pursued. In other words, it is to invoke the possibility of non-instrumental motivation. That said, citing a reason of the appropriate sort is only one way of fleshing out the answer to this question; others have been suggested above. In order to do a full taxonomy of the kinds of answers that can be given here a more general inquiry into the nature of motivation would have to be conducted. That inquiry could be expected to distinguish those types of motivation that proceeded from reasons, especially end directed reasons, and those which owed their provenance to other sources. From this it would then be possible to canvass those types of motivation which could be cited as ways of further explaining the negative by non-rejecting answer to the 'why?' question.

It has been shown that these negative answers to the 'why?' question share features with the positive answers that allows them to fill the same explanatory role in interpretations. Although there is no reason to expect a unification of these responses beyond their role in negatively answering the 'why?' question, the explanatory force of these motivators shows that instrumental considerations need not be the only source of motivation. Finally, I have shown that self-interpretation, coupled with sympathy and imagination, shows potential as providing one way of dealing with the collection of states that might be cited as providing an interpretation while still offering a negative answer to the 'why?' question. Since what is needed in

interpretation is an account of a personal level mental state which is motivating and under the appropriate agential control, it should be expected that interpretation won't be limited to merely offering intentions as explanations for acts and artifacts.

A note about simplification. The strategy I have pursued in this paper has been to divide sections according to a plausible taxonomy of answers to the 'why?' question. This has the effect of clearly contextualizing certain considerations and methodological assumptions. However, it does run the risk of implying that one can tell, in advance, which behaviors are earnest and which are not. Unfortunately things are not so simple in practice. Distinguishing the earnest, from the non-earnest and from the merely causal (the cases we turn to next) is part of the activity of interpretation. While I do believe these divisions are basic in the sense that an explanation must fall into one of these categories, that they are so fundamental should not be taken to show that a cursory observation will reveal what form of explanation is needed for some action or artifact.

1.4 REJECTING THE APPLICABILITY OF THE QUESTION.

There is one further type of answer to the 'why?' question that this essay has yet to consider. Those negative answers which reject the applicability of the question must also be considered. In one respect such answers are not relevant to the purposes of this essay. If interpretation is understood as providing the mental state that motivated the behavior, then this answer implies that in this territory there is no interpretation to be had. When the agent rejects the question by stating that she did not know she was engaged in that behavior, this is a fairly sure sign that there was not some conscious activity which motivated it.

However, the question is important to treat here for two reasons. First, interpretation seeks to account for the behavior of the subject. So being able to say that this behavior is and that behavior is not amendable to interpretation will be part of providing explanations of the right kind. When the interpreter is queried about why she fails to provide an explanation in such and such a case, she can answer that this is a class of behavior that does not admit of interpretation and point to this division of answers as part of the reason why. Second, since this behavior does not admit of interpretive explanation it will be natural to treat it as amenable to scientific explanation. Further still, since the behavior at issue will be outside of the conscious control of the agent it may seem appropriate that it should have a causal explanation rather than an interpretive one.

The fact that this mode of explanation will be seen as applicable raises a potential problem for defending interpretation as a viable tool for explanation. Simply put, since neurological and other causal mechanisms will be adequate to explain certain kinds of behavior, and since it cannot be expected that the neurologist will recognize the distinctions generated by way of the 'why?' question, it may seem as though the neurologist has the greater claim to be offering the right explanation of all behavior. This will be furthered by noticing that the interpreter, by treating the behavior in question as failing to admit of the right kind of explanation, is already ceding some forms of behavioral explanation to the neurologist. It is natural to wonder if this concession doesn't already cede too much.

One possibility for defending interpretation as an adequate mode of explanation is to point to the purposes it is employed for. It might be the case that a neurological explanation can adequately describe some phenomena and yet be unable to provide the type of description that would be useful to someone who sought an interpretation.

There is a great deal of evidence which suggests that with respect to purposes the interpreter and the neurologist are in pursuit of quite distinct kinds of information. Notice that it is often very difficult to translate the highly technical microstructural claims of the scientist into information that is useful for predicting the more prosaic behavior of individuals going about their lives. When one is told that emotions are simply changes in stress hormones, involuntary muscle contractions and other visceral states that is initiated by the amygdala's response to stimuli that may not be consciously perceived, this kind of information cannot be readily assimilated into an interpretive explanation.[6] Indeed if one is interested in accounting for overt behavior with any degree of complexity it would seem that a great many of these microphysical explanations would need to be concatenated in such a way that the observable inputs and outputs could be correlated. It is not clear that there is as yet an extensive enough understanding of such matters that would allow for this kind of link to be established in the way that those who seek after interpretations would find useful.

However, the promise of such a style of explanation becoming useful as further progress is made cannot be ignored. As a first pass at dealing with this possibility I want to suggest that there are, broadly speaking, two directions that future progress might take. Each of these must be analyzed separately. First, it is possible that future research will find a way to assimilate casual and interpretive explanations. Second, it may be the case that future research makes causal explanations increasingly divergent from interpretive ones.

If future research tends to allow assimilation of the interpretive and the causal this will take place by causal explanations acting as answers to interpretive questions. This follows from the fact that interpretation has been shown to be demarcated its ability to provide an answer to a

[6] See Prinz 2005, *Gut Reactions*.

distinctive question. So in order for the causal explanation to be seen as a kind of meaning explanation (or vice versa) it will have to be the case that this style of explanation supplies answers to questions about meaning. Further, if this were to happen it seems that it could only be done by supplying an account of the subject's mental state in a way which allowed for this meaning to be understood as an outgrowth of the subject's conscious mental activity. This suggests that for causal explanations to be interpretive the direction of progress will be toward giving a causal explanation in a form which is easily integrated into our common-sense explanations of mindedness.

So if progress develops in the direction of assimilation this doesn't pose a threat to seeing interpretation as a distinctive and viable form of inquiry and explanation. It can still be seen as distinctive in the questions it answers and the answers to those questions will make use of terminology which can be readily assimilated into the existing vernacular talk about minds.

The more troubling development from the standpoint of seeing interpretation as distinctive and viable is given by the possibility of further divergence. While mere divergence by itself doesn't pose a problem, the threat is that increasingly powerful explanations developed along the course of that progress will come to hollow-out the distinctive subject matter of interpretive explanation. In effect it would be to show that questions about meaning can be collapsed into questions about whatever mechanisms the causal explanation deals in by showing that a merely causal explanation was, in all cases, just as informative and efficacious as its interpretive counterpart. Instead of being seen as an assimilation however, the causal explanation would not adopt terminology that could be assimilated into common-sense explanations of mindedness.

I concede that such progress would be a threat to seeing interpretation as distinctive and viable. Essentially such a development would be the vindication of the monoist approach to mind. However, I think the threat can be seen as relatively remote. The reason for this is that not only would the progress have to create a divergent set of explanations for the phenomenon of interest to the interpreter, this would have to be done in such a way that showed that interpretive questions were poorly formulated or otherwise not coherent. As noted above, mere divergence is not a threat since that preserves interpretive explanation, instead it is only a very specific kind of happening along that path of progress which would pose a threat to my account of interpretation. So while this possibility does pose a problem for the interpretive strategy, that threat is somewhat remote at present.

To conclude, I have shown that interpretive inquiry appears to be pursuing a distinctive form of explanation. I have further shown that principles used to conduct this inquiry require tempered forms of idealization and must be supplemented by knowledge of the unique character of specifically human agency.

2.0 DISTORTING THE INTERPRETATION: REPLACING THE PRINCIPLE OF CHARITY IN RADICAL INTERPRETATION

2.1 INTRODUCTION

The most well-developed and clearly spelled out philosophical account of interpretation is that first developed by Quine and later transformed into the doctrine of radical interpretation by Davidson[7]. Because some form of interpretation underwrites any attribution of mental content, the doctrine of radical interpretation has acted as a starting point for a great deal of work in epistemology, ethics, and the philosophy of mind. It is not difficult to see the attraction of these. Radical interpretation promises to provide a method powerful enough to make any thoughtful behavior interpretable. As a corollary it seems to show that whatever cannot be interpreted by this method doesn't count as thoughtful. The method is designed to produce a picture of the subject that is modeled on a picture of a rational agent. Its central constraint, the principle of charity, favors finding shared beliefs and desires between interpreter and subject and so it is biased in favor of finding the agent rational from the standpoint of the interpreter. For Davidson, this bias is built in as a way of ensuring that the subject's beliefs are interpreted as generally true. But this concern for truth comes at a high cost. It means that radical interpretation distorts action that is thoughtful but not rational. For example, common irrational actions, like yelling at a lazy

[7] See Quine 1969, 1970, Davidson 2001c, 2002a, 2004 and Lewis 1983

player when watching televised sports, are easily understood, but radical interpretation doesn't have the resources to account for how this understanding is achieved. I propose that a better place to find the commonality that makes interpretation possible is in recognizing that interpreter and subject share similar ways of thinking that do not always reduce to shared beliefs and desires.

Davidson's work on interpretation has had far reaching consequences for philosophical methodology. The principle of charity is used to attribute a core of rational thought from the totality of behavior which can be observed. To invoke the principle is to claim that it is constitutive of thought that it be rational. Ralph Wedgwood's recent book *The Nature of Normativity* (2007) gives an argument for the reality of normative principles that depends on the principle of charity. Wedgwood reads a respect for rational norms off of observed behavior. He uses this natural respect for norms to show that normative principles are causally efficacious and so real. However, if it can be shown that charity tends to distort the picture of the subject, Wedgwood's constitutive argument will need to use a different strategy to secure the existence of normative principles.

In what follows I argue that the principle of charity should be replaced with a principle that directs an interpreter to minimize inexplicable behavior[8]. Accommodating this new principle means expanding the explanatory base used in interpretation beyond empirically justified beliefs and desires. In support of this change I argue that Davidson's theory of radical interpretation is unable to give a plausible picture of irrationality. I show that this failure is a result of his inability to see irrational behavior as underwritten by thought. This ignores the

[8] Grandy 1973 proposes a kindred principle. There he argues that because of the distortions introduced by charity, an interpreter should instead adopt a principle of humanity that directs one to minimize the attributions of inexplicable error.

possibility of using mental states that are known first-personally to explain the behavior. I criticize him for producing a picture of a mind which contains an abundance of true beliefs which come at the expense of an explicable and recognizable psychology. The source of this distortion is Davidson's principle of charity.

This revision to radical interpretation has consequences as far reaching as the original principle. I show that a distortion-free picture of the subject contains no privileged set of dispositions that can be used to reveal that only rational norms naturally govern thought. While nothing I say shows that rational norms cannot exist, I show that relying on dispositions, as Wedgwood does, presents a poor method for revealing this set of norms. Further, I show that because a great deal of irrational human behavior is easily understood and made sense of, there is no reason for an interpretive method to be a priori biased toward only producing rationalizing explanations. So, drawing on my critique of Davidson, I show that Wedgwood's method for associating rational standards with interpretable attitudes is flawed. Not every attitude that can be identified has a place in a rational life. Some attitude types are simply irrational. It follows that the best interpretation of a subject will include forms of sensitivity that function in their proper role but are independent of standards of rationality.

This essay is arranged as follows. In section one I present a summary of Davidson's theory and examine how it is used to interpret irrationality. In section two I offer a critique of Davidson. I argue that the principle of charity fails to provide the best interpretations of a subject's behavior in some cases. I show that taking the phenomenology of mental life seriously provides a way to produce more plausible interpretations of the subject. In section three Wedgwood's account of normative facts and its dependence on the principle of charity is introduced. I show that his argument relies too heavily on the principle of charity and so will

tend to over-rationalize dispositions. I then show that it is possible to produce more plausible interpretations by giving greater evidential weight to first-personal reports of a subject's thought. Finally, in section four I show how to replace the principle of charity and how the constraints on interpretation will be altered to accommodate this.

2.2 DAVIDSON'S RADICAL INTERPRETATION

Davidson's theory of interpretation is premised on the idea that if one encounters and attempts to decipher apparently intentional behavior, then one must begin by assuming that the subject is rational. To understand the subject as rational, in turn, requires that one see him as exercising oversight over his system of beliefs in such a way that those beliefs form a coherent network and where the contents of any particular node are specified by the rational relations connecting it to others. On this understanding, the agent monitors his network of beliefs and will either correct or form beliefs according to his view of what the balance of reasons supports.

Davidson describes interpretation that proceeds from this starting point as a way of extending a principle of charity to behavior that appears intentional but is not yet understood. However, Davidson does not view the extension of such a principle as optional. He takes it to be constitutive of mental content that it exists in a network of rational relations. As such, the only way to begin to understand the actions of an agent is to posit that he is rational and has a stock of beliefs and desires that prompt his actions. Because the principle of charity is acting as a constitutive ideal, Davidson is offering a metaphysical claim about what mental content is. It is that which the best interpretation specifies it to be. According to Davidson then, it is the nature of mental content to be rationally interpretable.

The biggest obstacle to offering an interpretation of this kind is getting started. After all, Davidson's imagined investigator is interested in understanding purportedly intentional action. But this can only be done by constructing a theory about what the subject believes and takes to be a reason for action. However, any theory which could ground claims about what the subject believes and takes to be reasons would surely be underwritten by the very behavioral evidence that the interpreter seeks to understand. How then does one get started? The answer here is deceptively simple; one begins by assuming that the subject is like oneself. Any being contemplating doing some interpretation of another being is already minded and so rational according to Davidson. As such, it is the interpreter's experience with his own rational life which sets the rational standards for the interpretation to conform to. That said, although Davidson does suggest using one's own experience with the unfolding of one's mental life as the model for interpreting others, exactly what parts of one's experience get attributed as part of this baseline differ with respect to the different goals one is trying to achieve in interpretation. So for instance, in *On the Very Idea of a Conceptual Scheme* Davidson summarizes the central issue involved in providing a Tarski style translation for an alien language.

> "What matters is this: if all we know is what sentences a speaker holds true, and we cannot assume that his language is our own, then we cannot take even a first step toward interpretation without knowing or assuming a great deal about a speaker's beliefs. Since knowledge of beliefs comes only with the ability to interpret words, *the only possibility at the start is to assume general agreement on beliefs.* … Charity is forced on us: whether we like it or not, if we want to understand others, we must count them right in most matters." Italics added (Davidson, 2001b 196-197)

It is helpful to note that "counting them right" here means counting them in agreement with oneself. In other work, which discusses the possibility of making sense of another human being's thoughts and emotional responses, as opposed to offering Tarski truth sentences for an alien language, Davidson restates the principle of charity.

"We start out by assuming that others have, in the basic and largest matters, *beliefs and values similar to ours*. We are bound to suppose someone we want to understand inhabits our world of macroscopic, more or less enduing, physical objects with familiar causal dispositions; that his world, like ours, contains people with minds and motives; and that he shares with us the desires to find warmth, love, security, and success and the desire to avoid pain and distress" Italics added (Davidson, 2004a 183)

So Davidson directs anyone who would make sense of the action of another to begin by assuming that the subject is a rational agent with beliefs and desires roughly similar to one's own. This ensures that the beliefs attributed will largely conform to the interpreter's view of what is true. To see behavior as interpretable then is to see it as underwritten by thought which conforms to the interpreter's understanding of the constitutive ideal of rationality. This means that the principle of charity does double duty; both securing a commonality to ground the interpretation and ensuring that the attributed beliefs will be generally true. This principle of charity then acts as a constraint on any attempt at interpretation. This should not be taken to imply that any attempt at interpretation is bound to terminate in an account of a maximally rational agent or that irrationality simply can't be interpreted. Indeed, explaining some piece of behavior by describing it as irrational is not to abandon the principle of charity. Davidson's argument is to be taken to show that it is only against a background of broad agreement, on a variety of supporting facts, that a person could be understood as believing anything: rational or not. That is, it would not be possible to ascribe to Smith a belief that he was Napoleon or that his wife was an eggplant, unless he could be taken to understand, in some sense, what it meant to be Napoleon or an eggplant. [9] Without this kind of agreement these claims would be mere word

[9] Bizarre examples aside, Davidson's commitment to the constitutive ideal of rationality is very strong. Indeed it seems to function so strongly as a constraint that the conditions necessary for ascribing a really strange belief, like those given, could almost never be realized. Consider that if Smith understood just what his wife and an eggplant were, it is very difficult to imagine how he could ever sincerely claim that an identity relation existed between them. By taking Smith's initial claim at face value and then pressing him for the details necessary to remove the initial befuddlement, an interpreter might very well come to decide that Smith is not in possession of very coherent ideas as to what kinds of things his wife or an eggplant actually are. (See Davidson 2004b, 196-197)

salad on Davidson's account – vocalizations that sounded like English but on investigation turned out to be parrot-like chatter.

This possibility of bracketing the principle of charity to explain irrationality raises the issue of how exactly such cases are going to be explained. It is clear that Davidson is aiming at an account that comports with the kind of understanding that the interpreter has of himself. However, as I will argue, this aim is not met by Davidson. To explain irrationality he posits that the subject is split-minded. The upshot of this posit is that transitions in thought are described as merely causal. The control and awareness essential to the attribution of mindedness has been lost. While such a drastic loss of control over one's own mind is certainly possible, it counts against any method of interpretation if this is the only kind of account of irrationality it can offer. Later in this paper I will show that this case can be plausibly and prosaically explained by giving an interpretation from the first-person that does not obscure the subject's control over and awareness of himself.

Davidson argues that cases of self-deception can only be explained by the presence of conflict that gives rise to metal division. He gives an argument, which he acknowledges is roughly Freudian in its outline, for the claim that self-deception is essentially a successful motivation to believe p that is sustained as a result of a belief that not-p. (Davidson, 2004a 206) More specifically, he claims

> "an agent A is self-deceived with respect to P when … the thought that he ought rationally to believe p motivates A to behave in such a way as to cause himself to believe the negation of p … the state that motivates the self-deception and the state that produces it coexist; in the strongest case the belief that p not only causes but sustains a belief in the negation of p." (Davidson, 2004a 208)

This is problematic, as Davidson recognizes, because, prima facie, one cannot offer an interpretation of another as committed to a contradiction without abandoning the principle of

charity. If it is one's own experience of oneself, as a rational being, which is supposed to drive and underwrite one's understanding of others, then one couldn't offer an interpretation of another doing some thing unless one could understand oneself as doing it. But it is virtually impossible to conceive of what it would be for a rational being to knowingly hold a contradiction. As such, it is impossible to imagine oneself doing so. If one's status as a minded being is constituted by being an epistemic monitor of his greater network of beliefs, then being aware of tensions should make it impossible to sustain an attitude of belief toward both of the opposing propositions. An attitude of suspended belief (or, failing that, a belief in one at the expense of the other) is forced on the agent by his status as a rational monitor of his beliefs. One can no more commit to a contradiction than one could believe oneself to not be a thinker, and so by the principle of charity, one should not be to attribute such a state to another rational being either.

