MODELING RATIONALITY:

A NORMATIVE OR DESCRIPTIVE TASK?

3For final version, see *Modeling Rational and Moral Agents*, ed. Peter A. Danielson, Vancouver Studies in Rationality, vol. 7. Oxford University Press, 1998, pp. 119-134.

The Rationality of Suicide: Two problems.

Suppose your friend, who is wholly committed to rationality, proposes to commit suicide. How -- without moralizing or exhortation -- might you try to dissuade her?

The standard method, springing from the type of Bayesian decision theory propounded by Richard Jeffrey (1965) and others, goes something like this.

First establish, on the basis of past decisions, a general picture of the subject's preference rankings. Then compare the present choice to that established picture. If they match, that indicates the choice is rational. If not, it isn't.

Obviously, this procedure makes several assumptions: that the subject hasn't changed her mind; that the inference from choices or reports to the original preference ranking was accurate; that the information from present choice is correct, etc. These might well be mistaken. They might also be virtually impossible to verify. But these are not problems I want to dwell on here. The issues I want to draw attention to stem from two additional considerations.

First, suicide seems especially resistant to rational assessment. The past, in this case, can strictly speaking afford little guidance as to the rationality of the present. The reason is that the choice of death is a choice that negates *all* the alternatives among which previous choices were made. Previous preference for a Chevy over a Lancia, or for eating lemon ice over viewing a Picasso, can provide no guidance where one alternative is *neither*. Nor can the choice of

pleasure over pain afford comparison with *nothing*. So here, it seems, the Bayesian schema loses its grip. Yet surely the choice of suicide can sometimes be assessed for rationality.

The second problem is that suicide involves, among other considerations, an assessment of what we might call *foundational values*: values -- if there are any such -- which are chosen for their own sake and not as a consequence of their relation to any other values. Is there a special way in which such foundational values are related to the Bayesian scale? And is there any adequate evidence embodied in past choices that can determine uniquely what they are?

Normative/Descriptive Ambiguity in Rational Models.

These questions are actually a symptom of a wholly general problem. Roughly put, the problem arises from the fact that any normative model of rationality presupposes a corresponding descriptive model. In the case just considered, the model lost its grip precisely because there was no applicable description of previous suicide behavior. In fact, we should rather say that every normative model is *identical* with a descriptive model: the difference depends merely on context of use. Yet they may draw us in different directions. In one mode, I infer my subject's preference ranking from her actual past decisions. In another, I infer from her choices that she is irrational, on the ground that they are inconsistent with her preference rankings as previously ascertained. (This is what we can't do in the case of suicide). But what guarantees that the original assignment was correct? Perhaps it was a mistaken inference, reflecting irrationality in the original choice. In *inferring* from it to the existence of a certain preference structure, I must assume it was rational. This assumption rests on some sort of "principle of charity'' (Quine, 1960; Davidson, 1982). Otherwise I might not have been able to make it look coherent. On the other hand, when I use the scheme to *criticize* irrationality, I take a previously established structure for granted. Then, for my inability to fit the subject's present choice into that structure, I choose to blame my subject rather than myself.

This ambiguity of the normative and the descriptive is pervasive in efforts to model human rationality.[1] But what exactly is the relation between them? The history of philosophy affords examples of two reactions: *naturalism*, which attempts to reduce the normative element to some sort of natural process, and *normativism*, which claims that within all attempts to model actual reasoning processes there must be an ineliminable element of normativity.

Naturalism must be distinguished from two neighboring positions.

First, it mustn't be confused with physicalism. A functionalist, for example, need not be a physicalist with respect to mental entities, while still claiming that all there is to be said about the norms of rationality can be accounted for in terms of natural truths. What marks out truth (as the proper object of belief) from the proper objects of other mental states is that truths alone must pass the test of simple consistency. The proper objects of other mental states, such as wants, need pass no such test. Consistency does not require that the propositions we *desire* be *possibly all true* together: it is sufficient that they be *possibly all good* together.[2]

Second, however, the contrast between naturalism and normativism may not be assimilable to the fact-value distinction. That rationality involves norms should not preempt the question whether all norms of rationality refer to the good. I hazard no judgment on that question here.

