WISE CHOICES, APT FEELINGS

A Theory of Normative Judgment

ALLAN GIBBARD

Harvard University Press
Cambridge, Massachusetts

1 • The Puzzle

Why ponder our lives? At one extreme, the question is not a live one. We are a pondering species—and not each by himself; we are conversants. Silence is a discipline; too much is torment, as children learn in school. Sometimes we discuss earnestly, and in any case we banter and tease, quarrel and sulk. We gossip and tell stories, with verve if we can. Do these things engage us because they have point? Are they ways of working matters through with each other, or playing them through? Even when we are not pondering outright, we are caught up in equivalents.

With human beings it has always been so. The !Kung of the Kalahari are hunter-gatherers, and perhaps in them we can see what our hunting-gathering forebears were like. "Conversation in a !Kung encampment is a constant sound like the sound of a brook, and as low and lapping, except for shrieks of laughter. People cluster together in little groups during the day, talking, perhaps making artifacts at the same time. At night, families talk late by their fires, or visit at other family fires with their children between their knees or in their arms if the wind is cold."[1] The !Kung criticize each other, they gossip, they make oblique hints; they tell about events, about comings and goings, and about past hunts. They plan their hunts, and the successful hunter may consult on the proper distribution of his kill. Occasionally they quarrel, and frequently they talk about gifts and their suitability.

Conversation, then, is far more than a carrier of information. In talk we work out not only what to believe about things and events and people, but how to live. We work out how to feel about things in our

1. Marshall (1976, 351–352). The !Kung San live in Namibia (South West Africa); the San (or Bushmen) are among the few remaining hunting-gathering peoples.

lives, and in the lives of others. Not that we strive always to get to the root of things: we think not so much how to live and to feel on the whole, but about one thing or another, as it catches our attention.

Socrates was different. Plato has him saying to the jury, "If . . . I tell you that to let no day pass without discussing goodness and all the other subjects about which you hear me talking and examining both myself and others is really the very best thing that a man can do, and that life without this sort of examination is not worth living, you will be even less inclined to believe me. Nevertheless that is how it is" (*Apology*, 38a). When Socrates spoke of a life without such examination, he warned of a way a person might really live, that most people perhaps do. Socrates drank the hemlock for pressing too far these questions of how to live. What he did, and the Sophists before him, was start from the materials of common thought and speech, and refine them. Euthyphro could speak of piety, and he could reason about it; he could not have lived the life he did without reasoning. We can be tripped because we walk, and Euthyphro could be tripped because he reasoned, and had to.[2] Socrates' path is avoidable, and yet not always with ease. We think and talk by nature, and whoever thinks and talks can be led, unless he is careful, into wider questions. That may not be what makes everyone's life worth living, but it grows out of parts of our being we could not be without. The !Kung are not Socratic, so far as I know, but Socrates and the !Kung both are of our species.

In this book I ask about Socrates' quest. To ponder how to live, to reason about how to live, is in effect to ask what kind of life it is *rational* to live. I offer no special answer to this question; my first worry is what the question *is*. What does it mean to call an alternative rational, or another irrational? That is the puzzle of the book, and my hope is that from working on it, we can learn things worth learning about ourselves and about our questions.

In part the question how to live is moral—perhaps in whole. What kind of morality, if any, would be worth heeding? Or does the rational life do without morality? Here again my puzzle will be what the questions are. Again I want to know what 'rational' means, and in addition I want to understand this talk of morality. What are moral questions;

2. Euthyphro, as Plato presents him, is also a self-righteous fool, bent on a shocking course of action (prosecuting his father). Still, it is not this abnormal foolishness that leaves him vulnerable to Socrates's techniques. Euthyphro is abnormally foolish but normally prone to be tripped in his reasoning.

what do they mean? What, if anything, do they have to do with rationality?