Davidson evades this conclusion by offering an account of what it would mean for a rational agent to have a split-mind. As he describes it, a split-mind would bring about the conditions necessary for a certain kind of monitoring failure to occur. In the particular case he uses as illustration (that of a bald man who is self-deceived into believing himself hirsute) the split-mind allows for the belief that he is bald to be safely portioned off from the totality of his other beliefs. Although Davidson does not delve very deeply into the psychology of the imagined case, he implies that the split-mind segregates one, or more, of the following: the epistemic norm to consider all the evidence, the uncomfortable belief, or the contrary evidence. This prevents them from being brought to awareness and so blocks the recognition of the contradiction. At the same time, the evidence for his not being bald is kept on the conscious side of the partition, and so, when reviewed by the agent, is available for underwriting his optimistic assessment of himself. Davidson describes his result as follows:

"The irrationality of the resulting state consists in the fact that it contains inconsistent beliefs; the irrational step is therefore the step that makes this possible, *the drawing of the boundary that keeps the inconsistent beliefs apart*. In the case where self-deception consists in self induced weakness of the warrant, what must be walled off from the mind is the requirement of total evidence. What causes it to be thus temporarily exiled or isolated is, of course, the desire to avoid accepting what the requirement counsels. … In the extreme case where the motive for self-deception springs from a belief that directly contradicts the belief that is induced, the original and motivating belief must be placed out of bounds along with the requirement of total evidence" Italics added (Davidson, 2004a 211-212)

This solution allows Davidson to respect the injunction to maximize true beliefs by the interpreter's lights, but he fails to notice that it does so at the expense of the assumption that subject is much like oneself. The split operates by preventing awareness of certain aspects of the subject's psyche. In order to be split-minded he must remain unaware that he is split-minded and this goes for anybody in such a state. But this means that even if the interpreter was split-minded he could not be aware of that fact within himself; he couldn't experience it first-personally. That is to say that his own condition would be invisible to him.[10] If he attributed such a state to the subject this could only be a theoretical posit, he couldn't make the attribution on the basis of personal experience with a state like that. As such, the posit of a split-minded subject saves the phenomenon of a mind that, by and large, believes the truth at the expense of a first-personally recognizable psychology. Here assuming that the subject is much like oneself and assuming that the subject believes the truth come apart. Davidson has sought to ensure that it is the injunction to preserve the truth which is retained while seemingly showing no awareness that this is occurring at the expense of understanding the subject as the interpreter understands himself.[11]

[10] Contrast this with telling a child that the peculiar sensation she describes is a result of striking her funny bone for the first time. Here the distinct first-personal experience provides a route to interpreting her feelings.

[11] For Dennett an intentional description comes with the attribution of some function that the agent aims at. But in this case no such attribution could take hold because the function of the split-mind is to conceal itself from the agent. Dennett 1990, 1996.

2.3 ALTERNATIVES TO THE SPLIT-MIND

Although Davidson's account does seem to describe what would be necessary for rational creatures to accept a contradiction and succumb to self-deception it does so by way of positing the existence of a strange mental phenomenon: a split-mind. It is odd that Davidson never explains how it is that conflict is able to bring about the failure of self-monitoring that results in self-deception. After all, not every episode of conflict or indecision results in a failure to self-monitor. A man can wish he was six inches taller without coming to believe that he is. Nor does Davidson ever specify how a mind could be both committed to some proposition while working to suppress the truth of it within itself.

More troubling is the fact that Davidson is here positing a mechanism which one could never have first-personal experience with. The split-mind of the bald man is a hidden force that could never work under the direction of the person because it is beyond his awareness and control. A man cannot be self-consciously split-minded. It would seem that Davidson violates a key constraint in offering this interpretation: no interpreter could be aware of his own split-mind. So, attributing one to another person fails to see the subject's thought as much like the interpreter's.

There will always be first-personal ways of understanding prosaic cases, like the bald man Davidson uses, that do not require the posit of two minds with one the victim of the other's machinations. Further the less prosaic and more deviant the explanation, the less it looks like an instance of irrational thinking as opposed to a loss of control over one's mind. In Davidson's example, it is far more plausible for the man, when confronted with the oddity of his belief, given his otherwise apparently normal behavior, to describe the odd belief as something that just occurred to him or as something he found himself saying. He may even go on to say that while

he knew that the claim would be contentious, he felt confident that he had thought it through or that a case could be made for it. Alternately one might explain the fact that what he says conflicts with the evidence by imagining that the bald man is driven by vanity. If he seeks only to have others hold a certain favorable view of his hairline he may be unconcerned with the strict truth of what he says as long as others assent. In short, there are a variety of attitudes that could explain his lack of concern for the truth of what he says. He might be self-biased, or fearful of ridicule, or, simply obstinate and unwilling to yield to the view that others hold of him. There is no obstacle to understanding such attitudes first-personally because there is no need for these motives to diminish the awareness of the person in order for them to motivate action.

This kind of prosaic irrationality reveals a person with enough of a grip on what counts as evidence that he is a candidate for belief ascription. Though this is obviously a picture of a man thinking badly, it is clear that he is still a thinker. In contrast, Davidson's picture presents a picture of a man who is unable to bring his thought in line with the demands of rationality. As such it is not clear that there is enough of a mind left in this picture to read beliefs off of. In fact, prosaic responses look so ordinary that it becomes difficult to imagine the kind of alternatives that Davidson's solution to the problem would require. Does Davidson expect his reader to imagine that the bald man, when directly confronted, just cannot process the requirement of total evidence or that he fails to understand the conflicting evidence? This is what would be needed if the mind was to be split in the way he intends. But this would be incredibly bizarre and so it seems far more plausible to suppose that whatever has gone wrong with this man is capable of engaging with his rationality and producing perhaps bias or a self-serving view point, but nothing as radical as a split-mind.

Imagine that the mind really did sequester both the requirement for total evidence and the evidence which would show that the man was bald in the way Davidson describes. That is, both of these (in normal cases readily acknowledged) aspects of one's psyche were blocked from awareness. It is difficult to imagine by what mechanism such a block could be accomplished. At best, and Davidson's account does suggest this, perhaps some form of selective forgetting could accomplish this task. While this may be enough to prevent the awareness of these items while the person is alone or not engaged with the issue, this mechanism certainly won't be sufficient to prevent awareness in a dispute about whether the man is bald. Indeed, it is almost impossible to conceive of how the mind might be able to insulate the person from these facts, when he is confronted with the word at large, while still preserving his status as a minded creature. This result should be obvious from Davidson's own account of mindedness. For Davidson, beliefs exist in, and are individuated by, their position in a rational network. Each node is connected to others by the recognition that there is a justificatory relation sustaining the connection. Further, others recognize this and this is exactly what it exploited in discourse about what is correct to believe. Other people then can be expected to raise exactly the contrary evidence and the requirement for total evidence in their challenge to the bald man. So being exiled from awareness doesn't look like a state that could be maintained, or if it could this could only occur through the most bizarre and baffling behavior of the subject. He would have to be seen as failing to understand the challenger or being unable to speak to the content of the challenge. These, it would seem, are the only ways that the awareness of these issues could continue to be blocked. None of this is to say that a person cannot retain odd or irrational beliefs in the face of challenge. But it is to note that this is usually done in familiar enough ways that are easily made sense of. A person may acknowledge how compelling his opponent's challenge is before replying that it is

not wholly convincing or may beg off and promise to think more about the issue. In this way then it becomes possible to retain the problematic beliefs without requiring that what is contrary is not present to awareness.

Thus I charge Davidson with failing to provide an interpretation since he attempts to locate the irrational phenomenon in a description that fails to conform to the constraints of interpretation. This failure to adequately locate the phenomenon can be traced to the view of mind borne from an over-riding concern to ensure that the subject be attributed true beliefs. Davidson's picture of an agent is that of an epistemic monitor of his network of beliefs. But, assuming that the bald man believes what he says means that picture can't be maintained in the face of such blatant contradiction. As such, a dramatic 'split' is posited which gives a picture of each mind being governed by rational principles. In so doing the unique role of the agent as monitor is fragmented and turned against itself. The control the agent once enjoyed over his thought has been diminished to the point of absence; he is no longer sensitive to the evidence. But, it is possible to avail oneself of a viewpoint in which epistemic standards, narrowly conceived, do not have the overriding constitutive character that Davidson's theory requires. Instead, it is possible to begin with a picture in which an agent sees various 'moves' in thought and action as compelling and attractive, without seeing a single epistemic standard as the ideal to which all moves must conform. This provides a way of understanding contradiction which grounds the phenomenon in the larger context of how one understands himself, but does not require a purely rational perspective over one's mental states coupled to a divided mind to make sense of deviations from the rational.

While it is possible to imagine a dramatic split afflicting a person's mind, it is imperative to ask if this is the best interpretation available to account for the person's behavior. I have

suggested that better interpretations exist. Because it is common to first-personally recognize propensities toward bias, emotion and error, it seems far more plausible to invoke these to explain prosaic irrationality. Such explanations ensure commonality because they cite features of personal experience which are readily recognized. Moreover such explanations are already recognized as thoughtful. This makes them better candidates for framing the explanation because their applicability to mental explanation does not require buttressing by a contentious psychoanalytic theory. By refusing to offer explanations of the subject which make his behavior explicable but require that he not be in control of himself, it is possible to provide an interpretation to which the subject can assent. This is just to recognize the role of these propensities in giving rise to one's mental states. But, these mental phenomena are not beholden to the narrow epistemic principles Davidson takes to be constitutive of mindedness.

So it seems clear that there is some tension between interpretations gained by cleaving closely to a principle of charity and those gained from a first-personal vantage point. In Davidson's example, charity led to positing mechanisms which afflicted the subject with a strange failure of normal awareness. It was suggested that there were better ways to explain the case. Imagining the subject as biased, excessively vain or self-consciously offering a rhetorical defense of his hairline explained his behavior at the expense of making him indifferent to the truth and justification of some of the things he said. If these are more plausible explanations, then charity has the potential to make irrationality appear to be outside the control and awareness of the person. In short, it tends to distort irrationality by making it pathological.

2.4 WEDGWOOD'S RELIANCE ON CHARITY

While charity can make the irrational pathological, it can also over-rationalize. Wedgwood argues that because irrational behavior can be shown to be pathological, this can be used to show that non-pathological behavior reveals the existence of normative facts. I show that moves of this type threaten to distort normal behavior by making it aim exclusively at the rational. In this section I show that Wedgwood's constitutive account of the metaphysics of norms is also premised on a distorted view of the human being. I then go on to show that replacing charity with a principle that respects first-personal experience solves both kinds of distortion.

Davidson's claim that the principle of charity is the governing principle for attributing a mind to a subject amounts to the claim that the mental is, by its very nature, rational. Indeed put this way it is easily seen as an extension of his better known dictum "belief is, in its nature veridical".(Davidson, 2002a) This view of mind has had widespread effects in philosophical methodology because it licenses the claim that mental states will, in general, conform to the strictures of rationality. From there one is able to uncover rational governing principles and argue that real and robust standards for thought or behavior can be read off behavior.

Wedgwood offers a metaphysics of normative properties that is premised on understanding those norms as operative on the subject in the way he is disposed to think. He operates with the idea that all concepts and attitudes can be associated with rational standards. In support of this he assumes a symmetry between concepts and attitudes and then argues that a concept is individuated by the rational role it plays in thought.(Wedgwood 162) Wedgwood draws evidence from responsiveness to reveal the rational role of the concept in question. It is

here where the reliance on charity is most needed. In order to attribute concept possession to some subject he must be able to read rational responsiveness to the concept's semantic structure off from the subject's dispositions. But the subject's dispositions to respond are likely to reveal at least some measure of irrationality. How then can the existence of normative properties, which can only be seen in conformity to rational standards, be predicated on the basis of an irrational disposition?

Wedgwood is looking for an explanation of how a thinker's rational dispositions might be distinguished from whatever gives rise to his irrational behavior. He begins by noting that while some of a subject's dispositions to reason with a given concept will be irrational, it becomes absurd to try an imagine a subject who was disposed to always reason irrationally with a given concept. To imagine this scenario would be to imagine conditions in which it was inappropriate to attribute possession of the concept to the subject, not to imagine that the attribution could be made on the basis of such a flawed disposition to respond. So, each subject who is a candidate for concept possession must be disposed to show sensitivity to the rational role of the concept. Thus, there must exist some rational disposition to use the concept correctly. Indeed Wedgwood argues that it must be possible for a perfectly rational person to possess any concept that can be identified. (Wedgwood 168) But if this is true, argues Wedgwood, it shows that attributions of concept possession are to be made on the basis of rational dispositions not irrational ones. Because irrational dispositions are associated with a failure of concept possession, Wedgwood can conclude that it is the rational dispositions that determine concept possession.

This argument does not imply that all concept possessors use the concept correctly. Wedgwood argues that rational dispositions are manifested *ceteris paribus*. As such, when

conditions are abnormal rational dispositions may be masked or blocked. The situation that rationally should trigger the response may be unusual enough that it evades identification by the rational sensitivity or other unusual features of the situation may trigger responses that mask it. Wedgwood notes that the unusual features may not be isolated occurrences. He imagines it is possible for entire communities of people to live under abnormal circumstances for long periods of time. Wedgwood contends that rational dispositions will be those which are best understood by an interpreter. Because he imagines that all human sensitivity tends toward rational responsiveness, the fact that an interpreter readily understands his subject's responses is taken as evidence that the disposition being displayed must be rational.[12] In this way the epistemology of interpretation provides a guide to the metaphysics of the normative properties.

The identification of rational standards implicit in the subject's responsiveness licenses Wedgwood's conclusion that normative facts are real natural facts that are not reducible to responses alone. Stated this way it becomes clear that Wedgwood's argument is dependent on an assumption of charity. If some forms of responsiveness cannot be interpreted charitably, and cannot be seen as pathological or abnormal, then Wedgwood will have to acknowledge that some concepts and attitudes cannot be associated with rational standards. This need not be fatal to his argument. He could argue that norms only govern some central set of concepts and attitudes. But this response significantly complicates his project. The reason for this is that it requires him to make a principled distinction between those concepts and attitudes that he takes to be central and those that are not.

---

[12] Wedgwood argues that certain rational principles are essential to the capacity to have a given attitude. He then goes on to say that this is reflected in making sense of a person as having that attitude. "..it is unsurprising that we find manifestations of this disposition particularly intelligible and easy to understand, especially in contrast to more unusual dispositions… attitudes that cannot be explained by appealing to these universal dispositions will seem puzzling and harder to explain." 2007: 238-239

There are two sources that he might draw on for that distinction. First, he could argue that the norms themselves, when properly examined, show themselves to only govern those concepts and attitudes which are central in the right way. But this type of response threatens to upend his constitutivism because it would require an independent account of the norms that was principled enough to draw this distinction. Since the only account of norms he has offered has been premised on an analysis of dispositions, it is not clear that there is any such account of normative principles available. After all, his central argument for the existence of such normative properties takes dispositions to provide a guide to them. And if it were possible to offer an independent account of them, presumably this style of argument would be unnecessary. Unless he can distinguish rational from irrational dispositions in interpretation, Wedgwood needs to offer a new route to the analysis of normative principles.

An alternative strategy then looks more attractive. Wedgwood can attempt to argue that the epistemology of concept and attitude attribution allows for the distinction to be made. However, this strategy will require pointing to some feature of the attribution and arguing that this signals that the concept or attitude is deviant in the way required. Wedgwood has relied on the idea that some attributions are more easily made than others and argues that this shows that those are the rational and central cases. As such, his defense against irrational dispositions will be vulnerable to a demonstration that ease of attribution does not reveal rational governance.

So it would make trouble for Wedgwood's argument if it could be shown that there were a set of easily identified motivations that would allow for the attribution of some concept or attitude without also allowing that a set of rational standards could be read off of the attribution. This would undermine his reliance on charity as a route to rational standards because it would be to deny that there was some central set of concepts or attitudes that revealed what

standards were in play. Instead, it would be seen that concepts and attitudes can exist and be recognized without associated rational standards.

It seems clear that the most fruitful place to look for such dispositions is in those that would be used to attribute emotional attitudes. Wedgwood treats concept and attitude possession as on par.[13] That is, both are attributed on the basis of dispositions and these dispositions reveal the rational standards that ideally govern the use of the concept or attitude. However, while Wedgwood discusses concept possession in great detail, he says comparatively little about attitude possession. Further when he does mention attitudes he frequently makes use of examples which treat them as evaluations of some stimuli. His example of an instance of admiration is notable in this regard – it is a reaction to the value of a piece of music being played well.(Wedgwood 236) However many attitudes do not seem to function in this way, so this domain provides a rich vein to mine for counterexamples. For example phobias do not respond to an acknowledged lack of danger and rage will often lead to a loss of rational control. So, many attitudes that are easily interpreted do not seem to reveal a rational underpinning.

The place where Wedgwood's treatment of attitudes and concepts appears most obviously flawed is in his willingness to treat attitude possession in the same way that he does concept possession. He notes that a perfectly rational being must be capable of possessing any concept which might be identified. Further this being will display no irrational tendencies and so the attribution of the concept will be on its best footing in such cases. While this seems innocent in the case of concept possession because concepts are individuated by the rational role they play

[13] For Wedgwood both concept and attitude attribution is done on the basis of seeing a rational structure in dispositions. "..the sort of disposition that a thinker must have, if she is to possess that concept or be capable of that type of attitude, is a disposition to use the concept in ways that the principle specifies as rational." 2007: 164 It follows that a perfectly rational being should be capable of possessing any concept or attitude type that could be identified. "..it seems that any concepts that you have could be possessed by a perfectly rational being who had no irrational dispositions at all" 2007: 168.

in discourse, there is no similar individuation principle for attitudes and so one might suspect that they are not always so amenable to rational reconstruction. Further, there are some attitudes that it seems wrong to attribute to a perfectly rational creature. This is particularly true of attitudes-of rather than attitudes-that. Attitudes-that take a specific proposition as the object of the attitude. For instance, one can be proud *that* his technique saved the company thousands of man-hours and one can be afraid *that* he will be bitten by a snake. Attitudes-of take more nebulous and generic situations as their object. Generally an attitude-of takes a term as its object, rather than a proposition. For instance, one can exhibit a fear *of* snakes, but this cannot be reduced to a fear *that* all snakes are dangerous or even *that* one might be bitten by a poisonous snake. Fear *of* snakes persists in the face of knowing the particular snake one is confronted with is benign. Rather, it is something like the alien serpentine body that elicits the anxiety and disgust.

Phobias are perhaps the most easily recognized case of attitudes-of but they do not exhaust the phenomena. There are a number of attitudes that take relatively nebulous terms as their objects and so reveal a disposition to respond to every situation of a given type with the characteristic response: they are not evaluative. Apathy can be described as a generalized lack of interest in the business of living or in some particular activity. Similarly both paranoia and euphoria are generalized attitudes toward the totality of one's experiences; they are not restricted to a particular domain. They don't apply to a well-defined set of objects. They are so promiscuous that they can attach to anything. It can also be the case that attitudes which are most readily recognized as attitudes-that may appear in more generalized forms. For instance one can be envious *that* Rick got the promotion but it is also possible to be envious *of* Rick more generally. The highly general nature of such attitudes makes them good candidates for irrational

attitudes. Something so powerful, yet so easily elicited, resists being characterized as a rational evaluation of some feature of the situation.