---

1. . This is so not only in those models that apply to behaviour. Hintikka (1962) exemplifies the problem purely cognitive domain. His notion of "virtual consistency" evades the issue, however, since it presupposes complete logical transparency. But of course a description that would fit only an ideally consistent subject is merely factual description of any real belief set.

2. . I have argued this in de Sousa (1974). The gap between simple consistency and rational policy is attested by the lottery paradox. (Kyburg 1961 pp. 196 ff.) It seems rational to believe of each lottery ticket that it will lose, while believing that one will win. Yet these are strictly inconsistent. The question of rationality here concerns a policy about believing, not the closure of the set of propositions believed. The issue in the lottery paradox is whether the necessity that some of the propositions in that set be false -- which is not controversial -- constitutes a sufficient reason for holding that the policy of jointly believing all of them must be irrational -- which is.

Some related forms of the problem of naturalism.

The present problem is an ancient one. Its first avatar in our tradition is Plato's puzzle about error in the Theaetetus: If a representation (belief, reference, or perception) is the imprint on our mind caused by the thing or fact which is its object, then how can it ever be mistaken? (*Theaetetus*, 187d ff.) If for R to be a representation *of* A is to be *caused by* A, then R must either represent A correctly, or represent something else correctly, or else fail to represent anything at all. In no case can it erroneously represent A.

In more modern forms, the problem of misrepresentation has been widely discussed. (Fodor 1987; Dretske 1986; Millikan 1993) If I mistake a cow for a horse, doesn't this mean my word 'horse' really means (to me) 'horse-or-cow'? (In which case I wasn't mistaken after all.) A normativist might view this as a piece of evidence for the irreducibility of the normative. Once all the mechanics of causation have been accounted for, it will still be possible to draw the distinction between what is correct and what isn't.

To counter this, the naturalist's strategy must first be to show that the "rightness" or "wrongness" associated with biological functions can be accounted for without normative residue. Naturalism need not eliminate teleology: it need only tame it. We need to show that it makes sense to claim that something is *meant* to be this rather than that, without resorting to ineliminable normativity.

A couple of examples will remind you of the flavor of the resulting debates.

i) The "frog's eye" problem: if what sets off the frog's eye is a black moving speck, must we say that it just accidentally finds flies, or rather that *by* means of the capacity to detect moving specks, it *serves* to detect flies? (Dretske 1986, Millikan 1991).

ii) In Elliott Sober's sorting machine problem, a series of sieves finds green balls, because the balls's color happens to be correlated with size. But is the sorting machine to be

described as having the *function* of sorting green balls, or that of sorting for small ones -- which sometimes happen to be green? (Sober 1984, pp. 99 ff.)

On the first problem, Ruth Millikan quotes an unpublished tract of Fodor's as remarking: Nature "doesn't sort under any labels". (Millikan, 1991, p. 159) If that were strictly true, then the plight of naturalism would be far worse when we are judging rationality, for there labels are *almost* all. Without labels, there is unlikely to be any way of arriving at a sufficiently unambiguous ascription of belief or want to prove rationality or convict of irrationality. (de Sousa 1971) Even without positing that nature sorts under labels, however, we can hazard hypotheses about the "real" function of these processes, by looking at the *causal* origin of their teleology. The mechanism of teleology in these cases may be difficult to demonstrate conclusively, but it seems reasonable to assume that it's not *magical*: that there is some naturalistic explanation for it. Can we distinguish, from a biological standpoint, between stimuli that are apparently causally equivalent in actual situations?[3]

The answer favored by both Sober and Millikan is that we can, providing we delve into the history of the selection process. In the case of the ball-selecting toy, *what* the device selects are green balls, but what it selects *for* are small ones. The reason is that the color of the balls is causally irrelevant to the selection even though the effect of the selection is to select green ones. The selectionist equivalent for the frog's eye is this: the frog's

---

3. . In the face of counter-examples raised, e.g., by Boorse (1976) to the classic Wright-type analysis, (Wright 1973) Robert Nozick has suggested that genuine teleology has to obey a second order condition, combining Wright's insight that if G is the function of X it explains why X exists, with Nagel's analysis of homeostatic systems: "The Nagel and Wright views can be combined, I suggest, to present a more complete picture of function. Z is a function of X when Z is a consequence (effect, result, property) of X *and* X's producing Z is itself the goal-state of some homeostatic mechanism M satisfying the Nagel analysis, and X was produced or is maintained by this homeostatic mechanism M (through its pursuit of the goal: X's producing Z)." (Nozick 1993 p. 118, referring to Nagel 1961).