The tie of morals to reason supports the whole of moral theory—perhaps. In any case, moral theories abound that say what the tie is or what it is not. They tell us whose good, if anyone's, reason commands we promote, but different theories tell us different things. Hume said that "Reason is, and ought only to be the slave of the passions, and can never pretend to any other office than to serve and obey them" (1740, bk. 2, part 3, sec. 3). Kant said the commands of morality are categorical demands of reason, and "reason issues inexorable commands without promising anything to the inclinations" (1785, 405). Sidgwick said the good is happiness, and that "as a rational being I am bound to aim at good generally,—so far as it is attainable by my efforts," and not at one's own happiness in particular (1907, 383). He thought also, though, that an egoist could evade this claim, and that in a "recognized conflict between duty and self-interest, practical reason" would be "divided against itself" (497, 508). Recent writers are divided against each other. Philippa Foot writes, "if justice is not a good to the just man, moralists who recommend it as a virtue are perpetrating a fraud" (1959, 100). David Gauthier maintains that reason demands maximally satisfying one's own desires, but under constraints that would be agreed to in certain ideal conditions (1986, esp. chap. 6). Thomas Nagel argues for pure impartiality between oneself and others. "In any situation in which there is reason for one person to promote some end, we must be able to discover an end which there is reason for anyone to promote, should he be in a position to do so."[3] Thinking about reason leads people to claims that are sharply at odds. It would be good to see what might be at issue.

My puzzle, then, is about Socrates, but it is also about the !Kung. I ask about moral philosophy, but also about everyday, non-philosophic life and thought and talk. The two kinds of talk are not the same, but one grows out of the other. As part of a human way of living, we think and discuss what it makes sense to do, and how it makes sense to feel about things. This thought and talk nudge us toward refinement. Wise choices and apt feelings figure in talk at both extremes: in refined, self-

3. Nagel (1970, 90). The list could be carried much further. See, for example, Grice (1967) for many claims about reasons. Parfit (1984, 313) argues at length against the claim that a rational person is "*equally* concerned about *all* the parts of his future."

conscious philosophizing and in everyday banter and quarrel. I want to know what is at stake in normative talk of both kinds.

A Glance Ahead

My proposal goes like this: Start with morality. We can understand the term broadly or narrowly. Broadly the moral question is how to live. Narrowly, we might try saying, morality concerns moral sentiments: the sentiments of guilt and resentment and their variants. Moral wrongs are acts to be avoided on pain of these sentiments. Morality in this narrow sense is a narrow part of life, but still, perhaps, something we need as a set of constraints. It seems worthwhile examining these constraints and asking whether they should matter.

To feel guilt or resentment is not in itself to make a moral judgment. A person can feel guilty and yet think he has done no wrong. He then thinks it makes no sense to feel the way he does, that his feelings are irrational. Narrowly moral judgments are not feelings but judgments of what moral feelings it is rational to have. Feelings, we think, can be apt or not, and moral judgments are judgments of when guilt and resentment are apt.

Saying this requires that we understand rationality. We must talk not only of wise choices, but of apt feelings too. What would a theory of rationality tell us? What does it mean to call something rational or irrational, a choice wise or foolish, a feeling apt or off the mark? My question shifts, then, from morality specifically to rationality in general.

To call a thing rational is to endorse it in some way. That suggests a scheme for getting at the meaning of the term. Instead of trying to define a property "rationality" by giving conditions under which a thing would have that property or lack it, start with the use of the term. Fix on the dictum "To call a thing rational is to endorse it," and search for a sense of 'endorse' for which the dictum holds true.[4]

The word 'rational' has a learned flavor, but the notion I have in mind is familiar. It is the one we use when we talk about "what it makes sense" to do or to believe, or when we speak of "the wise choice" in a situation. It is the one we use when we ask what we

4. My single and double quotation marks need a quick word here. Single quotes I try to use in the strict logician's way, to construct a name for the expression the marks surround. In the sentence this footnote tags, I name the word 'endorse'. Double quotes I use in the many looser ways quotation marks can be used, often to mention a word and use it in the same breath.

"ought" to do, or search for the "best thing to do", in a way that does not already presuppose we are talking morality. With feelings, it is the notion we use when we talk of anger, say, as warranted, or of pity as apt or misdirected. There does seem to be a common thought involved in all these turns of phrase, even if shades of meaning differ; one test is that to apply one phrase to an action or feeling while denying another is to invite puzzlement. I shall be using the learned term 'rational' in this broad way. It carries a kind of direct and flavorless endorsement, taken from the point of view of the person whose acts or feelings are being assessed. The rational act is what it makes sense to do, the right choice on the occasion. A rational feeling is an apt feeling, a warranted feeling, a way it makes sense to feel about something.[5] The term 'rational' may carry narrower suggestions, but this broad, endorsing reading is the one I need.