In addition to the attitudes which take nebulous objects there are also attitudes that seem to wear their lack of rationality on their sleeve. That is, for some attitudes it may be that there are simply no circumstances in which it would be 'correct', in the sense of its forming a part of an ideal plan, to have the attitude[14]. Rage seems to function in this way. When one is enraged there is an aspect of the state which opposes rational control over one's actions. To be enraged is, necessarily, to act in a way that does not take account of the way that one ought to act. This is not to say that there are not cases in which it might further one's goals to become enraged. For instance, intimidating others is more easily accomplished when enraged and so it might make a kind of strategic sense to bring it about that one become enraged to facilitate the effort to intimidate. But this looks like a pale imitation of a rational attitude. This follows from the structure that rage would have to exhibit for it to be rational. It would have to be rational to adopt an attitude which reduced one's ability to act rationally only when doing so could be counted upon to advance one's rational goals. But this means that in rationally deciding to abandon oneself to rage, one risks losing the prize that he aims at by allowing his behavior to be dictated by something other than rational principles. So, rational rage would depend on a kind of good fortune in securing its goal. One would have to hope that the loss of rational control did not diminish the ability to meet his goals.

It seems likely that there will be real difficulty in coming up with scenarios in which the rational standards governing the use of attitudes like schadenfreude, jealousy, mean-spiritedness or pettiness will become apparent. All of these attitudes appear, like rage, to be prima facie

[14] A concept or attitude featuring in some agent's ideal plan is taken as the paradigm case of respecting a concept's or attitude's rational role in Wedgwood's account.

defective. Identifying one in a person could be taken as grounds for describing the subject as irrational. Indeed these seem even less effective at securing one's goals than rage. However, even if one assumes that there is a set of core rational dispositions that point the way to the standards of rational use for these attitudes, it is implausible that this set of dispositions provides for the core cases of attributing the possession of these attitudes to a subject. Given how difficult it is to imagine a rational end that adopting these attitudes would reliably promote, coupled with how readily recognized and attributed these attitudes are, it is likely that these attitudes can be ascribed on the basis of dispositions that are not rational but are not prevented from being readily recognized for that reason.

It is essential to Wedgwood's argument that it be possible to identify a core group of rational dispositions to respond to normative facts. It is this which shows that the normative facts are causally efficacious and supports his realist claims. Thus if Wedgwood is unable to offer principled distinctions between rational dispositions and various disruptive influences that might distort their aims it will threaten his argument that there is a rational core to be found which supports the realist view of norms.

He offers an argument to show that, despite appearances to the contrary, it is possible to see a rational core to the dispositions witnessed even in those cases in which the dispositions apparently fail to display the relevant sensitivity. The key to dismissing the irrational dispositions lies in seeing irrational disruptions as reactions to abnormal circumstances. The abnormal response 'masks' the normal rational disposition to respond. Showing the principles governing human thought naturally and rightly extend beyond the merely rational requires overcoming this 'masking' argument. If this argument were successful it would allow Wedgwood to claim that

the responses being masked revealed those 'central' concepts and attitudes that are governed by normative principles.

I want to note first that there is already something suspicious about the strength of the constitutive claim Wedgwood advances. Taken very strongly, that all instances of irrationality can be explained by invoking abnormal circumstances, it cannot be falsified. Absent a rigorous definition for what constitutes abnormal, and normal, circumstances it will always be possible to point to some feature of the situation and describe it as the abnormal circumstance in question. Indeed it is exactly this kind of strategy that was used by Davidson to show that the irrationality of the bald man did not threaten the general thesis that charity was the right way to proceed in interpretation. [15]

Since it is always possible to frame a specific case in terms of disruption of the normal sensitivity and so argue that the anomaly can be dismissed without affecting the general thesis, the invocation of abnormal circumstances must be principled. Otherwise, it will just appear as empty redescription.

Evidence does not wear its credentials on its sleeve. For something to be counted as evidence it must be interpreted as evidence. Where the theoretical structure that it is being incorporated into purports to describe the structure of something abstract, such as the principles governing mindedness, the interpretation itself will be the most robust piece of evidence that can be appealed to. Since there is little possibility of gaining direct empirical contact with the mind that would reveal unambiguously whether the interpretation is correct, it is the interpretation's ability to fit with other data points that will determine its strength and applicability. Because a

[15] The invocation of abnormality here is, to be sure, not quite the same as the invocation of a split-mind. But, the outcome is the same in both cases: a merely causal mechanism is used to explain the irrationality.

starting assumption is just such a datum, there will be a tendency to favor interpretations that cohere with starting assumptions as opposed to those that don't.

It seems clear that if it could be shown that a situation was in no way abnormal yet the attitude a subject possessed did not display rational sensitivity this would be a decisive piece of evidence in favor of overturning Wedgwood's starting assumptions. Can such a case be shown? It certainly seems possible since a set of criteria have just been outlined that a situation would have to meet to warrant abandoning the staring assumption. However concealed within that set of criteria are terms that are loosely defined and so ambiguous enough to allow virtually any empirically described scenario to fail to satisfy the criteria. The relevant terms are 'abnormal' and 'failure to display rational sensitivity'.

The problem can be made acute. Any interpretation of a person's behavior in which the subject himself is presented as acting in a way that is irrational contains evidence which would potentially damage the assumptions underwriting charity. To deflect that damage it will be necessary to mitigate that interpretation. Two options exist for doing so. One, reinterpret the situation in ways that allow the behavior to be seen as rational – at least from the vantage of the subject. Two, reinterpret to ascribe the irrationality to a defective aspect of the subject's psychology. It is the latter strategy that the masking argument makes use of.

Such reinterpretation essentially ends in a stalemate. For someone unconvinced by the starting assumptions implicit in charity, the move will seem to smack of a kind of gerrymandering. The subject's behavior will seem to be arbitrarily repurposed or reassigned so that it satisfies a conclusion favored by his opponent. Similarly, for the interpreter, who thinks that charity does provide the only method of making sense of the subject's behavior, presenting

the subject as irrational will just look like a refusal to offer an interpretation at all. How might this impasse be resolved?

In theory, direct confrontation with the mind itself would be able to resolve this dispute. If one knew the principles that were called to mind and the thoughts which prompted the action it would be possible to produce an interpretation that was not vulnerable to redescription. So for instance, if sufficiently penetrating insight were gained into the thoughts of Davidson's bald man the issue could be resolved. It might be learned that he takes mention of his hairline as an insult and so his contrary stance can be explained as an angry defense of himself. Alternately it might be learned that he simply vacillates in his opinion of his hairline. On his 'good' days he finds he takes joy in his appearance and draws attention to it.

What is most interesting about these interpretations is that they are all available first-personally. By imagining insight into the subject that was sufficiently compelling that it could stand as evidence against the strategy of reinterpretation, a kind of first-personal description of the subject was produced. Contrast this with Davidson's vividly third-personal diagnosis of the subject suffering a split-mind that he was unaware of.

This suggests that first-personal reports need to be taken seriously within the context of interpretation. There are grounds for doing so. In the first place it is already recognized, though it forms no part of radical interpretation, that special authority attaches to self-reports of one's own thought. It is taken for granted that the subject enjoys a kind of insight that no one else has. In the second place, because the narrative provided by a self-report is so well understood by others, first-personal reports can provide for the kind of commonality desirable in interpretation.

First-personal reports often specify the motivations that prompted the behavior in terms other than belief and desire. This means that they have the potential to offer descriptions that

cannot be accommodated by the principle of charity. If a person describes himself as taking risks because he was feeling great from an earlier success or if he claims it was a sense of whimsy that led to his choosing his peculiar shoes, the person invokes descriptions of his actions that do not obviously traffic in straightforward beliefs and desires. If there is no way to offer redescriptions of his behavior that don't attribute strange rational reconstructions of these prosaic ways of acting, this would count against the principle of charity's ability to offer the best interpretation.

It is generally agreed that individuals have a kind of access to their own thought that others do not enjoy. Further the issue here is a subject's dispositions to think in one way or another – that is in accord with the rational principles or in some other way. The principle of charity is biased toward reinterpreting first-personal reports when they describe self-consciously irrational thought. Instead I propose according first-personal reports a greater weight in interpretation. I want to treat them as a having a presumption of at least equal evidentiary weight in comparison to rational reconstruction. No argument from radical interpretation shows such claims to be unreliable. Further, both interpretive claims and those that cite first-personal experience have a role in grounding attribution. As such there can be no argument for ruling in advance that claims of one kind or the other should be favored in a dispute between them. Ensuring this requires that they be treated, in general, as equally compelling pieces of evidence.

It seems that if claims of this type are accorded greater evidentiary weight it is possible to break the impasse that the possibility of redescription creates. Further, the replacement of charity for a principle that can take account of the way that first-personal reports operate, produces a principle that only gains, and doesn't sacrifice, any of the explanatory power of the previous method. Since beliefs and desires are generally taken to be apparent to consciousness, taking first-personal reports seriously will still allow any purposive or rational behavior to be explained

by citing belief and desire. However, it will further allow other states, such as whim and emotion, to be offered as potential explanations of behavior.

2.5 REPLACING THE PRINCIPLE OF CHARITY

I have shown that a strong principle of charity in conjunction with an empirically driven view of evidence will introduce distortions of two basic types. It will rationalize non-rational behavior including some emotional responses. And, it will pathologize that which it cannot over-rationalize. It will over-rationalize because it will seek to explain purposive behavior by citing beliefs and desires which together provide the context for the behavior to appear rational.

In some sense this is all it can do. Because the mind of another is not observable, its workings and mechanisms can only be theorized about and, on the empirical view of evidence, all theorizing that is efficient in providing explanation is equally valid. But, radical interpretation makes an exception for belief and desire. These are provided a priori as the mechanisms of mental action. So the theory must take it for granted that the mind works by way of these mechanisms and attempts at explanation which invoke these structures are a priori better than mental explanations that do not. This leaves explanations in terms of emotion, whim or other first-personally described irrational transitions in thought at a competitive disadvantage when facing off against belief desire explanations. To make matters worse, citing these other kinds of motivating force will require that one invoke non-rational explanations for the behavior and such explanations are a priori suspect because of the principle of charity.

This last point also explains why there will be a tendency to make pathological that which it cannot rationalize. Under the constraints of radical interpretation the kinds of theorizing which will be most suspect (at the highest competitive disadvantage relative to other accounts) will be those which violate rational principles and invoke highly theoretical terms in explanation. Such accounts violate the constraints of radical interpretation because they show that there are ways of understanding a subject as minded without the constitutive ideals set by the theory. This is where the example of the bald man becomes so instructive. Since Davidson can't find a way to attribute the obviousness of the man's baldness alongside his avowals to be hirsute without violating the constitutive ideals set by the theory, Davidson is forced to describe the phenomenon in ways which effectively deny the mindedness of the subject. In this case the mindedness is doubled and each half is able to take up some of the attributions that Davidson wants to make. But this comes at the expense of seeing the person as in control of his mind. Davidson must posit strange losses of awareness and other mechanisms that are outside of agential control to explain the unique way this subject thinks. In this way Davidson is forced to offer what might be called strongly third-personal explanations – he invokes mechanisms that can only be seen from outside to characterize the person. It seems the best Davidson can offer is a diagnosis and not an interpretation of the subject.

There is nothing wrong with mental diagnosis as such. There are, no doubt, many circumstances where it is appropriate. But it should be noted that such explanations minimize the mindedness of the subject in the accounts they offer. This can be seen in the way that the subject's thought is obscured by the explanation. It is no longer possible to understand the subject in the way he understands himself. What has been offered is an explanation that attempts to supplant first-personal accounts of the subject. The account that the subject could offer is on

less certain evidential footing than the account the interpreter offers because the latter is based on observation. And, the account that the subject is likely to offer will be seen as irrational by the interpreter and the charity constraint militates against accepting such an account of the subject's mind. The challenge then is to make first-personal accounts competitive with rationalizing explanations while still allowing third-personal explanations when appropriate.

Because the observational evidence has been given preference in interpretation and charity prefers interpretation which makes the subject turn out to be rational, there is little opportunity for first-personal phenomenological descriptions which cite any irrational transitions in thought to compete with accounts that rationalize the subject's behavior.

One solution to this issue would be to simply credit first-personal phenomenological accounts of the subject's thought with more evidentiary weight in every case than observational evidence interpreted through the principle of charity. This threatens to be too strong because one must admit that there are times when it is right to override phenomenological accounts of the subject.

The issue is just that third-personal descriptions have been accorded too great a role in theories of radical interpretation. Allowing the interpreter's own standards to determine the attribution of mental states to the subject is already to build in a bias toward the third-personal. But this is compounded by making all of the legitimate evidence observational because this necessarily minimizes the subject's (as well as the interpreter's) own phenomenological experience. Taken together, the operation of both of these principles serves to act as a trump card which can overturn any account that offers a description of the subject's own thought as it unfolded specified in terms that invoke something other than belief and desire or which cite irrational transitions. This will be possible so long as a competing description of the kind favored

by the theory cannot compete with the first-personal description. In practice this means that first-person authority has no standing; it either agrees with the interpretation, and so is accorded weight on that basis, or it falls to an observationally based description.

It seems possible to correct this. The goal is to accord first-personal authority some weight while allowing that third-personal reports also have legitimacy. In order to do this I propose a different set of constraints on interpretation. First, I propose to expand the evidential base to treat first-personal phenomenological reports as legitimate evidence[16]. It should be treated as on par with the observational evidence. This change, in turn, necessitates a change in the principle of charity. The reason for this stems from the nature of first-person self-reports. These are often given in narrative style which cite occurrences, emotions, vivid pictures and other mental phenomena as that which motivated action.[17] This means that reports of this kind will emerge from a much broader ontological conception of mental states than the principle of charity can account for. So the proposal is to take as given the evidentiary value of phenomenological reports. This will mean that one must also license the terms employed by the phenomenological report as real. That is, they cannot be treated as dubious by comparison to beliefs and desires.

Thus the proposal to treat phenomenological reports as having equal evidentiary weight will require alteration of the constraints on interpretation. The constraint provided by the principle of charity is tailored to account for a view of mind in which belief and desire are the

[16] This should not be read to say that all first-personal phenomenological reports must be given by the subject himself. I expect to be able to draw a distinction between the kinds of report that could be produced by the subject reporting on his own conscious experience and that which could only be produced by citing theoretical mechanisms hidden from the subject. In principle there is no obstacle to someone other than the subject trying to give an account 'from the inside' of what it is like to be the subject. As such when I talk about finding a place for first person authority I do mean to include the distinctive kind of weight that such reports receive when they are self-reports, but I also mean to try and include a role for this distinctive 'internal' way of understanding the subject. I propose putting this kind of understanding at equal evidentiary weight to the purely observational reports.

[17] See Hursthouse 1991.

basic drivers of action and observation is the only legitimate from of evidence. If the constraints on interpretation will now need to be altered to accommodate the terminology in phenomenological reports, the new constraining principle should presuppose the existence of these mental phenomena in the same way that the charity constraint presupposes the existence of beliefs and desires. This means that, at minimum, it should recognize the existence of emotional and kindred states without presupposing that these can all be necessarily decomposed into sets of beliefs and desires. This follows from the fact that typically such states are talked about in ways which do not treat them as simply interchangeable with the language of belief and desire. It should also recognize the existence of whim and its motivating potential, as well as the role that habit and know-how play in generating thought and behavior. And the way that thought typically 'flows' in patterns of association without being beholden to rational principles. I loosely term all of these kinds of mental activity 'processes'.

Now it seems that there will be no way to incorporate a role for these kinds of thought while retaining an overriding charity constraint. This is because building them into the constraints on interpretation is to treat the principles that they operate according to as, at least partly, constitutive of thought. Further because there is no presupposition that these processes can be decomposed into those that are governed by purely rational principles they must be seen as operating alongside and interacting with the subject's more rational states. Both of these ways of seeing the constraints show that what is constitutive of thought is not purely rational principles and so the best interpretations of a subject will sometimes be those that make him explicable by citing these arational principles. The scope of the principle of charity has been reduced by acknowledging the action of other principles of thought.

This should not be read as proposing that only those phenomenological reports that are issued by the subject about himself are to be accorded this evidential weight. Recall that one role that the principle of charity played was to establish sufficient commonality between subject and interpreter for interpretation to take place. Similarly the constraint I propose must also provide for commonality. Without this role it will not be able to get traction on the behavior. One will be told that certain kinds of accounts must be accorded equal evidentiary weight, but no method for generating accounts of that type will have been provided. As such I propose that the interpreter use his own experience with the unfolding of his mental life, including his own experience in providing phenomenological self-reports, to ground a description of the subject. This proposal is more powerful than the principle of charity because one can recover descriptions that the principle of charity would be able to offer by using it.

To see this consider that while phenomenological self-reports can extend beyond citation of mere belief and desire in accounting for some thought or behavior, not every report of this kind will offer an account framed in those other terms. So, since beliefs and desires are objects of conscious experience that are assumed shared by both interpreter and subject, any rational behavior can still be explained using the kind of terminology that was central to the principle of charity. That is, the interpreter still has recourse to use beliefs and desires to account for the behavior of the subject and will have the ability to describe the behavior as unfolding according to rational principles. But, the interpreter also gains the ability to describe the subject using other terms and to have these descriptions carry the same weight as those that were previously given. By recognizing the authority of self-reports one is bound to recognize that reality of the terminology that those are given in. Recognizing the reality of these requires that they be given a role in describing other minds and not just the speaker's own.

To be sure the ability to issue such reports does not result in the interpreter having the same kind of authority that the speaker himself does. That is not what is being claimed. Rather, the thought is that one recognizes the authority of first-personal reports. In doing so, one also recognizes that the form and content of that kind of explanation, as it is typically given, cannot be recognized by the principle of charity. The interpreter must make choices about whether to invoke emotional and similar states in an explanation or stick with beliefs and desires and cite rational principles in accounting for how some behavior is generated. This new principle doesn't presuppose that the best explanation will also be one which rationalizes the behavior.

2.6 CONCLUSION

Now that a replacement for the principle of charity has been sketched out it is useful to see how it impacts the issues raised by Davidson and Wedgwood.

A more plausible way of accounting for the bald man's engagement with the evidence is to see it as motivated by something other than a concern for the truth. And here it is easy to think through all of the self-serving kinds of things he might say. Moreover, it is possible to do so in a way that allows one to make sense of how it is with the person as well. This is an understanding of an agent making use of his reasoning to come to a conclusion that he antecedently favors. And making sense of this is no great leap; it is just what is done all the time in charging a person with bias or accusing him of being self serving. Instead of considering what is sanctioned by the conceptual connections to ideas, or what is required by the standards of epistemic excellence, he considers what might be said in favor of a proposal and in so doing can be expected to reason in

ways that are not captured by the very strong constitutive ideal of rationality that Davidson advances.