This condition excludes Boorse type counterexamples. It may, however, be too stringent, since in most cases of natural selection there is little reason to think the processes involved were homeostatic, insofar as that implies centering on some fixed point.

detection mechanism was selected *for* finding flies, but it selects mechanisms that find both flies *and* specks. We can safely insist that it must have an unambiguous meaning (even in the absence of labels subjectively assigned by a language speaking creature). We need only say that it *means* whatever it has been *caused* to find: " 'Selection of' pertains to the *effects* of a selection process, whereas 'selection for' describes its *causes.*" (Sober 1984 p. 100).

The moral is that evolutionary considerations are probably capable of assigning a definite function to some mechanisms or processes, without resorting to some sort of externally imposed normativity. One use of this idea, is that it might be possible to *explain,* in evolutionary terms, why we have certain propensities to follow given strategies. But can the reference to an evolutionary story actually avoid the question of normativity?

Some, in the tradition of Hume (1975) or Goodman (1983), have answered that it can. In fact this is arguably Hume's essential lesson: JUST SAY NO to the demand for justification. Instead, *change the subject*: don't ask *why* we do it, just ask *what* it is we do.

Sometimes rejecting a question is a good strategy. Witness Newton and Darwin: Newton's genius was to insist on *not* answering the classic question about what keeps the arrow in flight. Darwin's was to insist on *not* answering the classic question about what is the cause of biological diversity. After Newton, we don't ask why the arrow keeps going, we ask why it stops. After Darwin, we don't ask why living things are so diverse or why they fail to be true to type, we ask instead what makes them cluster around apparent types.

Many people feel cheated by Hume's answer: "So maybe I do it naturally, but why is this a reason to do it? Wasn't it you, Hume, who famously told us you can't go from an is to an ought?" There are obviously cases where *don't ask* is just an evasion. What makes it the right strategy in some cases and not in others?

One possible answer is that it must be the right strategy if it is *wired in,* i.e. if it is embodied in the system's "functional architecture". This is Pylyshyn's (1984) term for a level of explanation at which some mechanism carries out a function merely in virtue of its

physical configuration: given that it is set up in just such a way, physical laws have just that effect. The fact that we perform *modus ponens* is due to some such basic functional architecture. We *just do* what we are programmed to do.

But why doesn't this beg the question? Isn't it an empirical question whether natural selection results in mistakes?

Consider Richard Dawkins's (1982) discussion of the case of the digger wasps. Digger wasps hide paralyzed prey in burrows; when they fight over a burrow the time they spend fighting is proportional to their own efforts in stocking the borrow, not to the "true value" of the burrow measured in terms of the number of prey it contained. At first sight, these wasps appear to be committing the "sunk costs" or "Concorde fallacy." But this example too embodies the sort of descriptive/normative ambiguity I have been discussing. The digger wasps wouldn't be there if their policy hadn't worked out as well or better than available alternatives. So who are we to carp? Dawkins's recommendation is instructive: "assume that an animal is optimizing something under a given set of constraints ... try to work out what those constraints are."(Dawkins, 1982, p. 48) Sure enough, the digger wasps commit no fallacy under the constraints entailed by their epistemic position. In other words: the biologist's real task is to explain why the apparent alternatives that might have avoided the "sunk costs fallacy" were not actually "available alternatives". So even if one agrees that "sunk costs" reasoning is a normative mistake, the constraints of descriptive adequacy will not let us actually blame the wasp (or natural selection) for committing it.

If one can't even blame wasps for being irrational, what are the prospects of making charges of irrationality stick to the "rational animal" *par excellence*? Someone might object to this whole discussion that talk of evolutionary rationality is irrelevant. The questions we should be raising concern *rational agents*, where the word 'rational' just means 'capable of irrationality' Natural processes can maximize this or that parameter, but they can't exhibit irrationality.

The special rationality of persons, then, consists essentially in the capacity to be irrational. Some have argued that humans couldn't ever by *systematically* irrational. (Jonathan Cohen 1981; against him, see Stich 1990) But if so, what accounts for our actual irrationality in particular cases? And how, from a naturalist point of view, can we be convicted of such irrationality? If our models were strictly descriptive, would the appearance of irrationality not merely indicate inadequacy in the model? This, then, returns us to my original question: Are the models constructed to account for human rationality purely descriptive, or must they contain an irreducibly normative element?