What is it, then, for an act or a way of feeling to be rational? In what way does a person who calls something rational endorse it? Cryptically put, my answer is this: to call something rational is to express one's acceptance of norms that permit it. This formula applies to almost anything that can be appraised as rational or irrational—persons aside. It applies to the rationality of actions, and it applies to the rationality of beliefs and feelings. We assess a wide range of things as rational or irrational, and it is puzzling how this can be. The analysis offers an answer.

"To call something rational is to express one's acceptance of norms that permit it." This is only a first sketch, and it raises many questions. What is it to accept norms? Acceptance is a state of mind, and a good way to identify this state of mind might be to exhibit its place in a rough psychological theory. The capacity to accept norms I portray as a human biological adaptation; accepting norms figures in a peculiarly human system of motivation and control that depends on language. Norms make for human ways of living, and we can understand our normative life as part of the natural world.

Normative talk is part of nature, but it does not describe nature. In particular, a person who calls something rational or irrational is not

5. The phrase 'it makes sense to' strikes my ear as carrying the right kind of endorsement, but there may be dialectical differences. If I say "It makes sense to read the footnotes later," some of my informants hear only a weak and grudging admission that postponing the footnotes would not be utterly crazy. For them I can say "It makes *most* sense to read the footnotes later," and that does carry the meaning I want. Alas it also carries an unwanted claim of uniqueness, that it makes more sense to read the footnotes later than now.

describing his own state of mind; he is expressing it. To call something rational is not to attribute some particular property to that thing—not even the property of being permitted by accepted norms. The analysis is not directly of what it is for something to *be* rational, but of what it is for someone to *judge* that something is rational. We explain the term by saying what state of mind it expresses. In this sense the analysis is *expressivistic*, and in too big a mouthful, I shall call it the *norm-expressivistic analysis*.[6]

The analysis is non-cognitivistic in the narrow sense that, according to it, to call a thing rational is not to state a matter of fact, either truly or falsely.[7] None of this leaves normative language defective or second-rate. The analysis explains why we need normative language, and as it takes shape, it ascribes to rationality many of the features on which theories of normative fact insist. In many ways, normative judgments mimic factual judgments, and indeed factual judgments themselves rest on norms—norms for belief. Normative discussion is much like factual discussion, I shall be claiming, and just as indispensable.

Part of my concern, then, is to understand morality, but my total concern is much wider. Morality narrowly glossed is a part of broadly normative life in general. Diverse aspects of life are governed by norms: action, and also belief and feeling. A good account of humanity will include a story of norms, a story of psychic mechanisms that make for their acceptance. The story will show these mechanisms in interaction and exhibit their adaptive rationale. We need, however, more than a theoretical account of ourselves as a part of nature, an account as if from afar. As we lead our normative lives we need a sense of what we are doing, a picture of ourselves that can guide us. The ways we see norms should cohere with our best naturalistic accounts of normative life, and it is here that an expressivistic analysis becomes useful. We experience our lives in normative terms, in terms of things it makes sense to do, to think, and to feel. The analysis joins this experience to the detached, scientific perspective. It tells what we can see ourselves as doing as we engage in normative inquiry and discussion.

6. Ayer (1936, chap. 6) and Hare (1981) are both expressivists in this sense. Ayer thinks that moral utterances express feelings or moral sentiments. Hare thinks they express preferences of a special kind: preferences, all told, that are universal, in that they do not depend on who occupies which position in the situation to which they pertain (1981, 107).

7. Non-cognitivistic treatments of moral language emerged as a distinct alternative with Barnes (1933), Ayer (1936, chap. 6), and Stevenson (1937). Later Hare's version stood out (1952, 1963, 1981).

The analysis is non-substantive, in that one could accept it and yet have no idea how to go about inquiring into what sorts of things are rational. In some ways that is a virtue, for the analysis is not itself meant to capture a substantive view on the nature of rationality. It is meant, rather, to capture the common element in dispute when people disagree about the nature of rationality. To address substantive questions, we should know what the questions are.

The analysis has worth, though, chiefly if it does help us with normative inquiry. It should help us say what sorts of things really are rational or irrational, right or wrong. In this book, however, I take up substantive normative questions only cursorily, and only at a high level of abstraction. I do have substantive hopes for the analysis, but they are indirect. The analysis will not yield a mechanical procedure for answering normative questions; I doubt even that it will lead us to think about good lives in ways that are strikingly novel. It should help us make sense of ways we already inquire, when those ways are reasonable. It should help us sort ways of inquiry that make sense from ways that do not. My hope is that if we proceed with a clear, plausible view of what we are doing, then we can progress—not automatically, but with less murk than otherwise.