Replacing charity as a guiding principle of interpretation also has consequences for Wedgwood's argument. When first-personal reports are admitted as legitimate it can no longer be taken for granted that interpretations will tend to reveal rational sensitivity. Instead it will be seen that subjects can be interpreted as motivated by whim, emotion and a variety of other states not beholden to rational principles. Further subjects will be readily understandable when described in such terms. Such responsiveness need only be seen as apt or within the range of normal human reaction to be made understandable. This requires Wedgwood to give further distinguishing criteria so that the normative principles he argues for can be read off of the entire range of observed behavior. Without some further method of distinction, a study of dispositions will reveal a range of normal motivating principles but it won't be able to specify which of them are rational.

If radical interpretation is to offer an explanation that relies on the way one understands himself to bring out the right way to understand the subject, then it should only rarely offer the kind of strongly third-personal interpretation Davidson gives. This strategy only explains by denying the subject's agency and, in so doing, ensures that any identification with the subject is impossible. It is possible to incorporate tendencies toward irrationality, of a limited sort, into an interpretation of oneself. The real distinction between the first and third-personal vantage points just seems to be the extent to which the more extreme cases can be maintained. But, there are grave difficulties involved in a person taking seriously the idea that their mind was 'split' or having that description serve as a way of understanding of how it was with the person at the time. These ascriptions are too strong to serve as interpretative aids because they don't

describe an agent. What is needed is some way to give an adequate account of how prosaic forms of irrationality can be interpreted and reliance on the strong constitutive ideal of charity will not serve this purpose.

3.0 QUALIFYING CHARITY: SHOWING THE RISK OF DISTORTION THAT DECISION THEORY INTRODUCES TO INTERPRETATION

3.1 INTRODUCTION

Interpretation of an agent's behavior is a powerful philosophical tool often invoked to show that there is no problem accounting for some set of rational standards applying to an agent. It asserts that it is in the nature of certain attributed states that such standards strongly apply. One can't both be considered a believer and be indifferent to the rational standards that govern belief because to ignore those standards is to fail to be a believer. However, there is a risk in this strategy. The threat comes from interpreting the totality of the subject's behavior as underwritten by rational considerations. This threat is particularly acute when there are no principles for determining when rational explanation is apt as opposed to explanation in terms of the subject's non-rational states. In this respect David Lewis was something of a pioneer. He recognizes the need to explicitly qualify the role of charity in his method of interpretation. However, he seems unable to see that charitable assumptions inform his use of decision theory and so these go undiagnosed and unqualified. This essay aims to put that oversight right and in so doing assert the need for qualifying charitable assumptions in interpretation, whatever form they take.

I argue that Lewis's method is an improvement over other kinds of interpretation because he is explicitly concerned to give an account which includes a plausible understanding

of the way that an agent understands himself. In so doing, Lewis seeks to put an understanding of thought that is not tightly constrained by the constitutive ideal of rationality, at the forefront of his theory. This allows him to explain deviations from the rational ideal as the result of personal characteristics such as hasty reasoning, self-serving rhetoric, or having an incomplete knowledge of irrational inferences. [18] While this is an improvement over some interpretivist theories of mind, I argue that his theory is still problematic because it is too tightly constrained by decision theory.

Decision theory functions, in Lewis's account, as a mechanism for explaining behavior. Although he sees reason to weaken the constraining power of the constitutive ideal of rationality in explaining thought, he fails to apply similar reasoning to behavior and so decision theory's constraining power is left too strong. As a result of this tight constraint then, Lewis introduces a problematic asymmetry into his theory. The effect of this is to reduce the explanatory resources available to account for behavior. I show that a great deal of interpretive work is done to secure decision theory's applicability to behavior and this is the place for principles which qualify the role of charity to operate. The work of John Broome is also focused on critiquing some of the same features of decision theory that I identify as problematic. I use his criticisms to show that Lewis must incorporate explanatory resources not contained within decision theory to provide a plausible and nuanced account of a subject's behavior.

This paper is structured as follows. Section 3.2 provides an account of the unproblematic portion of Lewis's theory - the principle of charity. Section 3.3 introduces the problematic element of Lewis's account – the rationalization principle – and highlights the features which

[18] Contrast this with Davidson. Davidson's strong commitment to the constitutive ideal of rationality forced him to explain fairly prosaic cases of irrationality by positing brute causal mechanisms. His explanation of the man who believed, against the obvious evidence, that he was not bald invoked a causally induced 'split-mind'. The irrationality was then explained by one 'half' manipulating the awareness of the other. Davidson 2004a.

result in an asymmetry. Section 3.4 shows the asymmetry over-rationalizes behavior in Lewis's account. I then argue that that the actual practice of interpretation outruns the resources provided by Lewis's theory. The final section outlines some work on decision theory by John Broome with an eye to modifying Lewis's account. An examination of Broome's account provides a clear picture of how Lewis's theory needs to be supplemented.

3.2 PRINCIPLE OF CHARITY

Lewis provides a theory of radical interpretation that has a mechanism for making sense of thought which is less than perfectly rational. This mechanism goes some distance toward providing a theory that can bring a greater array of explanatory resources to bear in accounting for the subject than previous accounts of interpretation which sought only to rationalize him. [19] What allows Lewis to do this are two qualifications to the way that the principle of charity is normally understood. These make sense of the way that a person might self-consciously come to have or maintain an irrational state. This can be done in a way which both explains the phenomenon and which serves as a description of the subject that seeks to understand him in the way the he understands himself. Lewis's view is that, ideally, an interpretation should allow one

[19] Davidson asserts that an ideal interpretation would be one that the subject himself could self-ascribe. However he then goes on to offer interpretations of irrationality that only function because the subject is unaware of the irrational state. Lewis doesn't ask for anything quite this strong but he still recognizes that it is important to have the subject's own perspective form an integral part of the interpretation. While it is true that some very strong forms of irrationality are, in principle, not self-ascribable, it is often true that weaker forms of irrationality can be readily self-ascribed. It is also true that subjective descriptions of the transitions a subject makes in thought are vital in making him intelligible. Since strong forms of irrationality are radical enough to threaten the idea that there is an agent in the picture at all (and so not properly the object of interpretation), it seems as though the best interpretations will be those that invoke explanations that the subject could, in principle, self-ascribe. That said, there are many non-rational obstacles to complete self-awareness and so not every explanation of this type will in fact be self-ascribed.

to understand the content of a subject's thought "in the way that he could express it in his own language". (Lewis 108)

Lewis's paper begins with an interpretive challenge. For some arbitrary person that a full physical history is given for, call him Karl, the challenge is to use this to provide an interpretation of his mentality that makes sense of Karl's motives and reasons. That is, Lewis wants to understand what Karl believes, desires and means. But there is more; Lewis claims that he is searching for an account that describes these facts about Karl from his own perspective as well as an account that describes how these attributions would be parsed from the standpoint of the interpreter. So, Lewis seeks to provide his reader with both a first-person understanding of Karl and an understanding of him from the third-person to which he may or may not be able to self-ascribe. The end result is an understanding of Karl framed from his own perspective coupled with some way of attributing states to him from an outside perspective.

There is a sense in which the project outlined by Lewis is much more ambitious than others in the neighborhood. For instance, while some of Davidson's work is concerned with making sense of a person's speech behavior when it is taken for granted that he speaks the same language as that of the interpreter, Lewis is interested in a much more radical form of interpretation. He wants to know how it is that the physical facts determine the psychological and semantic facts about his subject. He doesn't assume that he can tell in advance, for example, that the sound a person makes is an instance of throat clearing rather than, say, speech. In what follows I will assume that Karl's language is sufficiently well understood that the interpretation need only make reference to brutley physical facts about him when they are appropriate. This is just to assume that I am in the enviable position, from Lewis's perspective, of having a greater stock of facts from which to derive my analysis than the position that he begins from.

Lewis sets his project as one of coming to understand Karl. And he specifies that coming to this understanding will include: understanding the truth conditions of Karl's sentences as specified in the interpreter's language, as well as, coming to have a sufficient semantic grasp of Karl's language that one can understand what his sentences mean to him. Lewis then specifies a set of principles that act as constraints on the interpretation. He does this while acknowledging that there does not seem to be any single method of interpretation which will produce fully determinate and satisfactory interpretations of every subject. Indeed, he acknowledges that the method of interpretation is something of a balancing act where the interpreter tries to balance sets of competing demands in such a way that a plausible account is produced. Lewis offers three methods for producing such an interpretation, though he allows that there may be more. The first two methods divide the problem of interpretation into sub-problems which are then solved in sequence to produce an interpretation. The third is described as a fall-back 'non-method' which simply seeks to balance all of the principles in play. My focus will be with the fairly strong way in which all of these methods constrain the interpretation of behavior so that it conforms to the ideals given by decision theory while highlighting the way that the standards of rationality are relaxed in the interpretation of thought.

Although all of the principles have some role in determining an interpretation, many are taken up with the linguistic issues that attach to the project of figuring out which of the sounds and gestures Karl makes are purposeful and so can convey meaning. Since that is not part of the particular project at issue here, I will be focusing on those principles most pertinent for making sense of a person's thought, speech, and behavior; when it is taken as given that the things he says are already intelligible. Thus I will focus on the principle of charity and the principle of rationalization.

The principle of charity is a constraint on understanding Karl's thought. This principle exists to ensure that the ascriptions to Karl contain those beliefs and values that he ought to have, given a common sense theory of persons that Lewis assumes everyone shares. Lewis remarks that an idealized version of this principle might simply ascribe the interpreter's own set of beliefs and desires to Karl. On the assumption that the interpreter takes himself to be rational, this ascription would attribute fully rational beliefs and desires to the subject. However, he recognizes that such an ascription would fail to take into account for the way that Karl's differing life experiences may have shaped his beliefs. Since a set of beliefs and desires cannot be assumed to be common between persons, so some other commonality must be found that is capable of grounding that interpretation of Karl. What is needed is a method for specifying how Karl's background experience has come to shape his current outlook. Lewis describes this as a way of incorporating the common-sense-theory-of-person's view on training and life experience into the interpretation of Karl; to the extent that the common-sense theory has something to say about such matters. Thus, unlike other methods, Lewis's account will not always begin by counting his subject 'right in most matters' by default.[20] But, if the common-sense theory of persons does not include an account of what the effects of training and life experience are, the interpreter must fall back on idealized principles of attribution that present him as fully rational.

Lewis argues that the common-sense theory of persons can be made into a practical, though non-idealized, principle of charity by introducing two qualifiers. He posits underlying systems that appear to be roughly constant across individuals but which will be responsible for producing different sets of beliefs and desires given different formative experiences[21]. He calls

---

[20] See Davidson 2001b: 196-197

[21] Lewis speaks only of Karl's 'life history of evidence', ' training' and 'systems of belief'. However, it seems that he is gesturing at a sufficiently broad set of circumstances that the term 'formative experiences' is an apt way of

method M a 'common inductive method for belief formation'.(Lewis 113) It is 'common' in the sense that the mechanism is conceived as one which would give rise to both the interpreter's (i.e. one's own) and the subject's beliefs given the differing life experiences and habits of thought of each. In this way, Lewis's account allows the interpreter to understand his subject in the same way he understands himself, though without requiring that the same beliefs are ascribed to both subject and interpreter. Instead it is the common method which can serve as the shared starting point for analysis. Similarly, U is described as a set of common underlying values that would be capable of producing, roughly, the differing sets of desires of both Karl and his interpreter, given their differing formative experiences and habits of thought. (Lewis 113)

Lewis does not go into much detail regarding the mechanism or structure of these functionally described posits. That said, it is clear that Lewis does not intend for them to be understood as sets of beliefs and desires, since he describes M as "an inductive method" and U as "a system of basic intrinsic values". It would also seem that he does not intend for them to be understood as operative in everyday reasoning and decision making. In the first place, their commonality would then preclude the possibility of disagreement between individuals who agreed on the facts of the situation; the commonality of method in this case would fail to generate contrary claims. In the second place, because Lewis attributes everyday action and beliefs to processes like deliberation; and not to posits M and U. Instead, it seems best to see these posits as mechanisms for the generation of habits of thought and modes of sensitivity from the varying formative experiences that they take as inputs. This makes sense of their commonality while preserving the possibility of disagreement on fundamental matters such as

---

capturing those features of Karl's life that are relevant to determining the particular patterns of thinking he makes use of. It is clear that Lewis is citing such 'formative experiences' in the account in an effort to provide a plausible story about why some patterns of thought that might be indentified in the subject only loosely conform to the standards of rationality.

'what counts as evidence' or 'what is truly desirable'. In these cases one can make sense out of common mechanism working to produce different convictions by taking different experiences as input. Thus I will understand these posits as working at a meta-level. As generating: dispositions to think in certain ways and sensitivities to features of one's environment.

Lewis tells his reader that these posits might be used to make sense of a person who was, say, never warned against 'hasty generalization'.(Lewis 113) Presumably then, common and easily understood errors, like this, can be explained by reference to the particular way that M, in conjunction with experience, has shaped the subject's thought. In other words, dispositions toward errors of certain types are explicable by way of this posit. If this is right then it ought to be expected that the particular patterns and tendencies displayed as a result of M will not always be available to the subject for endorsement or articulation. Since it is not a requirement on a common-sense understanding of another person that he be able to articulate, say, the Peano axioms in order to call him proficient at arithmetic, I will not be understanding Lewis as implying that every pattern of reasoning that the subject might be seen to engage in will reflect an abstract understanding of some justification for reasoning in that way. However this should not be taken to preclude the possibility of the subject being able to give an account of his thought and motivations when queried.

3.3 PRINCIPLE OF RATIONALIZATION

Adding these posits to the method of interpretation is an improvement over those methods which are tightly constrained by the constitutive ideal of rationality. This is because it allows for the possibility of making sense of an agent, who deviates from the ideally rational,

from his own standpoint. That is, on Lewis's system one can make sense of certain kinds of irrationality by showing how it would be possible for someone in Karl's position to have the kinds of dispositions that would explain it. So, an initial examination might tempt one into thinking that Lewis has managed to give an improved method of interpretation by minimizing the role of constitutive ideals through posits M and U. However, this would only be half-right. While Lewis's principle of charity is charitable in the way it interprets thoughts, his principle of rationalization seems to bring a problematic set of constraints for the interpretation of Karl's behavior. Lewis's rationalization principle brings a tightly constraining constitutive ideal for the interpretation of behavior in the form of decision theory. The result looks like an asymmetric theory of interpretation. While deviations from some ideal can be explained in thought by citing the features of a subject's history, the same cannot be said for behavior. Lewis's commitment to descriptive decision theory requires the attribution of a subject's choices and preferences directly from her behavior. Thus exactly those qualifiers which allow Lewis to overcome the problem of tightly constrained methods are lost when Lewis seeks to explain behavior. Lewis is committed to the idea that in offering an interpretation one is making sense of the subject's thought and resulting behavior. Moreover, he takes it axiomatic that behavior which cannot be fit into the model is incoherent. I mean to show that it is possible to make sense of behavior which decision theory would regard as irrational and so show that to ascribe some forms of irrationality is not yet to call the subject uninterpretable.

The principle of rationalization is designed to ensure that good reason exists for the things that Karl does. It requires that Karl's mental state be construed in such a way so as to allow him to make comparisons between various alternatives. Further, Lewis claims it should ensure that given a set of *"mutually exclusive and jointly exhaustive propositions about Karl's*

behavior; of these alternatives, the one that comes true…[should be the one] with maximum expected utility according to the total system of beliefs and desires ascribed to Karl" (Lewis 113). In addition, it seems clear that Lewis intends for this principle to be one of the strongest and least open-ended constraints on the resulting interpretation. This can be seen in his description of the principle as a way of ensuring that it will be decision theoretic principles which govern Karl's behavior. Indeed, Lewis doesn't see decision theory as 'esoteric science' but instead as "*the very core of our common-sense theory of persons, dissected out and elegantly systematized*" (Lewis 114).

Before examining the consequences of this principle it will be useful to consider how the qualifiers work in the interpretation of thought. Lewis's theory does allow one to see Karl as, say, biased in his assessment of himself or those things which are important to him. And, it is possible to recognize the source of this bias in familiar facts about how he is disposed to reason; perhaps self-interestedly. In this way the principle of charity allows a plausible account of human foible and emotion by treating at least some thought as not fully governed by rational principles. This is because it makes it possible to see how Karl could have convinced himself that he was being unbiased or that others were simply confused in their assessment of him without requiring that he is simply cut off from, and unaware of, the contrary evidence or whatever rational requirements are in play. So Lewis's qualifications bring a greater range of explanatory resources and so a more nuanced and realistic view of the subject's thought.

But this result only emerges in the interpretation by building the role of formative-experience into the non-idealized version of the principle of charity. This explains how a person might come to have some set of beliefs and values and why he might come to make use of these in thought in, perhaps, idiosyncratic ways. So Lewis allows for less than ideally rational thought

to emerge from his interpretation by applying a fairly loose rational standard to the subject. They are 'loose' in the sense that they don't constrain every description of thought, as such. In some descriptions it is possible to invoke arational mechanisms as explanation without denying that something thoughtful was being explained.[22]

That said, Lewis's account as a whole still turns out to be too strong. As I turn to the principle of rationalization it will be seen that a similar loose constraint would be desirable to have here as well, but Lewis fails to provide one. This means that he can't capture the plausibility of the idea that a person may act in ways that are just as idiosyncratic as the way he thinks. Indeed the asymmetric treatment of thought and action means that it won't be possible to see less than fully rational actions as a manifestation of thought with the same deficiencies. The overly strict principle of rationalization will rule out these idiosyncrasies in behavior even as the theory allows them in thought.

The principle of rationalization is concerned to provide preferences that explain the subject's behaviors. It does this by treating interpretable behavior as motivated, in all cases, by the subject's desire. As such, the principle of rationalization has, as its sole focus, making sense of behavior through the lens of decision theory. It is this treatment of behavior that introduces the asymmetry because it does not admit of qualification in the way that the treatment of thought does. While M and U can be used to account for deviations from the ideal in a subject's thought, no similar qualifier is available on the invocation of decision theory. So, even if charity allows the interpretation to make sense of a subject's odd thoughts, the interpreter is bound to find the subject conforming to the predictions of decision theory when accounting for his behavior.

[22] While such posits are of obvious value in explaining error, it is not as well recognized that they will also be useful for explaining success. The sensitivity of the chess-master, the intuitions of the experienced doctor and the ease with which the engineer make abstract ideas vivid by invoking mechanistic metaphors will all tend to be over rationalized in explanation without these posits.

Further, conforming to decision theory will require rationalizing the subject's behavior. It just won't be possible to see the subject as irrational or arational through this lens.