Four Classes of Models

Models of rationality fall into two pairs of distinct classes: (i) *strongly* or (ii) *weakly compulsory;* and (iii) *weakly* or (iv) *strongly optional.* These cases are significantly different with respect to their origins, to the role played in their determination by natural selection, and to the way in which they are subject to the two problems of descriptive/normative ambiguity and foundational status. The third and fourth class are particularly interesting.

In the remainder of this paper, I propose first to tease out some characteristics of the two compulsory types, particularly the duality of descriptive and normative aspects, and then to examine the special role of emotions in relation to both sorts of optional principles.

(i) Strongly compulsory. Example: *modus ponens/tollens*. One who doesn't observe these rules is straightforwardly irrational. Nevertheless, a non-question-begging justification of the rule hasn't yet been stumbled on. This fact makes it plausible to argue that these principles are irreducibly and categorically normative. For if they were merely conditionally normative, one would be able to offer the conditions on which their prescriptive force depends.

The fact is, however, that even in strongly compulsory cases the normative force of the argument falls far short of "logical compulsion". To see this, consider *Modus Ponens*. Two facts stand out: the first is that in this case *normativity entail naturalism*. The second is that even in this case *no argument ever compels*.

Why Normativity entails naturalism. A naturalistic theory is exactly what we need at precisely the point where normativism is supposed to triumph. The proof lacks freshness, because it's really an amalgam of Hume (1975), Quine (1966), Goodman (1983), Lewis Carroll (n.d.), and Wittgenstein (1951).) But here it is, in terms of Carroll's classic dialogue between Achilles and the Tortoise

Achilles: "if p then q, and p. So you must accept q."

Tortoise: "WHY must I -- oh, never mind, I know you'll never satisfactorily answer that one. Don't even try. [Hume] Instead, let me accept this imperative WITHOUT JUSTIFICATION. Let me accept it, in fact, as a categorical imperative of thought. Or if that sounds too grandiloquent, let's just call it a CONVENTION. [Carnap (1956)] But, please, write it down for me.

Achilles: All right, then (writes):

p and (if p then q)

But

If p and (if p then q) and (if p and (if p then q) then q) then q.

See? NOW you MUST accept q.

Tortoise: I've agreed to your rule, but now how do I know that this is a RELEVANT INSTANCE for its application? [Quine] I need a principle of INTERPRETATION that will indicate to me when and how I must apply this categorical rule or convention that I have agreed not to question. [Wittgenstein].

Hence the Quine/Wittgenstein/Carroll dilemma:

> **EITHER** you will need to give me a rule of interpretation for every new case -- and then you will have a doubly exploding process: for each new case will not only require a new rule of interpretation but also an additional rule of interpretation to interpret the application of the rule of interpretation -- ad infinitum.
>
> **OR** the answer has to be at some point that *we don't follow a rule at all: we just naturally do this*. In short, it's just the way we are wired.

Why no argument compels. I have called *modus ponens* (and its converse) the *most compelling cases*, because their violation requires heroic twists in the application of the principle of charity. Even in this most compelling case, it's important to see that no one is actually compelled to believe the conclusion of a valid argument. Arguments are maps, not guides. The most any deductive argument can give us is a set of alternatives: believe the conclusion together with the premises, or continue to reject the conclusion, but then also reject one or more premises. And in this situation, what is the most reasonable thing to do? The most reasonable thing to do is, surely, to believe the least incredible alternative. But what can determine which that is? Since not everyone will agree, the relevant determinant must be something essentially subjective. At best it can be discerned, at the end of a process of reflection, by the place at which a "reflective equilibrium" is reached. But if we need to appeal to a reflective equilibrium even in the most compelling case, then *a fortiori* we shall need to understand what it is that guides our choice of rational strategies in other cases. That, I venture, is where the unique structural role of our emotional dispositions comes in. More of this in a moment. Let me first complete the sketch of my taxonomy.