Above all, I hope, the analysis will help us understand why it matters which acts and feelings are rational. Deciding what sorts of things are rational is deciding what norms to accept in various realms. The point of the book is to ask what this involves. We can picture human normative life as a part of nature, but that might leave us at a loss as to how to engage in that normative life. My first step toward engagement is to try to puzzle through what normative questions are. The eventual goal is to address normative questions with a better sense of what addressing them consists in.

Substantive Analyses

According to any expressivistic analysis, to call something rational is not, in the strict sense, to attribute a property to it. It is to do something else: to express a state of mind. It is, I am proposing, to express one's acceptance of norms that permit the thing in question. There is a special normative element in talk of what it makes sense to do, to think, and to feel, and that element resides in a special state of mind.

This may seem perverse. Surely a descriptivistic analysis would be better; that is to say, if a person calls something rational, it would be best to hear him as describing it, as ascribing a property to it. Then we

could assess what he has to say and have clear grounds for judging it true or false—whereas on an expressivistic analysis, we can only react.

My broad response is that any such descriptivistic analysis leaves a puzzle. It misses the chief point of calling something 'rational': the endorsement the term connotes. More specifically, we find that even the best descriptivistic analyses fail. They yield meanings that are inadequate to the basic purposes to which the term 'rational' can be put. They leave the puzzle of what it could mean to call something rational—a puzzle we can solve if we bring in endorsement.

This is not something I can prove conclusively. New analyses may appear, and I have no general proof that they all must fail. The most promising current contenders do fail, though, and on a plausible diagnosis of their failure, others will fail as well. On that diagnosis, what descriptivistic analyses miss is a general element of endorsement—an element an expressivistic analysis can capture.

Some analyses are substantive: They start with a theory of what rationality is, and then say that that itself is what the term 'rational' means. They take the characteristics that go to make something rational, and have them constitute the meaning of the term. For action, the most sophisticated theory of rationality might be called the *Hume-Ramsey theory*. It is now orthodox among economists and decision theorists, and we might look to it for a substantive analysis. The Humean element is that instrumental rationality is all of rationality. The content of one's ultimate ends cannot be assessed as rational or irrational. Rationality lies in adopting appropriate means, in an uncertain world, to whatever substantive goals one may have. *Instrumental rationality* here means rationality in the pursuit of ultimate ends which are accepted as given. Ramsey's contribution is to interpret instrumental rationality as a kind of formal coherence among preferences and actions. The conditions of coherence are such things as that one's preferences form an ordering, and that one always does what one most prefers. On the Hume-Ramsey view, in short, rationality demands no more than a formal coherence of preferences, in a way that can be expressed by a set of axioms.[8]

8. Hume (1739, bk. 2, part 3, sec. 3) thinks that reason, strictly understood, yields only beliefs, chiefly about cause and effect. An irrational action, then, could only be an action founded on unreasonable belief—or, as Hume adds in his discussion of "unreasonable passions", one that chooses a means insufficient for its end. (See Sturgeon 1985a, 27–28.) The more adequate, axiomatic view of rationality in the pursuit of goals was developed by Ramsey (1931). Ramsey's theory at first got little notice, but Savage (1954, 1972) elaborated a similar theory, and was highly influential.

Now my purpose here is not to assess the Hume-Ramsey theory as a substantive account of what makes an act rational. My question is rather, could Ramsey's axioms, or some similar set of axioms, be taken as giving the very meaning of the term 'rational'? Could they constitute, in the strict sense, a definition or analysis of the term? The answer, I claim, must be no, and for reasons of the very kind that Moore put forth in his attacks on various "naturalistic" definitions of 'good'.[9]