In its current form the rationalization principle will prevent plausible interpretations from being considered and will have a tendency to eliminate the kind of detail that adds predictive power and understanding to an interpretation. This constraint will result in a picture of a subject who is highly rational. But this is purchased at the cost of seeing him as subject to emotion and irrational tendencies. Simply put, decision theory is unable to ascribe irrational preferences and so will be unable to offer interpretations of irrational behavior and emotional states. The issues I raise show that better interpretations would be produced by making use of posits like M and U in the interpretation of behavior. The upshot of such a change would be that this modified method of interpretation could predict rational behavior with the same efficiency as decision theory but would gain additional explanatory resources as well. I conclude that it is not possible to see decision theory as a common-sense theory of persons devoid of esoteric science. John Broome's research on decision theory echoes this conclusion. His work on the 'granularity' of discrimination reveals that decision theory has little to say about how preferences ought to be individuated. His solution to the problems raised by this is to only allow fine-grained discrimination in instances where such discrimination can be justified by the context of the interpretation. This solution comes with a cost. Broome is forced to admit that not all behavior that is intelligible can be made sense of through the lens of decision theory. It is exactly this feature of decision theory which Lewis cannot incorporate into his account. Broome is admitting that the understanding of decisions theory he is forced to adopt could not be the 'very core of our common-sense theory of persons'. At best, decision theory can only model some kinds of intelligible behavior accurately.

3.4 RADICAL INTERPRETATION AND DECISION THEORY

In order to see what goes wrong by relying on decision theory as a method of interpreting behavior, it is important to begin with a clear picture of the way it functions in Lewis's method.

Lewis's use of decision theory is designed to allow an interpreter to read preferences directly off of behavior. That is, he uses decision theory to reveal the preferences that a person has. The idea is very simple. If one observes Karl going into the grocery store and choosing apples over bananas, then it is possible to ascribe to him a preference for apples, in choices between apples and bananas. If, on another occasion when no apples are available, he is seen to chooses bananas over cherries then it will be possible to attribute a preference for those, in that context. Ideally it will be possible to assign rankings and other measures of preference strength so that it could be predicted that he will choose apples when offered a choice between all three. In this way decision theory assigns a preference ranking by taking account of what the subject does in relation to those options that are available. For the interpreter to assign a preference ranking then it must be the case that the behavior itself can be discerned as aiming at some outcome and this must occur against a background of available options. Thus the interpreter has two tasks: first discern the options available, then note the choice that the subject actually makes. To assign a preference is then to assign a two place relation. It is to predict that the subject will choose X over Y.

While preference rankings are useful heuristic tools it is important to note that Lewis is not committed to the idea that the mathematics of preference rankings acts as a model for the thought of the subject. Preference rankings are only useful to make predictions about outcomes. They cannot be used to reveal the actual mechanism of the subject's thought. This restriction is somewhat counterintuitive given the fact that decision theory is used by Lewis as part of an

interpretive method. If it is not possible to understand the modeling that decision theory uses as revealing the structure of thought, how should Lewis be understood when he claims that decision theory forms the very core of our common sense theory of persons? It would seem that the best way to understand him is to see that revealed preferences as showing the options that the subject takes to be most worthy of choice. However the method used to generate those predictions involve assigning a preference ranking to each option based on previous observation. As such, it might be thought implausible to imagine that the subject has something like a formally consistent ranking of all possible preferences built out of the directions of his choice in the past. In the first place this would require a massive memory that could account for all the previous choices he made. But in the second place, the kind of updating needed would require significant formal skill. Thus the calculus of decision theory ought to be expected to track, but not model, the actual thought of the subject. It is merely a device for predicting the outcome of that thought and it frames these predictions using the language of preferences. Although this is to make use of mental terminology, it ought to be understood as indicating the direction of choice among the available options.

If decision theory is useful as a heuristic and predictive tool one might wonder to what extent it is capable of producing an interpretation of the subject. Understood as framing device- a constraint on interpretation- it functions to ensure that the totality of states ascribed hang together in an ordering that is internally consistent. Lewis's thought is that preference ascriptions that violate the consistency conditions of decision theory are baffling and so produce ascriptions that lead to baffling pictures of the subject. This needs to be avoided because a baffling interpretation is no interpretation at all. Decision theory is the tool for this job. This gives decision theory a fundamental role in the overall method of interpretation because it acts as a way of ascribing

preferences directly from behavior without providing an account of what preference formation is in the subject. But this forces a question. Lewis describes decision theory as the core of our common sense theory of persons and he implies that any new developments that might make for better explanations of human behavior will be folded into decision theory. Is this true? Is decision theory so central to our conception of human beings that any new explanation will merely augment the theory? Or, put more bluntly, is there reason to be confident that the best theory of human beings will be decision theoretic at its core?

Unfortunately for Lewis's account it seems that this level of confidence is unjustified. I will show that further examination of this method reveals substantial interpretive work already goes into fixing the categories that decision theory can work its calculus on. Indeed this pre-interpretation is so important that it can be used to show that the behavior in question is not of a type that decision theory can model. This shows that it is possible to achieve a substantial explanation of the subject's behavior which is uninterpreted by decision theory and these developments show that decision theory cannot have the fundamental and central role that Lewis imagines for it.

To see this it is useful to analyze the role of preferences within decision theory. For Lewis the term 'preference' must act as a theoretical term wholly defined by the mechanisms of decision theory. As such, preference ought to be understood as the two place relation that predicts the direction of the subject's choice by reference to the available options. Understood this way it should be clear that preferences, in the context of decision theory, will not always behave in the way that would be expected from a vernacular understanding of the term. So for instance, if a person were conditioned through to always perform a certain action when promoted by specific stimuli, that action would be counted as a preference by decision theory and would be

accorded a stronger weight than other options that could be seen as available at the time of action. But, in the vernacular it would be odd to talk of person 'preferring' the outcome of a classically conditioned or merely habitual action. However, because the act produces its outcome in the context of other available options, it can be ranked against those. Further, if it were the case that the stimuli which elicited the response were always tied to circumstances in which it was possible to see other available options, that response would be expected to receive a very high preference ranking relative to these other options.

It seems then that the decision theoretic calculus is ready to treat any movement toward some outcome, in the context of other available moves, as a 'preference' and in this way departs from the vernacular use of this term. This need not be a problem for decision theory. After all, philosophers love to tinker and it is easy to modify a formal calculus to better capture the terminology it seeks to model. Moreover there would be various clues that might alert an astute observer to the fact that this was a somewhat deviant case of preferences. For instance, if the habitual behavior was intermittent the preference ascriptions would almost certainly become inconsistent. All that would be required it that the behavior be something that is generally available as an option but is intermittently performed. In that way a preference ranking that was generated from the bulk of choices would reveal a preferences for X over Y, but in some cases Y would be performed when X was an option on the table; seemingly in violation of the preferences previously observed. Odd preference rankings could also be achieved. It might be the case that a habitual behavior is performed regularly in a variety of situations. If so, it could be assigned a ranking that counted it as so preferable that it was at, or near, the top by a large margin of an ordering of all preferences. Clues such as these might cause interpreters to make further inquires into the behavior in order to see what was generating the odd preference ranking.

It is telling that this would be an inquiry into whether or not decision theory got the case wrong. As such, it is clear that there must be other tools of interpretation that stand outside of the constraints of decision theory. Otherwise it just wouldn't be possible to second-guess the results obtained in these odd cases.

Before considering this investigation, I want to briefly explore sticking with the status quo. It might be thought that the decision theorist could simply bite this bullet. It would be possible to rest content with the odd preference rankings that such a behavior would produce. But, doing so puts the theory's status as the best theory of behavior in jeopardy. This follows from the fact that an opponent who was armed with a method for distinguishing between preferences the theory could model and those it couldn't would necessarily have additional resources for giving more accurate predictions of the subject than the more coarse-grained and unmodified decision theoretic approach. Resting content with the theory's deliverances then is not an option Lewis can avail himself of since the possibility of this investigation shows that decision theory's ability to interpret is limited to a subset of behavior.

So it would seem that it would be desirable to augment decision theory with tools to investigate and so disambiguate. Since what is being resolved here is a kind of ambiguity, it must the case that further inquiry aims at resolving the case in one direction or another. It is easy to imagine the forms that this further investigation might take. The subject himself might be questioned with an eye to gaining a description of that which motivated the behavior. One might try to consult his medical records or his past history with an eye to uncovering possible sources of the behavior which would help settle the ambiguity. Indeed any investigation which seeks a more fine-grained understanding of preference than decision theory can achieve would be useful here.

The shape the inquiry would take is interesting. In making use of these other methods the interpreter is doing an investigation into the applicability of decision theory to the phenomena at issue. Essentially he asks: is this the sort of thing that my theory is designed to give an account of? But that the question takes this form implies that a substantial amount of interpretive work is being done by this investigation. It is bringing a range of explanatory resources to the inquiry that decision theory lacks because it must decide whether the theory is applicable and then offer some justification for this decision. In cases where this inquiry can determine that decision theory is applicable it need only invoke the resources that decision theory would use anyway. For instance, discovering that the behavior appears to be a voluntary choice made with knowledge of alternatives would definitively secure decision theory's applicability as that it just what the decision theorist takes to be relevant. But discovering a deviant case would mean saying something about why this was the sort of case that decision theory is ill suited to handle and an explanation of this type would go beyond the range of resources that decision theory has at its disposal. As such, if decision theory requires this level of supplementation, it threatens the idea that by itself it is the best method of interpretation available. In fact, it looks like it cannot be taken as a complete account of a subject's behavior.

There are two options for the theory when it deals with odd preference rankings that may require further investigation. Put bluntly, the theory can either embrace or eschew the behavior in question. That is, it can either revise the input conditions in such cases. Or, alternately, it can ascribe the behavior to some source the theory does not purport to treat (i.e. by denying that it should be understood as choice-driven). The first option is employed in cases in which a person acts so as to violate the theory's predictions. Surprisingly, these can be gerrymandered into support for theory by treating the lack of predictive success as indicating that the input

conditions were not as they were originally supposed. So, instead of being evidence against the theory, the case can be refashioned as one that speaks in the theory's favor. This would just be self-validation since the starting preferences are just fitted to the outcome. In the alternative strategy, deviant cases can be explained by way of a disruptive influence. In so doing, this cause would be shown as the kind of thing which removed the thing done from the category of behavior.[23] Thus the activity is removed from the category which the theory aims to explain. This strategy obviously works best for explaining cases of apparent defect or disease, since it is far easier to consign these motives to something other than rational thought. In this way the theory is biased toward seeing a person as rational because it can explain all cases of good motivation by making use of the theory's rationalizing resources, but consigns all cases which bear a superficial similarity and so could serve as counter-examples into a category that the theory does not purport to treat.

Noticing this is not yet to show that Lewis cannot use decision theory in that way that he wants: to underwrite behavioral interpretation. A defender of Lewis might simply stipulate at this point that he is only interested in offering an account of the kind of behavior that decision theory is capable of describing. He might go on to claim that the kind of augmentation of the theory that the disambiguation investigation requires is of a type that reveals the behavior belongs in these deviant categories. If this is so he might plausibly claim to be treating a subset of the totality of human behavior and then go on to claim that this subset has a special status as being the only kind of behavior that it is possible to offer an interpretation of. Perhaps in these deviant cases it is only possible to give a diagnosis or invoke some other explanation that assigns mechanisms rather than meanings as motivations?

[23] Or, at least, the category of behavior the theory is able to treat.

This strategy for defending decision theory appears suspect because it treats the theory itself as the arbiter of which kinds of behavior an interpretation can be offered for . It is not obvious that the range of behavior for which an interpretation, as opposed to some other kind of explanation, is appropriate is exactly that range which decision theory is well-suited to describe. Decision theory's applicability can only be definitively secured by discovering that the subject will think in ways the theory can model. Further, if that is the sort of information which must be discovered, and so cannot be merely assumed, then this is a subject in which contrary discoveries might also be made. One might learn that the subject is whimsically motivated and so the available alternatives play little role in his choices. Or, one might learn that entire courses of action appear in the subject's mind, fully-formed, and he feels drawn to carry them out with no concrete end in mind that would justify them. One might also learn that he is constantly rushed with little time for comparison; when one plan seems 'good enough' straightaway he carries it out. In learning any of these things the interpreter would be filling in principles that were analogous to posits M and U in their ability to qualify the constitutive ideal provided by decision theory. As such, it is these kinds of traits which ought to function as qualifiers on the application of decision theory by the lights of Lewis's own view of thought.

If it is appropriate to supplement decision theory with this kind of information, this reveals that it is appropriate to qualify the constitutive ideals that the theory makes use of to get its interpretation going. But this suggests that Lewis's method is merely dogmatic in its assertion that the best theory of human motivation will be decision theoretic at its core.

This dogmatism can be explained by noting that for Lewis the major theoretical terms in which the explanation is given are all analytically defined within the theory itself. There is no way for the theory to confront reality and modify its central principles or the resulting

terminology. 'Belief', 'preference' and 'choice' just are to be understood by their implicit theoretical definitions. Either the candidate preferences and choices ascribed fit these theoretical definitions or they are rejected as possible candidates for being captured by this terminology. Lewis describes that issue as follows:

> "…the fundamental principles of our common sense theory of persons implicitly define such concepts as belief, desire, and meaning. Actually I would like to claim something stronger: that the implicit definitions can be made explicit, and that the explicit definitions so contained would be analytic. If so, then our constraining principles would themselves have a status akin to analyticity: Karl might have no beliefs desires or meanings at all, but it is analytic that if he does have them then they more or less conform to the constraining principles by which the concepts… are defined" (Lewis 112)

Earlier I advocated understanding Lewis's posits M and U as structures located above the level of beliefs and desires which generate 'dispositions to reason in certain ways and sensitivities to characteristics of a situations'. If this understanding is correct, then similar posits will be capable of qualifying the rationalization principle in a way which diminishes its constraining power. That is, as long as posits M and U are seen as capable of producing dispositions to think in certain ways and dispositions toward certain kinds of sensitivity, then it will be possible to see similar posits as weakening the constraining power of the rationalization principle in a great many interpretations. But that possibility can only be realized if the new posits are seen as having the ability to influence the interpretation independently of the working of the decision theoretic calculus.

This weakening of the constraining power of the rationalization principle can be accomplished by citing features of the subject which might plausibly cause him to have dispositions or an outlook which would diminish his propensity to behave in ways predicted by the rationalization principle.

3.5 MODIFYING RADICAL INTERPRETATION

John Broome's article "Can a Humean be a Moderate?" is a discussion of what role Humean instrumental rationality ought to play in decision theory.(Broome 68-87) Broome first distinguishes two views of the role of rationality. On one view rationality provides no real guide to life because rationality is not concerned to set a person's ends. Instead reasoning is purely instrumental and so only provides a person with the means to fulfill whatever ends she has. Call this the extreme Humean view. The moderate Humean view, however, asserts that decision theory can act as a way of prohibiting some sets of ends by providing the standards of consistency for sets of preferences. Thus a person can have any preferences that she likes; however her sets of preferences must be consistent if she if going to be counted as rational. In this way a moderate Humean is committed to the idea that irrational preferences can only count as irrational given some antecedent preferences. And it is decision theory which dictates the consistency conditions for sets of preferences. It is Broome's claim that the moderate Humean view collapses into the extreme Humean view. This is an unacceptable result for Broome because it shows that there is just no such thing as irrational preferences, or anyway, that rationality provides no guide to which preferences one should have. His solution is to modify decision theory so that it contains specifications that dictate that some preferences are irrational simpliciter, and not just because they form part of an inconsistent set.

Obviously Broome and Lewis have very different ideas about the role that decision theory ought to play. Broome uses decision theory as a critical tool: its use is in specifying what sets of preferences count as rational. On the other hand Lewis employs decision theory in a descriptive way: it sets the framework on what can be attributed to the subject in interpretation. But there is sufficient overlap for the two views to be relevant to one another.

Broome must be in the business of attributing preferences if he is going to then pronounce a set of preferences rational or irrational. Further, Broome is concerned that some sets of preferences that appear intransitive will be prima facie incoherent unless it is possible to reconcile them with decision theory. So, at bottom, both authors are concerned that absent a way of making sense of preferences that appear irrational the behavior of the subject will just be baffling. Broome offers a method of finely individuating preferences that can help resolve some apparent cases of intransitivity. I show that the kind of investigation that justifies fine individuation of preferences can also be used to augment the explanatory powers of Lewis's use of decision theory.

I will argue that Broome's criticisms show that without knowing how a person individuates preferences it won't be possible to determine if her preferences are rational by the standards of decision theory. Moreover I will agree with Broome that decision theory is incomplete as a prescriptive theory unless it includes criteria for preference individuation. The upshot of Broome's solution is to incorporate a way to gain insight into the subject's own understanding of the world with an eye to using this to help individuate his preferences.

I then argue that Lewis would produce a better theory by attending to the subjective states of mind of the subject. This provides the resources to distinguish between different kinds of motivation and allows for a more nuanced picture of the subject. However, the richer picture of the subject comes at a cost for Lewis. The upshot of attending to these features of the subject mean that the resources of decision theory are inadequate to provide a complete picture of the subject's behavior. The solution I propose suggests that there is no reason for the asymmetrical treatment of thought and behavior at the center of Lewis's theory. Behavior should be as amenable to non-idealized descriptions as thought.

Broome is most interested in the way that some groups of preferences seem to violate the principle of transitivity in decision theory. Transitivity is a consistency constraint on the ordering of preferences. It states that if a subject prefers A to B, and B to C, then she ought to prefer A to C; assuming she is rational. Broome is concerned that some sets of preferences seems to violate transitivity because the individuation of the preferences appear to be indexed to the choices on offer. In the example he uses, Maurice has a variety of choices. He can stay at home, he can go mountaineering in the Alps, or he can go to Rome. Maurice prefers staying home to going to Rome. And he prefers going to Rome to mountaineering, but, oddly, he prefers mountaineering to staying home. His preferences appear intransitive and so it would seem that he is being irrational. "Not so" claims Maurice. Rome is boring so it is best to stay home when give the choice, he reasons. And heights are frightening to him so he would rather go to Rome than to the Alps. But, when given the choice between staying home and going mountaineering he regards the choice as a test of his courage and so, to avoid cowardice, he elects to go to the Alps. Maurice's ability to finely individuate preferences seems to have blocked the charge of inconsistency. Moreover, his fine individuation of preferences tells us something about how Maurice sees the world; he is not inclined to consider opportunities singly but instead sees them as part of choices. So understanding how Maurice sees the world requires adjusting the preference rankings that would have been assigned if an interpreter wasn't privy to this insight.

While this strategy may seem to give a plausible rejoinder to concerns about Maurice's consistency, Broome worries that as a general strategy it will allow a person to always evade a charge of inconsistency. All an agent would need to claim is that her preferences are, in fact, more fine-grained than was initially suggested by the failure of transitivity. Maurice's case provided a somewhat reasonable account of his fine-grained preferences, but Broome's concern

is that in general no such reasonable defense of the structure of preferences would need to be provided by an agent who availed herself of this defense. The moderate Humean view thus collapses into the extreme Humean view because it is no longer possible to charge an agent with holding an inconsistent set of preferences. This charge can always be evaded.