(ii) Weakly compulsory. Sometimes, it seems that *sub species aeternitatis* there is a clear answer to the question of what is *the* correct way to interpret a given situation and produce a rational outcome. Some of the Kahneman-Tversky problems seem to be of this sort: the usual claim about them is that we *tend to make mistakes* about them. (Kahneman and Tversky 1982)

Another good example is a problem that has been around for a while:

Three cards are face down; one is an Ace, two are Kings: you don't know which. I ask you to put one finger on one card at random. (You may hope it's the Ace.) I then turn one other card up, which is a King. Now I ask you to bet on which of the remaining cards is the Ace: the one you had your finger on, or the other one?

It is tempting to reason: since there are just two cards, it makes no difference. You could switch or stay at random.

But actually, if you switch, you stand to win, if you stay, you stand to lose, two thirds of the time. For of all the times you start playing this game, your finger will be on the Ace just one third of the time.[4]

In cases such as these, it's clear that our intuitive answers are just plain wrong. It doesn't follow, needless to say, that "evolution failed us", since one can imagine constraints under which the decision procedure in question might turn out to be the best of all possible procedures. Besides, while we are bad at working out probability problems, we are actually quite sensitive to frequency differences in practice. (Whitlow and Estes, 1979) In some cases, principles such as "anchoring" or "representativeness" may involve significant savings of cognitive resources, and yield approximately correct results enough of the time to outweigh their disadvantages in the cases generally highlighted in the literature. (Kahneman and Tversky 1982)

In these compulsory cases, we might expect that once the problem is sufficiently well defined, we can give conclusive reasons for the superiority of one argument or method over another. This class of examples differ from the "strongly compulsory" ones in that they

---

4. . I don't know the origin of this puzzle, which has been around for some years. A version of it became widely known a few years ago as the Monte Hall problem. Hundreds of mathematicians and statisticians, it was reported, got it wrong (Martin 1992 p. 43.)

make no claim to foundational status. As a result, they admit of (conclusive) justification. I call these weakly compulsory because an argument is required to see that they are correct principles. (By contrast, as we just saw, in the case of *modus ponens* the cause is lost as soon as one starts asking for an argument.) Arguments in their favor will be normative in tone; but once understood, they will be seen to be as compelling as any argument can be -- within the limits just discussed. Anyone (including, notoriously, a number of "experts") inclined to dispute the standard solution to the Monte Hall problem can be invited to put their money behind their principle, and soon come up against the necessity of admitting that *either* it is time for them to give up the ordinary principles of induction, *or* they must take their monetary losses as evidence of their mistake.

(iii) Weakly Optional. Optional cases are those where no "compelling" (in scare quotes, because of the qualification just made that no argument is really so) solutions can be shown to be correct. There are two separate reasons why this might be so. In one case, there are alternative solutions that have the feel of an antinomy: equally compelling arguments seem to line up on either side of the issue. Newcomb's problem, for example pits dominance arguments (nothing wrong with them) against probabilistic arguments (nothing wrong with these either.) Yet the arguments's conclusions are radically incompatible. (Nozick 1969)

(iii) Strongly Optional. In other types of situation, there is no definitely or demonstrably right answer to the questions at all. When offered a choice between betting and not betting in a zero-sum game, for example, there is, ex hypothesi, no Bayesian reason to choose either. Such a decision is a paradigm case of the strongly optional. In other cases there are competing goals that cannot be reconciled at any permanently optimal point. In the ethics of belief, for example, the two competing goals are "maximize truth", and "minimize falsehood". Either goal could be fully satisfied at the expense of completely ignoring the other. So any policy is in effect a compromise between contrary risks.

Still other cases seem to involve rules that are reasonable under certain evolutionary constraints. These are reminiscent of biological cases such as the digger wasp. Principles such as anchoring or representativeness may belong here rather than in the compelling class. For having a stable policy enabling quick decisions may lead, in the long run, to better results even if the policy is a relatively coarse one. As Mill once put it, a sailor would not get on better by calculating the Nautical Almanac afresh before every turn of the tiller. This consideration casts doubt on the claim that these are definitely *mistaken*. (Mill, 1971)

The necessity of biological economics

In all the above cases, it might be tempting to claim that there is no real possibility of discovering that the processes of evolution are "irrational". The normative correctness of these processes are built into the conditions of adequacy for their description. The reason is that the economic model is *more straightforwardly applicable in biology than it is in economics itself*. For economics applies to people only insofar as they can be construed as economic agents -- an idealization. In biology we can give the model a literal interpretation: probability of this gene's reproduction can simply be taken as the actual frequency (in some run considered long enough) of the gene, and the benefit can be interpreted as the difference between this frequency and the corresponding future frequency of its alleles (or some other acceptable measure of fitness). If there were constraints that prevented an organism from attaining some "ideal" condition, these are automatically included in the equation.