Take first the Humean component: that instrumental rationality is all of rationality, that one's intrinsic ends, wanting the things one wants for their own sake apart from what they bring, cannot be substantively irrational. According to Hume, reason must be the slave of whatever intrinsic preferences one may have. In his oft quoted words, " 'Tis not contrary to reason to prefer the destruction of the whole world to the scratching of my finger. 'Tis not contrary to reason for me to chuse my total ruin, to prevent the least uneasiness to an *Indian* or person wholly unknown to me. 'Tis as little contrary to reason to prefer my own acknowledged lesser good to my greater, and have a more ardent affection for the former than the latter" (1739, bk. 2, part 3, sec. 3). It is no less rational to seek pain than pleasure, it is no less rational to seek destruction than to seek cooperation, and it is no less rational to devote one's life to the collection of bottle caps than to the composition of symphonies—if one's ultimate preferences run that way. Now these are matters of controversy, to say the least. If, however, the Humean thesis is built into the very meaning of the word 'rational', then all these controversies are settled by the meaning of a word. What can the controversies be about?

Perhaps they really are about words and usage. People seem to be claiming something more substantive, though, when they quarrel with Hume. They seem to be making claims about how to live. Hume's own point may well be about words: he himself does not seem to be talking about which things are ultimately worth seeking; he is debunking the question. Still, his opponents do try to make claims of substance. They claim that there are things—accomplishment, for instance—that it is rational to want whether or not one does happen to want them.[10] They

9. Moore (1903, 10–21, esp. 10–12). Moore seems to have thought he had a blanket argument that all such definitions stemmed from a "naturalistic fallacy". It is hard to see quite what the argument is supposed to be; see, for example, Frankena (1939). What I imitate is his pattern of attacks on specific examples of naturalistic analyses.

10. Griffin speaks repeatedly of accomplishment as a chief prudential value. He lists

claim there are other things—vengeance, say—that it is irrational to want, for their own sake at least, even if one does in fact want them. These people use the word 'rational' in a way Hume finds unintelligible. It would be good to see if we can interpret what they could mean.

Any difficult analysis will be controversial, and so the point is not just that there are controversies. An analysis can be questionable and still be right. To refute an analysis by counterexample, we need a case in which a person not only doubts it, but accepts something inconsistent with it—without linguistic or logical confusion. Here is such a case: Octavia thinks reason demands that anyone give weight to his own future happiness. It makes this demand, she thinks, even on a person who is now indifferent to the future. Now whether or not she is right, if her thought is intelligible, if it is unconfused linguistically and logically, then the Humean thesis is wrong as a claim about meaning. Future happiness is no goal of Cassius's, imagine, and so by the Humean thesis, it is rational for him to give it no weight. According to Octavia, he rationally must give it weight. If the Humean thesis is right as a claim about meaning, then Octavia has meanings wrong or else she is logically confused. If her opinion is intelligible, then the Humean thesis is wrong as a claim about meaning.

No such counterexample will be airtight. It can always be asked whether the opinion put forth as intelligible really is intelligible. Is the thinker really free of all linguistic and logical confusion? The point, though, is this: An opinion seems intelligible; we seem to understand what the person is claiming. There is pressure, then, to interpret the opinion as intelligible—or at least to explain the appearance of intelligibility.

In short, then, the Humean thesis fares poorly as a claim about meaning. People think some things worth wanting whatever in fact one wants, and other things not worth wanting even if in fact one does want them. We seem to understand these claims, and it would be good to have an analysis that counts them as intelligible.

Instrumental Rationality

Perhaps the Humean's point is something like this: Instrumental rationality seems unproblematic. We quarrel about facts and we quarrel

as other prudential values understanding, enjoyment, deep personal relations, and such components of human existence as autonomy, basic capabilities, and liberty (1986, 67–68). His account of what makes something a prudential value is partly a full-information account. In addition, prudential values must be widely shared among people (113–120).

about ultimate goals. In disputes about ultimate goals, though, is talk of rationality anything more than a rhetorical bludgeon? Why not disarm, and confine the term 'rational' to talk of means? Rationality then is rationality of beliefs: beliefs about what means will reach what ends. So Hume seems to have thought, and so perhaps ought we.[11]

Instrumental rationality, though, is more than rational means-end beliefs. Any account of instrumental rationality must cope with interactions of reasoning agents one with another. It must cope with uncertainty. We need something like Ramsey's theory to capture these complexities.