Broome recognizes that this move is a threat to the status of decision theory as able to offer guidance. He seeks to solve this problem by requiring that fine-grained preferences be justified in a manner similar to the one Maurice invokes in the example. That is, Broome proposes that decision theory be altered to include a 'rational principle of indifference' that requires that a person be indifferent between two options so long as the difference is not enough to justify one option over another. Broome recognizes, however, that one cannot adopt such a principle without abandoning the moderate Humean position. So some options will not be irrational because they are inconsistent with antecedent preferences, but instead, because they are a preference for X over Y when there is not enough of a difference to justify a preference for X or Y. It follows that this option requires one to regard some preferences as irrational simpliciter; namely those that fail to conform to the rational principle of indifference. Thus the moderate Humean must either accept the principle of indifference or become an extreme Humean.

It may seem as though the moderate Humean can rescue the idea that preferences are only irrational by being inconsistent with antecedent preferences. This would allow her to evade the idea that some preferences are irrational simpliciter. However if she is going to avail herself of this defense she must allow that Maurice is able to make a very strange kind of comparison; he must be able to compare two options that he could never have a choice between. That is, Maurice must be able to form a preference between going to Rome when the other option is mountaineering and going mountaineering when the other option is staying home.

Broome recognizes how odd such choices would be. He calls these nonpractical preferences because they are choices between options that could never be presented as a practical choice. For the Humean to maintain her position she would have to argue that a person could feel drawn to one or other of these nonpractical options on the basis of the pull of her desires, independent of any rational considerations. Any rational considerations that need come into play here would just force her into the position of adopting something like a 'rational preference of indifference'. As such, to remain a moderate, only brute or unjustified desire can influence her preferences between such options here.

This problem gives Broome the leverage he needs to argue that the moderate should either adopt his solution or become an extreme Humean. Broome argues that while it is possible that some preferences can be formed through merely consulting the strength of one's desire, nonpractical preferences are not like that. Nonpractical preferences are too abstract and require too much intellectual effort to understand to be introspected on the basis of mere feeling. One can discern a direction of preference by weighing up considerations in favor of each alternative, but doing this is a rational process, he claims, and so by recognizing that this is the only way to distinguish that alternatives one admits to needing something like a rational principle of indifference to do the required work.

So on pain of becoming an extreme Humean and being forced to take nonpractical preferences seriously, Broome proposes the more attractive option of looking to the subject's own justification for the grounds for the fine-individuation. And it is important to note that in saying this Broome is imagining a restricted role for the term 'preference'. Preferences are not just what one is drawn to when other options are available, they are at least considered enough that a person can cite reasons for her choice. Further, the resources that decision theory needs to

justify fine-individuation in one case and not the other have their source in a subjective account of the subject. That is, it is the motivations that the subject himself cites, or fails to cite, which act as the difference maker in seeing the ascribed preferences as rational or irrational.

Broome seems to see his proposed solution as a rather minor fix to the overall decision theoretic method, but it seems that once it is admitted that a subjective account of the subject's motivations is relevant to the ascription of preferences, there is no reason to limit this move to only justifying fine individuation.

Earlier in this paper I suggested that it was possible to imagine a kind of reflexive behavior which was capable of producing intransitive preference orderings. At that time I argued that it would be prudent to begin an investigation that could disambiguate preferences of the right type –those that would be expected to fall in a rational ordering- from those apparent preferences that might be mistakenly read off of behavior. Broome's proposed solution to the intransitivity problem above also works as a solution to this problem. Because it would be expected that the subject himself would know, in general, if a behavior had been motivated by choice, a subjective account of the subject would be expected to reveal enough information to justify the disambiguation in one direction or the other.

Similarly it is possible to imagine other roles where the inquiry into the first-personal would help to secure the applicability of decision theory to the behavior. For instance, I have argued that preferences must be seen as a two-place relation. A person only prefers X by reference to Y. But this means that fixing the background availability is essential to producing accurate preferences. Here too it would be possible for simple ignorance to result in intransitive preferences. All that would be needed is a mismatch between the interpreter's and subject's view of what actions were available. The result would an inaccurate preference because the value in

the second place wouldn't reflect the subject's view of the available options. But, of course, this would also be easily resolved by a simple inquiry into what the subject thought were viable choices that could be made in that situation.

Broome recognizes that he can only adopt his view of the rationality of preferences by offering a redefinition of the way that the word 'preferences' is understood. But he is not concerned about this because he only sees decision theory as a way of specifying the rational structure of preferences where these are already understood as the kind of thing that one needs to apply rational considerations to forming. Indeed he claims…

> "It must genuinely be a condition of rationality that a person's preferences should conform to the consistency requirements of decision theory – transitivity in particular. We understand preferences in such a way that this is so. I think this condition provides a serious constraint on the notion of preferences, because if preferences are conceived of in some popular ways (as feelings, for instance) it is very hard to see why rationality should require them to be transitive." (Broome 78)

But it is clear that this is just a way of pointing out that decision theory ought not be taken as a total theory of human motivation. Broome is acknowledging that people may make decisions on the basis of arational motivations (feelings for instance) he is just unconcerned about such things because he doesn't see it as decision theory's job to specify constraints for such influences. Still later, Broome says more about how exactly this constraint on preferences requires him to adopt a nonstandard notion of preference.

> "We have at least two different concepts of preference. According to one – call it the 'evaluative' concept – a person prefers A to B only if she estimates the goodness of A above the goodness of B. According to the other – functionalist – concept, she prefers A to B only if she is in a state that typically leads her to choose A rather than B." (Broome 84)

Broome goes on to make use of the evaluative notion in specifying his solution to the problem.

With this in place it is clear that Lewis is running together two concepts. He employs the functionalist concept to claim that any time a choice has taken place it is possible to

discern a preference. But then he uses the evaluative concept of preference to go on to imply that the consistency conditions are applicable to the preferences read off of behavior. This kind of move would be innocent if one could specify in advance, as Broome does, that he is employing some restricted notion of preference that only applies to choices that are made as a result of evaluation.

In the context of a theory of interpretation, however, this kind of move cannot be made. The whole enterprise of interpretation is organized around the idea that one cannot tell in advance that some behavior or other is the product of some specific kind of mental action. Indeed, Lewis aspires to make sense of the totality of Karl's behavior. So, as long as one is willing to recognize, along with Broome, that some behaviors will be made sense of my pointing to a person's feelings or other motivational mental states instead of any evaluative mental states, this will be enough to show that decision theory can only have a role in making sense of some subset of behaviors. Thus decision theory cannot purport to be a complete account of the psychology of anything except some perfectly rational choice evaluator. Broome shows that decision theory must be qualified or supplemented because some choices will be motivated by something other than 'preferences' in his special sense of the word. Because it is clear that one can make sense of human behavior by citing feelings, emotions and habits, one must be committed to accounting for such features of human psychology within the context of an interpretation. Decision theory is inadequate to that task.

4.0 INTERPRETATION AND VIRTUE THEORY

This essay challenges the exclusive focus on rationality common to some virtue theoretic approaches to ethics. Specifically it takes issue with the idea that an act must be motivated by rational considerations if it is to contribute to a life of virtue. I argue that the focus on rational action as an inseparable part of virtuous action obscures the value of arational states in human life. Further, it fails to account for the fact that rationality is often an unwieldy and imperfect tool for the attainment of ends that can be achieved by arational means. I conclude by suggesting that a Platonic approach shows promise for evading the idea that rational justification is the sole criterion of good action. The result is an approach to virtue theory which has room for all of those whose natural traits, upbringing, and circumstance make the pursuit of flourishing through solely rational means untenable, while still allowing room for seeing the life of reason as one kind of flourishing human life.

This essay is organized as follows. First I will present a common method of introducing virtue theory: the analogy with health. I show that this analogy is intended to set the stage for virtue theory by presenting, in health, a purportedly singular and universal end that can be likened to the role that flourishing plays in defining which traits are to be counted as virtuous. I then challenge the idea that health can provide such an uncontroversial end and, by extension, challenge the idea that flourishing has a character which is similarly singular and universal. Next

I argue that just as different conceptions of the human being help to shape differing notions of health, an examination of flourishing reveals that there appear to be multiple routes to the achievement of this end –some very rational and some less rational. I go on to show that practical rationality, understood as a capacity for treating reasons as reasons, does not always provide the best or most direct route to flourishing. Instead, I argue that human beings have capacities to recognize and respond to situations in ways that do not require the recognition of reasons as reasons. These capacities for engagement appear too quick, intuitive, or emotional to be rational. However, whether through training or instinct, they appear well-suited to bringing about good outcomes. A virtue theory that cannot recognize this risks excluding some successful modes of living from being considered virtuous. The solution to this is a form of virtue theory that has the resources for recognizing than many parts of the person have a role in securing a flourishing life.

4.1 HEALTH

An analogy with health is often used to illustrate the central role that flourishing plays in virtue theory. It is designed to put a universal but uncontroversial standard of excellence in place that is applicable to any person simply by way of her biological makeup. The analogy takes the kind of advice that would be dispensed by a medical practitioner to be on par with the injunctions issued by the virtue theorist. That is, the analogy treats the injunctions of the virtue theorist as simple pragmatism for the task of living well. Because health is seen as desirable for all and applicable to all by virtue of their bodily constitution, it serves as a useful way of modeling virtue theoretic advice. Flourishing is similarly treated as universally desirable and applicable to all by virtue of the purported essential qualities of the human being. However, this

parallel treatment also reveals a specific kind of vulnerability. If it can be shown that medical advice, and the desirability of health, can be shown to be indexed to desires or contexts which are not universal, then this will threaten the universality of its injunctions. It would remove medical advice from the category of simple pragmatics for living a human life and place it in the category of injunctions in the service of some goal or other. While showing medicine to have this status wouldn't be enough to threaten virtue theoretic injunctions, if they do admit of the kind of parallel structure that many virtue theorists seem to assume for them, the concerns about the universality of medicine could provide the template for similar concerns about virtue and rationality.

It should be obvious that health and the practice of medicine go hand in hand. Indeed, with few exceptions, it was the expertise of the medical practitioner which granted the authority to say what was healthy and what wasn't. But what exactly does the doctor aim at when trying to restore her patient to health? It might be thought that health provided a singular standard that every doctor should aspire to, but in fact, what is taken to be healthy, or part of medical practice, admits of many interpretations. In what follows I want to examine some differing conceptions of health and the medical practices they animate. It will be seen that while some conceptions are more controversial than others they all share a common structure: they begin by taking some picture of the human being for granted and then develop the concept of health around that. While the picture of the human being that informs the resulting concepts is intended to be quite abstract – a stand-in for anyone who might show up in the waiting room - I show that it carries a great deal of content that is capable of artificially narrowing the conceptual networks it informs.

Consider Plato's *Republic*. Plato notes that in building the Kallipolis there will be a need for a certain kind of medical practice, but it is not the variety of medical practice which he thinks then dominated Athens. Plato complains that too often medicine is practiced so as to aim at the relief of minor complaints which are brought on by the lifestyle choices of the patient. So those who constantly indulge in food that is too rich, or who drink to excess, solicit doctors who will remove the hangovers and bowel complaints brought on by intemperance. Plato criticizes such a practice on the grounds that these 'small cures' do nothing to address the imbalanced mode of life that such people practice. Instead, he proposes a form of medicine which treats the temperate worker of the Kallipolis in a way that protects his essential identity. Plato's proposal is to employ doctors who simply won't treat those who practice debauched lifestyles. Further, the treatments that do exist will aim at the benefit of the worker, apparently on the assumption that the person is defined most essentially by his occupation. Plato describes the kind of treatment that a house builder with a head injury would receive. The builder would be told to rest for at most a day or two and then return to work. The injury will then either resolve on its own, or the man will perish. Most surprising is Plato's insistence that both of these options are preferable to a long drawn-out form of treatment that coddles the worker until his injury is healed but in the process makes him a less skilled or willing worker. Indeed, his description of the worker presents a man who so closely identifies with the job of building that he also thinks this form of treatment preferable to one that forces him to mind his illness and neglect his work. (406d)

Plato's medical proposal sounds radical to contemporary ears because his focus is on treatment of a worker rather than a human being. It goes without saying that to our ear there is a distinction between one's self and one's career that needs to be made. A person, we think, is not simply his job title. But noticing this puts the question I want to ask in sharp relief. What

would it mean for a form of medicine to aim only at the treatment of the human being? That is, would it be possible to come up with contentful medical policy that aimed at the human being simpliciter? I maintain that if such a policy were developed it would be so open-ended as to be empty of any real content. The reason for this is that terms which ought to function as generic universals appear to be laden with content that reflect certain biases and pre-conceptions unique to the context in which they are used. Eliminating these requires taking a skeptical approach to the essential properties of the human being, while retaining them fails to ground the universality that virtue theory or medical practice seems to aspire to. In what follows I want to briefly examine this point in more detail.

This first issue that must be settled in developing a medical practice is deciding what it is that one aims at in trying to create a healthy patient. In contemporary medical practice treatments that are considered viable are informed as much by the desires and circumstances of the patient as the disease. For instance, while it might be appropriate to recommend a somewhat painful treatment that maximizes longevity in an otherwise healthy and young patient, it would be problematic in a terminally ill patient that experienced chronic pain. Similarly, a treatment that risked a patient's fertility might only be considered a last resort in a younger patient, while an older patient's doctor might see the same treatment as promising. It is clear the healing the diseased is one principle that animates contemporary medical practice. But even this principle does not completely characterize the practice since sensitivity to patient discomfort, autonomy and expected life-plan also figure in our practice. Our contemporary practice also recognizes preventive medicine and cosmetic medicine as legitimate off-shoots. To recognize the legitimacy of these branches of medicine would seem to indicate that medicine appears to be charged with the task of maximizing ability. Indeed, it is common to tailor treatments to the patient's career or

expected life path. On this conception, medical practice aims at creating a human being who has no impediment, and perhaps even an enhanced ability, to function under the circumstances he chooses for himself or can expect to finds himself in[24].

While I do not have the space to offer a full defense of this conception of medicine, I think that it can be supported by noticing some features of contemporary medical practice. First it should be noticed that contemporary medicine has now largely abandoned that older form of paternalistic treatment that previously defined it and come to see patients as autonomous decision makers. The fact that such a break can even be recognized bears testament to the central role that autonomy plays in contemporary medical practice. Further still, this conception has helped to usher in a consumer centered medical practice in which patients are seen as 'shopping' for services that medicine is able to provide. Cosmetic medicine operates explicitly on this basis. Less noticed is the fact that traditional forms of medicine also operate by taking urgency of complaint to be a factor which determines how treatments will be carried out. For instance, superficial dermatological problems might be seen as worthy of treatment or not depending on how much dissatisfaction the patient expresses about them. Further still, it is not unusual for facts about a patient's profession or lifestyle to enter into the determination of how pressing a complaint is.

The contemporary practice then certainly breaks with the Platonic form just considered. Plato's goal was to produce a functioning worker and he was willing to ignore complaints from those who engaged in a life of vice. Our contemporary practice takes complaints seriously and considers occupation alongside other roles that the patient self-

[24] Julian Savulescu argues that the distinction between treatment and enhancement is arbitrary. He notes that enhancement aims at the same goals treatment and argues that medicine's focus on promoting human well-being requires acknowledging enhancement as a legitimate form of treatment. See J. Savulescu 'Genetic Interventions and the Ethics of Enhancement', 2007.

identifies with. The contemporary and Platonic forms are incompatible precisely because they begin with different conceptions of the human being and frame the rest of the practice around that. Probably, to the contemporary ear, the Platonic conception looks defective because its central principle results in forms of harm that our system avoids. But, treating autonomy as the most important feature of the patient also carries risks. The most obvious is the deaths and ill-health that result from being required to take a patient's choices seriously. In addition, there will be instances of regret, forms of alteration that only evade the charge of maiming by the fact that they came at the patient's request, and doctors becoming unwilling assistants in alleviating the consequences of their patients' debauched lifestyles. Neither system then can claim superiority on the grounds that it prevents all possible harms. From the outside, only a judgment that this or that system better captured the essence of the human being could justify a preference between the two. But my purpose isn't to hold one from or other up as superior, it is simply to notice the relationship between the conception of the human being and the resulting medical practices it supports as legitimate.

One striking feature of medical history is the fluidity with which different conceptions have been allowed to inform various aspects of medical practice. There is evidence that pharmaceutical developers have always lobbied to see the creation of 'diseases' to which their potion offers a cure. Such lobbying, of course, requires a way of defining the untreated state in such a way that it appears defective. For instance, it would seem that treatments such the pre-frontal lobotomy were justified by certain paternalistic attitudes that treated the expression of distress itself as disordered. There is evidence that the various treatments used to combat the ennui of 1950's domestic life owed much of their character to assumptions about the proper role for women and what activities they should derive satisfaction from. More recently, the popularity

and medical endorsement of testosterone supplementation seems to speak to a willingness to treat the natural consequences of aging as disordered. None of these practices are possible to justify without at least tacit reference to an ideal of the healthy human being.

By raising such examples I do not intend to begin a criticism of this form of medicine or that, but instead to notice the common feature which connects the Platonic and the contemporary: that which will be seen as a legitimate medical complaint, procedure or treatment will follow from the conception of the human being that informs that picture. If, for instance, the picture of the human being that the conception of medicine uses depicts the person as first and foremost a builder, then procedures which deprive the person of that ability will be seen as a form of maiming rather than a form of treatment. Or, if the background social mores in place do not allow for a married woman to find satisfaction outside of domestic life, the failure to find satisfaction there will suggest itself as a form of pathology. Similarly the person-as-autonomous-agent picture that informs contemporary medicine regards the lobotomy as a form of maiming and laments how common its use was in years past.

So what would a purely general picture of the human being look like? If what is suggested by these cases is correct, it would have to be incredibly flexible. Assuming that each of these conceptions discussed, and the myriad possibilities which were not, has at least some claim to be a legitimate conception of the human being requires being skeptical of any particular conception offering a complete picture. As such, when it is recognized that there is some quite determinate content being built into the conception, this can be construed as evidence that the conception may be too narrow; unless it can be shown to be general enough to accommodate other conceptions which might press a legitimate claim to be incorporated. A purely general conception must be capable of accommodating any legitimate claim. Unless one is prepared to

argue that the proper conception of the human being needed to inform the medical context is already known, any purely general notion will have to allow room for future developments. But something with that level of generality looks as though it would operate as a *de facto* form of skepticism. It would, in effect, be asserting that it was unclear just what features of the human being needed to be prioritized in decisions about medical treatment. That conception would have to await further information, probably that which emerged from particular cases, before making a determination about which ideal of health was most appropriate in the particular situation. Its general stance would be empty and platitudinous until it had more determinate content to fix on.