There is one qualification, however. Sometimes, we can see something in nature that we would have to rate "plainly irrational" if we thought God had invented it. That's because if God had invented it there would have been no special constraints on the mode of its engineering.

Take, for example, the ratio between the sexes among vertebrates in general and primates in particular. If you were God, you would surely arrange for that to be as close to 0:1 -- to parthenogenesis -- as was compatible with the gene-mixing function of sex. (Grant, for the sake of this almost-serious argument, that hypothesis about the function of sex) (Williams 1975). In a stable environment, the best bet would be to settle on a satisfactory model. That means parthenogenesis. A parthenogenetic species needs only half the resources for every offspring produced, and moreover the offspring, being clones, are of guaranteed quality. In unstable environments, however, all clones might be threatened at once. So we need variation, kept up by the gene shuffling of sex, to increase the chance of there being some variant pre-adapted to the new conditions. Males, however, are notoriously murderous and wasteful, and their presence in such large numbers clearly manifests that this is not the best of all possible worlds. One in a hundred would easily suffice. But the trouble is that the *mechanism* that secures the actual ratio takes no account of the normative considerations just adduced. It secures the result purely mechanically, for if there is a tendency for the genes to favor one sex, the members of the other immediately acquire an advantage, in that they will, on average, necessarily have occasion to contribute their genes to a larger number of members of the future generation.

Optional-Foundational Principles

In the case of human policies, in contrast to the biological cases just described, any constraints placed upon us by the facts of natural life limits only what we can do, not what we can judge to be desirable. Here, perhaps, is the crucial difference between biological models and models of genuine intentional behavior. Can we make a clear distinction between biological principles of rationality and those that are determined at the level of the actual life of the individual?

Cognitive science commonly posits two levels of brain programming. First level evolutionary programming determines innate operational mechanisms; its advantages are stability and early access. Second level evolutionary programming acts via a first level capacity for learning. Its advantages are flexibility, power, and lower evolutionary cost. These advantages outweigh the disadvantage of complete dependence at birth, as well as the risks entailed by the need for a developmental process crucially contingent on environmental circumstances. The hypothesis I want to advance is that in the case of the "optional" types, the range of principles of rationality available actually corresponds, at least in certain cases, to *emotional dispositions* that determine the *framework* of rationality rather than its *content*. They represent an analogue, in the sphere of evaluation and reaction in real life situations, of the evolutionary tradeoff just alluded to: when in the grip of an emotional state, we tend to act with a speed that easily becomes haste; our behavior tends to be stereotypical, but it is generally efficient. Emotions, like instinctual behavior patterns, trade flexibility for early access to a response. We can therefore think of them as analogous to genetic constraints on rationality, without being committed to the view that they are genetically programmed. innate

This, then, is the simple idea I want to promote: it is that the pattern of dual programming (the first for fixed responses and the second for the learning of new ones) can be extended to illuminate the "optional" principles of rationality. A partly acquired emotional repertoire might provide substitutes for innate rationality principles such as the ones governing strongly compulsory practices such as *modus ponens*. Thus, for example, an emotional fear or love of risk can determine the choice between betting and not; a tendency to attachment might anchor the policy of anchoring; and some sense of emotional identification with a certain kind of scenario might promote the policy of "representativeness." Moreover, emotions present characteristics peculiarly well suited to *temporarily mimicking* the rigidity of constraints acting on the economic models of evolution.

Emotions as Foundation-Substitutes

There are a number of parallels between the temporary role played by emotion in the determination of our cognitive strategies, and the evolutionary role illustrated above in the case of the structures determining functional architecture. I noted above that even in the *most compelling* case, the choice of what to believe is actually to be determined by subjective factors. In less compelling cases, such as the case of suicide, or of temptation (see Elster 1979, Ainslie 1993, Nozick 1993), present emotion is pitted against future or past emotion in ways that reflect individual "temperament" (our wired-in emotional dispositions) but also another source of variability tied to factors that depend on individual biography. In all cases, however, the emotions seem to play at least the following roles: (de Sousa 1989)

-they filter information, to the point of temporary exclusion of normally relevant facts;

-they offer strong motivational focus;

-they have quasi-foundational status, and can therefore be modified only by the kind of reflection that can change a reflective equilibrium. This is why emotions are said to "transcend reason" insofar as the latter is the mere working out of the means to the emotionally fixed goal. (*Reason is and ought to be*, in the words of Hume, *nought but the slave of the passions*.)