Ramsey treats rationality as a kind of formal coherence. His axioms require more than sheer logical consistency in belief. They require that one's preferences form an ordering. They require as well a "sure thing principle" that goes like this: Let *A*, *B*, and *C* be prospects, and let *p* be something that might happen—say, a coin's landing heads—about which, in itself, one cares not at all. Suppose one prefers *A* to *B:* then one prefers the compound prospect

A if *p* and otherwise *C*

to the compound prospect

B if *p* and otherwise *C*.[12]

Ramsey's account is widely accepted, but it is controversial—even for instrumental rationality. Proponents claim that an action is rational when and only when its agent satisfies Ramsey's formal conditions of coherence for preference. That incorporates Hume's thesis, but even for instrumental rationality, the claim is hotly disputed. Now if Ramsey's axioms give what 'rational' means, it is puzzling what the controversies could be. They appear substantive; they seem to be over whether a person can be rational without satisfying the axioms. They do not seem mere disputes over English usage, as they would be if Ramsey's axioms defined the very meaning of the term 'rational'.

Here, at length, are two current examples of controversy.

You and your twin are arrested and placed in a prisoners' dilemma. The district attorney has isolated you from each other, and now con-

11. Don Loeb proposed this line of thought to me, and convinced me that it, or something like it, explains some of the attraction of Humean accounts of practical rationality.

12. I speak of Ramsey's axioms loosely. The points I want to make are not about Ramsey's specific formulation, but about the general form of theories like those of Ramsey (1931) and Savage (1954, 1972).

fronts you each with this proposition: You get a year off your sentence if you implicate twin in a robbery. Specifically, if twin implicates you, you get ten years in jail unless you have implicated him, in which case you get only nine. If twin keeps mum, then you get one year, on a trumped-up charge, unless you yourself have implicated him, in which case you go free. Twin is faced with the same proposition, and you both know it.

	Twin rats	*Twin keeps mum*
You rat	9 years for you 9 years for twin	Freedom for you 10 years for twin
You keep mum	10 years for you Freedom for twin	1 year for you 1 year for twin

You want to minimize your own time in jail, and you care not at all about how long twin spends there. Twin and you are very much alike, and so if you rat on him, that is bad news for you: it indicates that probably twin is ratting on you, making you spend nine years in jail. Likewise if you keep mum that is good news for you: it indicates that probably twin is keeping mum, giving you just one year in jail.[13]

How is it instrumentally rational for you to pursue your goal—your goal of spending as little time in jail yourself as possible? Some say it makes sense to rat. What twin does is now beyond your influence, and whatever he does, you save yourself a year in jail if you rat. This is the standard view of game theorists, yielded by systems like Ramsey's: ratting "dominates" keeping mum.[14]

Others say it makes sense to keep mum. That indicates twin is keeping mum, and twin's keeping mum will save you nine years in jail.[15] Others even say it would make sense to keep mum if the other prisoner were not your twin—at least if you each can expect the other to keep

13. The prisoner's dilemma is attributed to A. W. Tucker; see Luce and Raiffa (1957, 95–97). Nozick (1969) introduces twins, and points out that doing so makes the problem into an analog of Newcomb's paradox, the topic of his paper.

14. Gibbard and Harper (1978) and Lewis (1979b) are among the people who accept this argument. There is some dispute over the interpretation of dominance in systems like Ramsey's (see Levi 1975), but I take the interpretation favored by Nozick (1969) and Gibbard and Harper (1978).

15. This argument would be recognized as sound in the decision theory developed by Jeffrey (1965); see Gibbard and Harper (1978). Levi (1975) and Eells (1982) defend Jeffrey's kind of decision theory from attacks based on examples like this one, but they may want to deny that it has the consequences I am claiming for it.

mum. It cannot make sense, they say, for both of you to act in a way that jointly frustrates the goals of each.[16]

Now my puzzle is not which of these sides is right, but what is at issue. Those who think it rational to keep mum may know perfectly well that keeping mum violates axioms like Ramsey's; they think the axioms mischaracterize rationality. 'Rational' as they use the term cannot mean "satisfying the axioms", or these people would be trivially wrong. Do they mean something different by the term from their opponents? Then there is no real disagreement, just a word used in different senses. There seems, though, to be real substance to the dispute: hypothetically taking on egoistic goals, the two sides disagree about what to do.

Here is a second controversy, active among economists.

You are forced to play Russian roulette—but you can buy your way out. One bullet is placed in a six chamber revolver, and you must spin the cylinder vigorously and shoot yourself square in the head. Here is a question: What is the most you would pay to have the bullet removed? (Pay in installments over a lifetime if you survive. The payment is excused if you die.)