This much appears inescapable. There is no such thing as the place-holder-person, or purely abstract person, which can inform the development of a completely abstract conception of medicine applicable to any human being. The best we can develop is an open-door policy toward any future conceptions. But this renders purely general medical injunctions empty precisely because of the emptiness of the picture that motivates the practice.

Finally, it should be noticed that the picture of the human being which motivates a medical practice is not a product of that medical practice. That is, there is nothing about the practice of medicine which would allow for a conception developed there to have any special claim to be the correct conception. The arguments that would motivate that are philosophical, not medical. These are questions that must be answered before the practice of medicine can take shape. The picture of the human being is that which medicine must conform to, it is not the other way around.

4.2 WHAT DOES THIS SUGGEST FOR VIRTUE THEORY?

Once it has been seen that medicine works by taking a certain picture for granted and then developing its practices around that, it is easy enough to notice that the conception of the human being which stands at the center of a virtue theory works to develop the concept of a flourishing human life. However, while the philosophical justification may go largely unexamined in the medical context, ethics doesn't enjoy this luxury. It is up to the ethicist to supply and defend the conception of the human being that will inform the resulting theory.

In offering this justification many philosophers have been impressed with a certain Aristotelian conception of the person as a rational animal. For instance, Philippa Foot, in her 2001 *Natural Goodness* cites both Aristotle and Aquinas in addressing exactly which conception of 'good life' is appropriate to the human being.(Foot 55-56) She notes that Aquinas, following Aristotle, describes the human being as the only creature able to cognize an end as an end. She then argues that this feature of human beings is not a psychological fact, but instead is a way of revealing that to cognize something as an end is to see it as the kind of thing which is amenable to rational appeals. That is, to offer an end directed claim that such-and-such *should* be done, is to see that "should" as in need of justification. The end directedness of the *should* claim thus stands in need of the justification that would be cited by saying what was worthwhile about the end coming about. That only human beings have the rational ability to justify certainly puts them into a special class of creatures. Further, because that justification has a connection with the agent's motivation to act, human action enjoys some special connection to the rational.

Foot's claim is not only that human thought and motivation work in a specifically rational way, but also that these special abilities have a distinctive place in any conception of human flourishing. Notice that one would point to a pigeon with a withered wing as a bad

specimen of pigeonhood, on the grounds that such an anomaly deprives the pigeon of exactly those abilities needed for pigeon flourishing. In just the same way, Foot argues that human rationality supplies the human being with a distinctive mode of living her life. Further she argues that the good life for a human being is inseparable from the rational abilities put into the service of living that life (Foot 66). This inseparability claim is central to her view. It is important because it rules out an end-state view of flourishing that reduces it to some merely psychological or material state. Instead flourishing is to be understood as a form of happiness which comes as the result of the exercise of one's rational capacities.(Foot 88-90) On this view then one cannot conceive of the flourishing human life in terms that are separable from the exercise of one's rational faculties. She notes, for instance, that for some adults who have suffered a cognitive impairment the only kind of enjoyment possible is child-like. Such people will enjoy the gustatory pleasures as well as simple games and certain sensual pleasures but they will never know the kind of joy brought on by, say, romantic love, intellectual engagement, or self-sufficiency. This example shows that it would be inappropriate to point to the pleasure such a life contains and call it a good human life on that basis. It might be a life that is as good as is available for someone in such a situation, but this is a good qualified by being indexed to the deficient cognitive abilities of the particular person. (Foot 86)

It is Foot's concern to develop a notion of the flourishing human life which respects the central role that rationality contributes. In developing this Foot argues that human flourishing is a distinctive form of happiness with a close connection to a specialized notion of benefit. It is this form of happiness which provides the end for human life. She is concerned to show that this is a form of happiness that *a priori* rules out wickedness. (Foot 96) But this is very difficult as she describes the bare idea of happiness as 'protean' – first appearing one way and

then another. (Foot 97) In response to this she offers a way of understanding benefit which purportedly reveals in its grammar that the flourishing life is inseparable from a life governed by practical rationality. She further argues that practical rationality acts as a master virtue.

Foot argues that there is a specific understanding of 'benefit' that reveals the distinctive from of happiness she aims to describe. She considers the case of Fred and Rosemary West, sadistic serial killers who were able to avoid detection for decades. Foot notes that the Wests would not have been able to carry on their crimes for so long had it not been for the incompetence and neglect of many police, case workers, and others charged with the care of vulnerable people and investigation of missing persons. She claims though that it would be incorrect to describe the Wests as being *benefitted* by the conduct of these people.(Foot 94) While it might be the case that their evil plans were helped along – even benefitted by these actions – it would be wrong, claims Foot, to treat the Wests themselves as being benefitted by the inaction of the authorities. Her claim is that the grammar of the term doesn't allow it to be used in this way.

It is clear that this example is intended to reveal that there is something incompatible with wickedness in the flourishing life that benefit contributes to[25]. But this negative assertion doesn't elaborate on the positive conception of the flourishing life. Foot treats the articulation of this positive idea an ongoing project but it is clear that she does have some ideas about what it must include. In another example, this one involving one who chooses to live life as a gangster, she makes clear that certain modes of conduct –i.e. the exploitation of fear as an instrument to

[25] It can be difficult to follow Foot's grammatical arguments here. Indeed, it seems natural, to my ear anyway, to describe the Wests as benefitting from the inaction and incompetence of the authorities. Further, even if Foot's point is granted, this use of benefit doesn't appear to directly illuminate the special notion of happiness because it doesn't seem to violate the use conditions of benefit, in her special sense, to apply it to forms of happiness that are not dependent on rational action.

gaining material wealth- is incompatible with the life well lived. (Foot 53, 65)To be sure, it is difficult to conceive of how a person who was used to detecting and exploiting the weaknesses of others could be able to live the life of, say, a dedicated friend or dutiful spouse. It seems probable that it is in the nature of such attitudes to prevent the individual who has them from fully revealing himself and his weaknesses to another person. The inhibition and mistrust generated by such a lifestyle would then prevent real fellow-feeling and empathy and prevent those close intimate relationships that rely on this from forming in an undistorted way. The gangster's view of people can be expected to touch all aspects of his life. So it is clear that Foot's theory is able to show wicked thought and conduct as problematic on the grounds that it is incompatible with a flourishing life. Further, her use of this and other examples example suggests, correctly, that proper regard for others and respect for their autonomy is part of living the flourishing life. Foot's conception of the goods a flourishing life includes doesn't appear to be very contentious. It seems to allow room for friendship, romance, a strong work ethic, and attention to responsibilities. In short, an ordinary life. (Foot 88)

Still, part and parcel of Foot's claim that the flourishing life and the life of virtue are inseparable is the idea that flourishing cannot be reduced to a merely psychological state. If it could, there is a threat that any old kind of activity such as pill taking, surgery, or attending certain rallies might engender the flourishing life, absent any particularly virtuous conduct on the part of the agent. Further, goods such as romance are seen as good in themselves, not for any pleasurable psychological tickles that might result from them. So the inseparability claim provides a way to connect the products of a specifically practically rational life to the life of virtue. Foot claims that although it is common for philosophers to cite desire as providing reason for action, this is not the only source of reason giving force. Duty, necessity, expectation and

other motives can also be considered by practical rationality. However, because this faculty has the special from of happiness as its end, these motives will always be considered through that lens. So, practical rationality can frame its injunctions as independent of any particular person's actual occurring desires, but in doing so makes reference to a purportedly universal end. To see why only the deliverances of this particular faculty are capable of bringing about virtue it is important to note that for Foot the flourishing life must be a human life. Since Foot identifies practical rationality as the faculty which expresses the agent's humanity, its deliverances from an essential component of her conception of flourishing. To count as flourishing the life must be a happy one, in Foot's special sense, and that requires that it emerge from the actions of practical rationality.

Considered then from this perspective, certain actions look like they cannot be the result of an all things considered deliverance of practical rationality. [26] While she does little to specify exactly what basis the practically rational judgments make use of in delivering their verdicts, the examples she uses give some clue as to how the flourishing life is being conceived. In a particularly telling example she describes a group of students who protested against the Nazi regime by randomly sending letters to fellow citizens expressing disgust for the Nazi policies. These students are soon captured and, in her example, given the opportunity to renounce their dissent and go free. They are further promised that a certain pill will be given to them which will wipe their memory of the initial dissent and act of renouncement. Because these are intelligent and thoughtful people she assumes that availing themselves of this option would result in a life

[26] For instance, promise breaking on the grounds that breaking this promise won't cause harm. Foot describes promising as a uniquely human institution that allows us to harness reciprocity and duty to accomplish the tasks that require joint effort. She further analyzes the institution as one that allows promise breaking for certain purposes, though the fact that no harm would come from the broken promise is not seen as a legitimate reason for breaking a promise.

of prosperity, intellectual and social engagement with others and similar desirable outcomes. However, she describes this scenario as one in which the good life is simply unavailable to these poor captives.(Foot 95-97) This example illustrates, most strongly, the claim that the flourishing life is not merely psychological. Nor can it be understood as a mere end state. Here is a set of people who have the opportunity to live a life which, considered from a purely psychological perspective, would be considered a flourishing one. The reader presumes that they will use their natural talents, empathy and goodwill to make productive lives for themselves and those close to them. However, this otherwise happy scenario is marred by an action the agents have no memory of and which only impacts their current lives in fruitful ways.

Given this conception of the flourishing life and its dependence on an all things considered standard of rational behavior it is natural to wonder why someone would be moved to pursue such a life. After all, if practical reason demands that one be rational even when that brings psychological torment and pain that is apparently not offset by other goods, one might wonder why it is that he should be rational in the first place. This question however is one that Foot thinks is answered by pointing to human nature. As she sees it, rationality is not only a uniquely human attribute, it is that which foists a distinctive mode of living on human beings. Just as having gills forces the fish into an aquatic life, so having rational abilities pushes the human being into a specifically rational life. That is, we live in ways that demand reasons and justifications, according to Foot. So the rational life is, for us, inescapable. If nature has given us a certain tool that determines the conduct of our lives, we can examine that tool to determine which standards of conduct ought to be acknowledged by anyone making use of it. For Foot, the realm of moral reasons and the realm of practical reasons are one. That which is instinctual,

emotional or arational is incompatible with virtue because these modes of thinking and acting fail to treat ends as ends and so fail to make the right use of reasons.

A reliance on practical reason is a requirement for the human mode of life and so it is not feasible to ask after reasons for escaping that life. Further, even if such a question could be asked, it seems that for Foot only one answer to that question could be produced. If one asks 'why should I be rational?" and is understood to be asking after reasons to be rational, it is easy to point out that the asking of the question already reveals a prior rational nature. Otherwise it wouldn't make sense to be asking after reasons for being rational since reasons are already what move rational agents. On the other hand, if the question is not understood as asking for reasons it cannot be made sense of. This inability to ask after reasons for rational conduct would seem to show that the rational life is not optional for human beings.[27] (Foot 64-65)

4.3 RATIONALITY'S PLACE IN A HUMAN LIFE

I want to begin my engagement with Foot by examining the notion of practical rationality in play. Foot's idea of practical rationality cannot be solely concerned with outcomes. If it were understood in that way, any outcome which advanced the flourishing of the agent could be called rational on this understanding. But this wouldn't make sense of the inseparability thesis. The inseparability thesis works to individuate outcomes by way of the thought and action that gave rise to them. This is why, for instance, her Nazi example looks so controversial. On an outcome based view of what is desirable, it looks like the prisoners should recant their criticism. It is only

[27] If this question is understood as asking after reasons for being rational, in general, the point made seems to follow. However, the question might be understood as restricted to a particular domain. On that understanding it would appear that one is asking after the benefits of behaving rationally in some particular endeavor.

when the success of their life after recanting is understood as tainted by the recanting, that a case can be made for saying that such a life fails to be a flourishing one. The outcome is tarnished by being unknowingly premised on an act of self-serving cowardice.

So if a life of flourishing cannot be understood by ignoring the motives and thought of the agent, it must be the case that practical rationality, Foot's only route to the flourishing life, has something to say about the kind of thought that is needed to achieve flourishing. The implication is that practical rationality demands that one think in the right way about life if that agent is to achieve the life that practical rationality aims at. Here it is not difficult to imagine the kind of things that practical rationality will demand. First, anything that is to be called rational surely needs to respect the most basic rules of logical consistency and probably must avoid those fallacies that lead one away from truth. However, it is clear that Foot demands more than this of her conception of practical reason. Proper exercise of practical reason, for her, also seems to require that one make the right decisions for the right reasons. She suggests, through the requirement that one treat an end as an end, that this requires one to actively weigh multiple considerations and then make choices on the basis of which of these is most compelling. Thus the kind of judgment exercised by the virtuous agent is deliberative. One is actively considering, weighing and discarding considerations in virtuous decision making.

The requirements for having certain kinds of thought, then seems to exclude some forms of human action. If one were to act out of habit rather than deliberation, whatever they were doing couldn't be considered a paradigmatic exercise of practical rationality. Similar remarks could be made for acting on the basis of emotion, sudden urge, whimsy, fantasy and other motivations that are not understood to include the evaluation of reasons as reasons. I suspect that this differential treatment of these other modes of action stems from two sources.

First, it is thought that the human being is fundamentally rational. So on this understanding the exercise of emotional faculties in decision making wouldn't necessarily be the proper exercise of one's human capacities. Since the flourishing life is a human life, it must be the case that it involves the exercise of one's human capacities especially in the arena of choosing how to act. By treating emotional faculties as outside of the expression of one's humanity, they can be excluded from having a role in virtuous conduct. Second, it might be thought that these other motivations do not take account of reasons in the right way to guide virtuous action. Because of the inseparability claim, it is not possible to separate outcomes from the reasoning that led to them. As such, if an action is motivated by something other than the conclusion of properly conducted practical reasoning, this is enough to impugn its virtue. In what follows I want to challenge the idea that the human being is essentially rational, before going on to suggest that arational action can make a contribution to the flourishing life.

One challenge to the idea that the human being is fundamentally a rational creature emerges from developmental considerations. It should be clear that the kind of reasoning demanded of an agent by Foot's account will only be possible for those of sufficient age. It is uncontroversial that, say, a preteen could not be considered mature enough to regularly put this kind of reasoning into practice. Further, there are obstacles, such as the maturity of impulse control, that prevent even young adults from being particularly good at reasoning in this way or acting on that reasoning in the appropriate situations. At the other bookend of life there are obstacles as well. In the first place, it seems natural to point out that specific kinds of nostalgia, habit and inflexibility are only possible in the more mature stages of a human life.[28]

[28] Since nostalgia is a longing for some euphemistically described past it would seem to be more likely with an increase in age. Habit, being ingrained over time, shows similar features. Finally some forms of inflexibility seem to

And to this it might be added that the development of such traits can be seen as an expected part of the aging process. This indicates that human motivation and the actions that spring from it can be expected to have a different character at different stages of one's life. Further, it is entirely reasonable to expect a decline in mental acuity as the human being becomes superannuated. When fortune or technology reward a society with longevity, then it is reasonable to expect that a corresponding accrual of these age-related motivations will also be seen. It is true that impulsive or emotional action can lead to ill-considered courses of action, but it is possible to develop antidotes to this tendency. It is common to try and instill obedience to authority in the young to make the guidance offered by more mature persons more effective at steering conduct. Societies also set up systems of social sanctions that take maturity into account in disciplining the young and impetuous. Other measures look as though they hold the possibility of mitigating some of the outcomes associated with age related traits as well.

What this suggests is that while rational motivation may suffice for bringing about good outcomes, it is possible to balance the action of traits and external consequences in order to try and ensure good outcomes for those who fail to be motivated by rational considerations. As the picture of the human being is broadened to include stages of life in which rational thinking is difficult to implement or doesn't sufficiently motivate, this provides some pressure to see a diminishment of rational motivation as normal and not defective. The appearance of defect emerges from a picture that treats one particular developmental stage as paradigmatic. But such a restriction of the picture appears arbitrary if the subject matter of flourishing is the whole human life. So if rationality is to characterize human life, it seems that the picture of human life that is being put forth is really a picture of healthy adulthood. Just as it was seen that further

---

result from seeing the world as progressively more corrupt and inhospitable and so might be expected if the person feels 'left behind' by some wave of changes.

examination of the 'abstract human being' in the case of medical practice revealed a picture with much more content than may have been expected, here too it seems that a rather definite picture is being used as the model for the abstract human agent.

Noting that the standards for virtue seem best suited to an adult life ought not be taken to show that as long as we are dealing with adults the standards Foot adopts are appropriate. In fact, educational considerations also play a role here. Consider again Foot's example of the Nazi prisoners. In this scenario, it is clear that very little will be accomplished by refusing to recant. The refusal will cost the prisoner's life and will probably result in torture. It seems reasonable to point out that the refusal to recant is extraordinarily principled. But now one might ask 'what kind of developmental and educational experiences are necessary to produce so extraordinary a person?'. It seems clear that such willingness to stick to principles in the face of accomplishing little and sacrificing a lot is not a normal course of action. In order to achieve this kind of principled stance the right kind of education must be given. Indeed I suspect that this is the kind of action that is born out of an education that stresses a respect for legal principles, makes room for the value of dissent and martyrdom, and stresses selflessness. None of this is to suggest that there is anything wrong with this from of education, though we can certainly invent scenarios where it works against the flourishing of the agent. However, it does seem that we should recognize how unusual and intensive such an education is. Teaching people to think in this way is to aim at disciplining and elevating a particular faculty: practical deliberation.

While it might be hoped that we can live in a world in which such education was common place and available, the availability of this education doesn't really speak to the heart of the issue. The issue can be phrased the following way: if an educational program of the kind suggested is carried out, it can be expected to produce a dominant rational faculty, in part, by

diminishing contrary motivations. So, intuitive, emotional, or spontaneous reactions to situations will be less active as a result of this education. Further one can expect that an increase in the foibles associated with rational thinking will also be observed as part of the process[29]. This is just to say that educating someone to be rational, where that aims to alter their motivational structure, is a mode of changing the person and in that change the costs as well as the benefits will be borne together since they both stem from these educated aspects of the person.[30] I mean to suggest that these changes to the person are a way of narrowing the active traits of that individual. Indeed, any mode of discipline seeks to reduce the effects of what is spontaneous or unconsidered. This need not be problematic since at least some discipline is needed to live any human life, but they need to be registered since it is in this narrowing that there is the temptation to point to the product of this process as the purely general case. While educating a person into such a mode of life appears capable of putting them on a path to flourishing, the question I want to ask is whether this is the only path to flourishing. I suspect that some other avenues to flourishing may be foreclosed on in the creation of such a disciplined person.