To grasp the significance of this last point, it's important to see that emotions are not merely desires. They have motivational force of some sort, to be sure, but that force is structurally different from that of desires, because of their "foundational" status. Let me explain.

The recently dominant Bayesian-derived economic models of rational decision and agency are essentially assimilative models -- two factor theories, which view emotion either

as a species of belief, or as a species of desire. They make is look as if all behavior can be explained in terms of a suitable pair (or pair of groups) taken from each category.

That enviably resilient Bayesian model has been cracked, however, by the refractory phenomenon of *akrasia* or "weakness of will." In cases of akrasia, traditional descriptive rationality seems to be violated, insofar as the "strongest" desire does not win, even when paired with the appropriate belief (Davidson 1980). Emotion is ready to pick up the slack: it determines what is to count as input into the Bayesian machine. Emotions are often credited with the power to change beliefs and influence desire. But they can play a determining role even without doing either of these things. By controlling attention, emotions can fix, for the duration of what we suggestively call their "spell", what data to attend to and what desires to act on, without actually changing our stock of either beliefs or desires.

The Normative Factor in Emotions.

It remains to sketch how the role of emotions in framing optional rational strategies is compatible with naturalism, and how it can explain the appearance of irreducible normativity.

The hypothesis I have just sketched could be rephrased in these terms: that our repertoire of emotions constitute the *temporary functional architecture* of a given person's rationality rules. The way that they are built up involves playing out basic scripts or "paradigm scenarios," in terms of which the emotion is in effect *defined.* Individual temperament plays a crucial role in the writing of these scripts, and individual differences in temperament account for a good deal of the individual differences between scripts. But so, of course, does individual history: since roles are first played out in social contexts (albeit a society that may consist only of child and caretakers), these are also in large part conditioned by social sanctions. We learn to "conform", we learn to "rebel". Neither

concept makes sense without a social *norm*. And these norms are subjectively experienced as if they had objective reality. Felt norms, while in the grip of an emotion, will be experienced as compulsory, and so appear to be categorically normative: hence the temptation to reject naturalism. The qualification, *categorically* is important, because conditional injunctions don't really pose a problem for naturalism: those who insist on irreducible normativism need *categorical* imperatives. It may be difficult to prove that a certain strategy is best given certain goals, but it is at least clear how this could be a purely factual question. The difficulty of doing any more has led to a tradition of thinking of reason as essentially limited to the elaboration of means (Wiggins 1976). As we saw in the case of suicide, the hardest cases for naturalism are those that involve foundational choices, i.e. choices that are not themselves conditional on pre-existing choices. But once we see the emotions as forming the framework of our deliberations and the limiting conditions of our rational strategies, it is no longer surprising that there should be a category of "optional" models, strongly backed by social norms, which in certain situations might be experienced as categorical norms.

But what, in turn, is a social norm? I conclude by hazarding a coarse guess. A social norm is nothing more, I venture, than a collection of facts about the individual reactions (actual and counterfactual) of individual members of the society. To be sure, those individuals will refer to a norm in their reaction. That is because their reactions are partly internalized as immediate emotional states, which are experienced as guided by norms. But there is no reason to take this experience at face value. For if the norm itself is merely embodied in further counterfactuals about the reactions of members of the society, there is a self-feeding loop here that is capable both of accounting for the powerful appearance of irreducible normativity, and of explaining it away, as reducible without remainder to natural facts.

REFERENCES

Ainslie, G. (1992). *Picoeconomics: The strategic interaction of successive motivational states within the person*. Cambridge: Cambridge University Press.

Boorse, C. (1976). Wright on functions. *Philosophical Review, 85*, 70-86.

Carnap, R. (1956). *Meaning and necessity: A study in semantics and modal logic*. Chicago: University of Chicago Press, Phoenix.

Carroll, L. (n.d.). What the Tortoise said to Achilles. In *The complete works of Lewis Carroll*. New York: Random House: The Modern Library.