Next the scene shifts to a second version of the problem. You are forced to play Russian roulette with four bullets in the revolver. Answer a new question: What is the most you would pay in this version to have *one* of the four bullets removed, leaving three? More? Or less?[17]

Most people answer less. Standard decision theory says to answer more. Indeed it tells you to pay as much to reduce the bullets from four to three as you would pay to reduce the bullets from two to none.

The second scene, after all, is as if you were forced to play in two stages as follows: First you must play with three bullets. Then, if you survive, you have to play again, this time with two bullets—but you

16. See Baier (1958, 308–315) for reasoning that appears to support this conclusion, and Rapoport (1960, 174–178) for such a treatment of the prisoner's dilemma itself. Parfit attacks this conclusion (1984, 87–92). Gauthier (1986, chap. 6) takes a position something like this for certain circumstances. The circumstances are these: that the prisoners agree beforehand to keep mum, that each is disposed to keep the agreement so long as he expects the other to, and that each can tell fairly well that the other is so disposed. Then, Gauthier maintains, each acts rationally in keeping the agreement. This in effect is a claim about instrumental rationality, and others disagree.

17. Kahneman and Tversky (1979, 283) give this example and credit Richard Zeckhauser. The structure is that of the Allais paradox.

can pay to have them both removed. Work it out: All told, in both versions, if you don't pay, your chances of killing yourself are four in six. If you do pay, your chances of killing yourself are three in six—again in both versions. The two versions, then, are equivalent.[18]

How much are you willing to pay? In the two-round version if you die in the first round, it makes no difference what you are to pay on survival. Your question, then, is how to treat the second round: how much at most to pay to have both bullets removed, when they are the only ones in the cylinder. You would pay more, presumably, to have two sole bullets removed than to have one sole bullet removed. Therefore in the two one-round scenes, you should pay more to reduce the bullets from four to three than to reduce them from one to zero. So goes the argument.

Some people accept it. Others understand it but reject it. They stick with their initial claim that it is worth more to rid oneself of a one-in-six threat to life and get safety, than merely to move one's chances of safety up from two-in-six to three-in-six. The argument depends, they say, on a dubious principle. Take the stage with two bullets in the cylinder. How much do you pay to have them both removed? The argument assumes that this should not depend on what risks lead up to that stage. It should not depend on whether the choice is faced all by itself or whether another stage leads up to it. If that yields counter-intuitive results, so much the worse for the principle.[19]

What is at issue between those who accept the principle and those who reject it? Those who reject it cannot mean by 'rational' what axioms like Ramsey's say. They know the principle is one of the

18. David Lewis has argued along these lines.

19. The appeal here is to the sure-thing principle. It says that if two lotteries are equally undesirable, then when one is substituted for the other in a compound lottery, that makes no difference to how undesirable the compound is. The principle looks appealing, but the Russian roulette case shows it wrong, its detractors say.

A little more carefully, the argument is this. Take the most you would be willing to pay to have two bullets removed if they are the only bullets. Call living having paid that much being *poor*, and living having paid nothing being *rich*. By stipulation, you are indifferent between *poor* and what I shall call the *plunge:* namely, four chances in six of being *rich* and two of being *dead*. Now take the two-round game. This is a lottery with even chances of two prizes: *dead* and the *plunge*. The buyout substitutes *poor* for the *plunge* in the compound; the result is even chances *dead* and *poor*. By the sure-thing principle, that should make no difference: you should pay as much to have one of four bullets removed as to have two bullets removed if they are the only ones. But clearly, opponents argue, it is not worth as much to have one of four bullets removed as to have a bullet removed when it is the only one—and so much the worse for the sure-thing principle.

axioms[20] and they think rational choice violates it. Do its opponents and its proponents simply mean different things by 'rational'? Then there is no real dispute. There seems in fact to be a dispute of substance: The two sides differ on whether to pay more in the one case or in the other.