So, none of this need be taken to dispute the content of Foot's claims, It only serves to show how narrow a conception of education, reasoning and attendant concepts are required by Foot's account. Indeed, even this may be expected since virtue is sometimes presented as in principle available to all, though in practice achievable by only the most dedicated individuals. Still, noticing the narrowness of what counts as virtuous coupled with how difficult this is to achieve does raise some internal tension for Foot. She strongly suggests that the virtuous life, for most people, will be an ordinary life. While it is true that few of us will ever find ourselves faced

[29] I have in mind here things like a tendency to over-think, feeling alienated from one's self or one's emotions, or finding a strong urge to second-guess and revisit one's choices.

[30] This closely tracks the points made about medical practice above. The costs and benefits emerge together because they are both directly linked to the conception in play.

with so taxing a situation as the Nazis' prisoners, given that her account of the virtuous life requires a person to think or reason in accordance with very specific standards one might wonder if the narrow conception given by Foot could realistically be put into practice while pursuing an ordinary life. I will leave that as an open question because settling it is beside the point. It is no criticism of a moral theory to point out that it demands difficult action.

However, the narrowness of the conception of a virtuous life could serve as a direct attack on a moral theory if that narrowness was understood as seriously constraining a full and valuable expression of human life. Recall the analogy with medicine. If the abstract conception of the human being which is adopted is too determinate it cannot accommodate the possibility of alternate conceptions. However, it is possible to not notice this flaw and so proceed with an overly stringent medical practice. In the examination of Foot there is evidence that an overly contentful abstraction of this type is being offered. First, it was noticed that the demands of virtue seemed best suited to someone in the prime of her life. Further, favorable circumstances and education seemed required for virtue to be achievable. So there is at least some prima facie evidence that Foot's account is problematically narrow.

Over turning this *prima facie* evidence would require showing that that which was excluded by the theory was in some way incompatible with true virtue or the expression of human abilities. Perhaps a defender of Foot could attempt just this kind of claim as a rebuttal. However, I want to try and head off that move by introducing two additional lines of thought that seem to me to foreclose on the possibility of a successful defense using those avenues. Further they bolster the case that I am putting forth. First, I will offer an argument which asserts that Foot's conception is too exclusive. Her picture is one of a deliberative, thoughtful agent drawing on her education and clear-eyed view of the circumstances to act well. I think that it is worth

considering a broader picture of human attributes and noting the unique and often non-rational contributions they make to the success of the agent. If it can be shown that arational faculties are able to contribute to successful human action, and can do a better job than those under rational control, this will strongly suggest that flourishing's dependence on practical rationality is artificially narrow and has little room to account for arational routes to human success. Second, it should be pointed out that rational deliberation has recently come under attack as being a resource intensive and not particularly reliable decision making tool. One source of this criticism comes from studies which argue that some cases snap judgments and other 'gut level' methods of problem solving and acting may be superior to explicit deliberation. The reason for this may be due to certain innate capacities that human being have for tasks like detecting deception, prevailing in argument, discovering breaches of general rules and the like.[31] While this empirical research is important, it is also controversial and the categories most easily studied are not always those that best lend themselves to philosophical argument. Instead I want to highlight a few cases of rational pathology. These cases show ways that rational thought can lead one toward less than ideal outcomes and shows how those outcomes can be improved by the interjection of arational actors. This is not to say that there is no role for rational consideration. It is obviously a very general tool that can be used in a staggering variety of situations. But if it gains its importance from its versatility, and not from any innate authority, and if the human being is a creature endowed with other modes of engaging with the world, this should lead one to be skeptical of the idea that practical rationality is the master-capacity for motivating human action.

[31] See, for instance, Woodward and Allman's 'Moral Intuition: Its Neural Substrates and Normative Significance', 2007.

While I do not have the room to consider all of these alternate arational routes to human flourishing I hope that an examination of a few can stand in for a fuller account of the value of these capacities. Consider parental love. This capacity for a unique kind of affection is remarkable for its immunity to rational scrutiny. Being a parent enables one to sacrifice in ways that might otherwise be impossible. It allows for intuitive kinds of knowing that deepen relationships, foster opportunities for growth and can act as a early warning signal that something is amiss. To these roughly positive qualities it is also important to consider the self-deception and defensiveness that sometimes accompanies the parental relationship. The arationality of the parent's love contains the ingredients for incredible depth of feeling and for being at odds with reality.

But the unique relationship that holds between parents and children seems, to me, to be better for its irrationality. That is, strictly considered, it seems that one's devotion and dedication to one's own child is not easy to offer a rational justification for. This is not to say that rational considerations are unable to motivate depth of feeling engendered by less scrutinized motives. However, it is difficult to imagine that the level of devotion sustained as a result of rational considerations could be of the same character as the emotional. For whatever rational basis is offered would still be beset by qualifiers and doubts about the extent to which it would be rational or required to perform in ways that the emotionally devoted do. In this respect I agree with Foot that an action motivated by rational considerations is of a different character than a superficially similar action which stems from arational motives. What I mean to ask is whether a plausible picture of flourishing could have room for these arational motives. So, were one to engage with their children on a strictly rational basis that basis would be different for being rational than the kind of more emotional engagement even if it was similar in other respects.

To be clear, I don't think the issue here is one of what actions are performed but, instead, concerns the reasons why some action or other was performed. Strictly speaking, of course, Foot's commitment to the inseparability thesis builds the justification into the action's description. Even actions which appear 'the same' in the moment won't be treated as such by Foot if they don't emerge from the same motivating forces. So even if it were the case that a parent could be aroused by reason to do exactly the same actions as the emotionally devoted parent does out of an unconsidered motivational response, these acts wouldn't be seen as equivalent by Foot. Further, it is reasonable to suspect that rational considerations just wouldn't come to motivate a person in exactly the same way as another motivated in some other way. Foot's theory is one that requires that actions be motivated according to very strict criteria. [32] However, there is reason for thinking that these arational modes of engagement can contribute to the value of the relationship. For instance, the increase in trust that the arational mode of engagement allows could create a relationship that feels particularly safe to the child. The expectation that one's word is sufficient for belief, even in the face of evidence to the contrary, makes for a special kind of relationship. Further, the offer of unconditional love and support seems to be at odds with a mode of thinking which always asks after justifications and sets limits on how far certain attitudes are warranted. None of this is to suggest that there is not some potential for abuse in these relationships, but as has been shown, this is also a feature of rational conduct and the presence of that potential does not serve to impugn rational thinking generally.

In this example it seems clear that something other than rational considerations help to shape the parent child relationship. Foot's requirement that virtuous action emerge from considerations of practical rationality appears to artificially constrict the motivational precursors

[32] Allison Jaggar's 'Feminist Politics and Human Nature' contains an excellent discussion of the way that rational considerations can come to undermine emotional relationships by altering their character.

to action that she is willing to take seriously. In so doing Foot's theory is failing to accord a role to what seem to be natural capacities for action and motivation that are taken up some of the most basic aspects of human life. Instead of relying on these, she proposes that practical rationality step in and provide a roughly similar service. For all of the supposed similarity, the motivations provided by rational considerations are fundamentally different according to Foot. But this example challenges her claim that emotional motivation is unable and unsuited to underwriting action which contributes to the flourishing human life. Foot's theory narrows the range of possible motivations that can contribute to flourishing. In doing so she also exposes herself to the charge that her theory's stringency can be traced to a distorted view of human life and those capacities that exist for bringing about successful action.

The example of the parent child relationship seems to show that there is value conferred on a relationship that stems from the arationality within it. But cases that show this need not be limited to relationships. Religious faith is obviously capable of adding value to a life. Further it is able to compel action from faith that otherwise just couldn't be motivated. And, to describe these religious attitudes as a form of faith is precisely to mark their relative immunity to rational scrutiny. Actions performed from the motive of faith can give deeper meaning to the life of the agent performing them and have wider, and generally positive effects on the community at large. Another source of value that might be seen as at odds with rational thinking is the power of sentiment. A sentimental attachment to some person or object seems to require that the object is given an outsized personal importance. So these will be cases where the depth of feeling associated with the object will go beyond what is warranted by a more neutral consideration of the object's worth. Again I take it to be obvious that the feeling associated with sentimental

attachment can motivate people in surprising ways and the actions that spring from those motivations can confer value on the agent's life.

In addition to being able to bring about value directly, arational states can sometimes break the stalemate that purely rational actors with opposed interests can develop. It is well known that the cold-war doctrine of mutually assured destruction was vulnerable to irrational actors. Because the stability of this situation is imposed by a kind of mutual blackmail, all that is needed to break this stability is an actor whose desire to destroy the enemy is stronger than his desire for self-preservation. Indeed there is evidence that Nixon and Regan both tried to make themselves appear irrational to Russian diplomats in order to strengthen the position of the US at the bargaining table. Of course, one hopes that this was a ruse since the stakes were so high. But if it makes sense to feign irrationality in order to improve one's bargaining position in a high stakes game, then there will be cases in which irrationality itself can also accomplish this goal. Consider a lower stakes version of the same kind of stand-off. It is easy to imagine a scenario in which a rational actor has some inkling of his partner's infidelity that he is unable to confirm. Further, if his partner recognizes that he is only going to be motivated to act when he has strong evidence, and the partner wishes to continue the illicit activities, the partner can ensure the stability of the situation be simply refraining from producing strong evidence of the dalliance. But this whole dynamic changes if the actor is not rational. If the actor is subject to strong feelings of jealousy this may motivate him to act on a mere inkling rather than strong evidence. Further still, if the actor's partner realizes this and realizes that the response, because of the jealous streak, may be disproportionate to the actual offense, the unfaithful behavior is less likely to develop in the first place. The partner will come to reasonably believe that the chances of avoiding the consequences of being unfaithful are significantly lower than when paired with a

rational actor and this prevents the dynamic from developing in the first place. The irrational behavior of a single party is able to halt the descent into rational pathology.

It might be thought that this view unfairly disparages rational thinking. I intend no such thing. Indeed I take myself to be agreeing with Foot's basic point that the character of action and the motivation that precedes it is altered by the kind of thought that generated it. Where I depart from her view is in thinking that arational forms of thought are not defective forms of rational deliberation but are modes of sensitivity and engagement with the environment that have the potential to contribute to a well-lived life. Whether these are emotional responses to situations or simply snap-judgments made under conditions with limited time and information resources, the common element is that these decision makers do not seem to be reacting to reasons as reasons. That is, they are not considering discrete aspects of the situation, weighing those against other aspects and considering what actions are most favored by consideration of these points. That it is possible to recognize non-deliberative capacities for thought and action and that these capacities have the potential to guide an agent toward successful outcomes strongly suggests that the focus on practical rationality, and its requirement that reasons always be treated as reasons, is too stringent to serve as the sole model for a flourishing human life.

It may seem as though the position I am suggesting is to break radically with the traditional view of virtue. To be sure, the central concern for rationality has been a feature of many virtue theoretic views dates back to antiquity and so it is not just the contemporary literature which seems to have regarded practical rationality as the master virtue. However, I would like to suggest that there is precedent for the type of theory I am advocating. It can be found in Plato's *Republic*. Plato recognizes that the rational life is not suited to everyone. This is because not everyone has the capacity for the high-level abstract thought that he believed full

rationality required. It also seems clear that Plato thought something approximating the life of virtue was available to those who were not lucky enough to posses the intellect required of the fully rational life. Those citizens of Kallipolis that were not endowed with a powerful intellect could, through: trust in authority, a cultivated frugality and an appreciation for the community nature of the city, come to participate in a well-lived personal and civic life. To be sure, Plato never claims that this kind of prosaic life is the most pleasant of all possible lives, but he does suggest that such a life is the most pleasant of those available to these citizens. So while he retains the terms 'virtue' for the guardians' exemplary mode of living he does recognize that the best life available to a person is something to be valued and aimed at in daily life.

What makes this possible in *Republic* is two-fold. First, the view of the human being that Plato arrives at is one which doesn't treat the dominance of the deliberative faculty as common to all people or always capable of motivating the right choices in those it does exist in. Second, Plato treats the life of virtue as an extraordinary life that threatens to chafe against the desires and instincts of even dedicated guardians. This means that his injunctions are not universal to all people but instead are only applicable to a particular class. It is true that he refuses to see the lives of those outside of the guardian class as fully-virtuous, nevertheless he does not neglect them. Instead he offers advice that aims at achieving the most virtuous life available to them. And it is this aspect of his theory which I see the most potential in developing further. His view of the human being is not singular, it makes room for the recognition that, because of circumstance or character, it won't be possible to offer the same recommendations or the same ends to each person. It becomes possible for each person to flourish, to the best of their ability, despite being unable to achieve full virtue. This might seem to stand as an obstacle to development of virtue theory in this direction, but further examination might remove this.

Simply put, Plato's view of fully-virtuous flourishing is awfully strange. Intellectual work does provide some satisfaction and pleasure, but it is hardly of the permanent and nourishing character that Plato suggests and it certainly doesn't seem to always outweigh the pleasures of family life, sex, and private property. Perhaps then, it is possible to see Plato as something of a chauvinist in his enthusiasm for this ascetic style of life. If so, this removes some of the attractiveness of seeing this style of life as superlative and that in turn makes it more plausible to regard the life of the lower classes as displaying a kind of virtue because they flourish to the extent that they are able to. Plato begins with a multiplicitous view of human beings, but one variety is isolated and elevated. But there is no need to follow him in this move. By retaining the view of human beings as varied creatures with different ends suggested by differing characters and circumstances it is possible to go on to define a kind of flourishing for each and so replace the singular definition of virtue with something more open-ended. This is the kind of open-ended generalization that the examination of medicine suggested could function as a fully-general conception of the human being.

Contrast this with Foot. She sees virtuous life as ordinary life but then focuses on a conception of the person which is quite singular. While she convincingly argues that rationality is one route to living well, she doesn't see other capacities as able to even approximate this life. This is a puzzling aspect of her theory for it is by analogy with the oak tree and the deer that 'good life' for the human being is supposed to get its force. But it would be wrong to focus on a single trait of deer life, such as swiftness, and argue that this was the sole criterion of good deer behavior. I hope to have shown that a similarly focused concentration on rationality is just as exclusive and so must be broadened.

BIBLIOGRAPHY

Anscombe, G.E.M. *Intention*. USA: Harvard University Press, 2000.

Aristotle. *Nicomachean Ethic*. Trans. Irwin Terence. Indianapolis: Hackett Publishing, 1985.

Broome, John. *Ethics out of Economics*. Cambridge: Cambridge UP, 1999.

Davidson, Donald. *Inquiries into Truth and Interpretation*. Oxford: Clarendon Press, 2001.

- - -. "Belief and the Basis of Meaning". Rpt. Davidson 2001, 141-155
- - -. "On the Very Idea of a Conceptual Scheme". Rpt. Davidson 2001,183-199
- - -. "Radical Interpretation". Rpt. Davidson 2001, 125 -141

Davidson, Donald. *Problems of Rationality*. Oxford: Clarendon Press, 2004.

- - -. "Deception and Division" Rpt. Davidson 2004, 199-213
- - -. "Incoherence and Irrationality" Rpt. Davidson 2004, 189-199
- - -. "Paradoxes of Irrationality" Rpt. Davidson 2004, 169-189
- - -. "Who is fooled?" Rpt. Davidson 2004, 213-230

Davidson, Donald. *Subjective, Intersubjective, Objective*. Oxford: Clarendon Press, 2002

- - -. "A Coherence Theory of Truth and Knowledge" Rpt. Davidson 2002, 137-158
- - -. "First Person Authority" Rpt. Davidson 2002, 3-14
- - -. "Knowing One's Own Mind" Rpt. Davidson 2002, 15- 38

Dennett, Daniel. *The Intentional Stance*. Cambridge MA: MIT Press, 1996.

Dennett, Daniel. "The Interpretation of Texts, People and Other Artifacts" *Philosophy and Phenomenological Research*, Vol. 50. 177-194, 1990.

Dent, N.J.H. *The Moral Psychology of the Virtues*. Cambridge: Cambridge UP, 1984.

Dutton, D.G. and Aron, A.P. "Some Evidence for Heightened Sexual Attraction Under Conditions of High Anxiety" *Journal of Personality and Social Psychology*. Vol. 30. No. 4. 510-517, 1974.

Foot, Philippa. *Natural Goodness.* Oxford: Clarendon Press, 2001.

Gazzaniga, M.S. and LeDoux J.E. *The Integrated Mind.* New York: Plenum, 1978.

Grandy, R. "Reference Meaning and Belief" *The Journal of Philosophy*, Vol. 70, No. 14, 439-452, 1973.

Hursthouse, Rosalind. "Arational Actions" *The Journal of Philosophy*, Vol. 88, No. 2, 57-68, 1991.

Jaggar, Allison. *Feminist Politics and Human Nature*. Rowan and Littlefield Pub, 1988.

LeDoux Joseph E. and William Hurst, eds. *Mind and Brain: Dialogues in Cognitive Neuroscience*. Cambridge: Cambridge UP, 1986.

Lewis, David. "Radical Interpretation" *Philosophical Papers Vol. 1*. Oxford UP, 1983.

McCulloch, Gregory. *The Life of the Mind*. London: Routledge Press, 2003.

Mele, Alfred R. *Self-Deception Unmasked*. Princeton: Princeton UP, 2001

Moran, Richard. *Authority and Estrangement*. Princeton NJ: Princeton UP, 2001.

Plato. Republic. Trans. Reeve, C.D.C. Indianapolis: Hackett Publishing, 2004.

Prinz, Jessie. *Gut Reactions*. Oxford: Oxford UP, 2004.

Quine, W.V.O. *Word and Object.* Cambridge MA: MIT Press, 1970.

Quine, W.V.O. "Ontological Relativity" *Ontological Relativity and Other Essays*. New York: Columbia UP, 1969

Rorty, Amelie. "Enough Already with 'Theories of Emotion" Solomon 269 -278

Sacks, Oliver. The Man Who Mistook his Wife for a Hat and Other Clinical Tales. New York: Harper and Row, 1987.

Savulescu, Julian. "Genetic Intervention and the Ethics of Enhancement". *Ethical Issues in Modern Medecine.* Ed. Steinbock, Bonnie, John Arras and Alex John London. Mayfield Publishing, 2006.

Solomon, Robert. *Thinking about Feeling*. Oxford: Oxford UP, 2004.

Velleman, J. David. *Practical Reflection*. Princeton: Princeton UP, 1989.

Velleman, J. David. *The Possibility of Practical Reason*. Oxford: Clarendon Press, 2000.

- - -. "On the Aim of Belief". Velleman 244-281
- - -. "The Guise of the Good" Velleman 99-122

Wedgwood, Ralph. *The Nature of Normativity*. Oxford: Clarendon Press 2007.

Williamson, Timothy, *The Philosophy of Philosophy.* Malden, MA: Blackwell Publishing, 2007.

Wilson, T., *Strangers to Ourselves: Discovering the Adaptive Unconscious.* Harvard University Press, Cambridge, MA, 2002.

Woodward, J. and Allman J.M. "Moral Intuition: Its neural substrate and normative significance". *Journal of Physiology Paris.* In Press. (http://saki.caltech.edu/publications.html)