Cohen, J. L. (1981). Can human irrationality be demonstrated? *Behavioral and Brain Sciences, 4*.

Davidson, D. (1980). "How is weakness of the will possible?". In *Essays on actions and events* (pp. 21-43). Oxford: Oxford University Press, Clarendon.

Davidson, D. (1982). *Inquiries into truth and interpretation*. Oxford: Oxford University Press, Clarendon.

Dawkins, R. (1982). *The extended phenotype: The gene as unit of selection*. Oxford: Oxford University Press.

de Sousa, R. (1971). How to give a piece of your mind, or the logic of belief and assent. *Review of Metaphysics, 25*, 51-79.

de Sousa, R. *The rationality of emotion*. MIT Press, 1989. A Bradford Book.

Dretske, F. Misrepresentation. In R. J. Bogdan (Ed.), *Belief: Form, content and function.* Oxford, New York: Oxford University Press.

Elster, J. (1979). *Ulysses and the sirens: Studies in rationality and irrationality.* Cambridge: Cambridge University Press.

Fodor, J. (1987). *Psychosemantics*. Cambridge, MA: MIT Press. A Bradford Book.

Goodman, N. (1983). *Fact, fiction, and forecast* (4th ed.). Cambridge, MA: Harvard.

Hintikka, J. (1962). *Knowledge and belief*. Ithaca, NY: Cornell University Press.

Hume, D. (1975). *Enquiry concerning human understanding* (L. A. Selby-Bigge, Ed & introd.) (P. H. Nidditch, Revised by & notes by) (3rd ed.). Oxford: Oxford University Press, Clarendon.

Jeffrey, R. C. (1965). *The Logic of Decision*. New York: McGraw Hill.

Kahneman, D., & Tversky, A. (1982). On the study of statistical intuitions. In *Judgment under uncertainty: Heuristics and biases*, ed. Kahneman, D.; Slovic, P.; Tversky, A. Cambridge and New York: Cambridge University Press.

Kyburg, Henry. (1961). *Probability and the logic of rational belief*: Middletown, CN: Wesleyan University Press.

Martin, R. M. (1992). *There are two errors in the the title of this book: A sourcebook of philosophical puzzles, problems, and paradoxes*. Peterborough, Ontario: Broadview Press.

Mill, John Stuart. (1971) *Utilitarianism. In Essential works of John Stuart Mill, ed. Max Lerner. New York: Bantam.*

Millikan, R. *(1993). White Queen psychology and other essays for Alice*. Cambridge, MA: MIT Press. A Bradford Book.

Millikan, R. (1991). Speaking up for Darwin. In *Meaning and mind: Fodor and his critics* (pp. 151-165). Oxford UK and Cambridge MA: Blackwell.

Nagel, Ernest (1961) *The structure of science*. New York: Harcourt, Brace and World.

Nozick, R. (1986). Newcomb's problem and two principles of choice. In *Essays in honor of Carl G. Hempel*. Dordrecht: Reidel.

Nozick, R. (1993). *The Nature of Rationality*. Princeton: Princeton University Press.

Pylyshyn, Z. (1984). *Computation and cognition*. Cambridge, MA: MIT Press. A Bradford Book.

Quine, W. V. O. (1960). *Word and object*. Cambridge, MA: MIT Press.

Quine, W. V. [. (1966). Truth by convention. In *The ways of paradox and other essays*. New York: Random House.

Sober, E. (1984). *The nature of selection: Evolutionary theory in philosophical focus*. Cambridge MA: MIT Press.

Stich, S. (1990). *The fragmentation of reason*. Cambridge, MA: MIT Press.

Whitlow, J. W. Jr., & Estes, W. K. (1979). Judgment of relative frequency in relation to shifts of event frequency: Evidence for a limited capacity model. *Journal of Experimental Psychology: Human Learning and Memory*, 5, 395-408.

Williams, G. C. (1975). *Sex and evolution*. Princeton: Princeton University Press.

Wiggins, D.R.P. (1976) Truth, invention and the meaning of life. *Proceedings of the British Academy 62:331-378.*

Wittgenstein, L. *(1958). Philosophical investigations* (G. E. M. Anscombe, trans.). New York: Macmillan.

Wright, L. (1973). Functions. *Philosophical Review, 82,* 139-168.