It is easy to think that instrumental rationality, at least, is unproblematical. The problems seem to lie with the substantive rationality of goals. Instrumental rationality seems just a matter of beliefs about consequences, and of doing the thing whose consequences one most wants. The two examples—prisoner's dilemma with twins and Russian roulette with buyouts—each show that things are not so straightforward. In both cases the facts are clear enough; they are given by stipulation and known to the agents. The puzzles of instrumental rationality lie not just in the facts. Agents seek their goals in interaction with other agents who have different goals. Agents pursue their goals knowing they are ignorant. Instrumental rationality includes coping with both these features of life, and there are disputes about how it makes sense to do so. In a broad sense, Ramsey's decision theory treats rationality as a kind of coherence, but this coherence is not just logic narrowly construed, and it is not just straightforwardly choosing what one most wants.

We need, then, something as complex as Ramsey's axioms to characterize instrumental rationality. Such characterizations are controversial. Choosing sides in these controversies is not a simple matter of saying which axioms are satisfied and which are violated. People who agree on that can disagree on which acts are instrumentally rational. Choosing sides consists, it seems, in a kind of hypothetical endorsement: we hypothetically adopt certain ultimate ends, and then settle what to do.

The Ramsey axioms may, of course, be offered as something other than a definition of 'rational' in the sense I am after. In that case, nothing I am saying speaks for or against them. The axioms may be offered as a substantive account of the nature of rationality. My puzzle then becomes, What is the account claiming for actions that satisfy those axioms? The axioms may, on the other hand, be used to stipulate a

20. Again, the reference is not specifically to Ramsey's formulation, but to systems like it. The principle is one of the axioms or it follows from the axioms, depending on how the particular set of axioms is formulated. The name "sure-thing principle" comes from Savage (1954) but his formulation is somewhat different.

meaning for the word 'rational', as mathematicians stipulate what shall constitute a "group" or a "well-ordering". In that case, there is no great room for controversy: nothing is at stake in a mathematician's appropriation of a term but suggestiveness, clarity, and fecundity. What I am asking is what the term 'rational' means in genuine controversies, if such there be, on the nature of rationality.

Full Information Analyses

On a view widespread among philosophers, rationality is somehow a matter of full awareness of the facts. Brandt elaborates a view of this kind. He announces that he will "pre-empt the term rational to refer to actions, desires, or moral systems which survive maximal criticism and correction by facts and logic" (1979, 10). Brandt's account could be put as follows. First, we characterize the intrinsic desires that it would be rational for a person to have. They are the ones the person would have after repeated representation of all relevant, scientifically available information, in an ideally vivid way, at appropriate times (11, 113). Brandt calls this vivid, repeated representation of facts *cognitive psychotherapy*. Second, we define the rationality of an act as a means, or its "instrumental rationality". Take some system of intrinsic desires as given. An act is *instrumentally rational* as a means to the fulfillment of those desires if and only if the following holds: if a person had those intrinsic desires, and he had all relevant scientifically available information "present to awareness, vividly, at the focus of attention, or with an equal share of attention," then he would be willing to perform that act (11). An act, then, is *fully rational* if and only if it is instrumentally rational as a means to the fulfillment of the intrinsic desires it would be rational for the agent to have (149).

One problem for any "full awareness" account such as Brandt's is that rationality, in the ordinary sense, often consists not of using full information, but of making best use of limited information. Acting in full awareness of all relevant facts suggests not rationality, but something more like "advisability". Whereas rationality is a matter of making use of the information one has, advice can draw on information the advisee lacks. Suppose, for instance, I am lost in the woods without a map or compass. With full information, I would take the shortest, easiest route out of the woods and to where I want to be—but I don't have full information. The rational thing for me to do, then, is to pursue one of the standard strategies for getting out of trackless woods:

walk carefully in a straight line by sighting along trees, or go consistently downhill. Some such strategy is rational, even though I am confident it will not take me along the route I would travel with full information. What is rational, then, is not what an agent would do with full information. It is not even what he would do with all the information he should have obtained: my being lost in the woods without a map or compass is likely the result of previous irrationality, but rationality now consists in coping without the information I ought earlier to have secured. If I unexpectedly meet someone who knows the woods, then what I shall want is advice. I shall not want to know what it is rational for me to do; I already know that the rational thing is to ask for directions. The expert can then tell me what is advisable: he can tell me the shortest way out of the woods. In his advice, he draws on full information. Of course, once I have his advice, it then becomes rational for me to do what he says—but only then. Rationality, we may conclude, is related to advisability, but the connection is this: in the special case in which I know all that bears on my choice, what it is rational for me to do is what it is advisable for me to do. Otherwise the two may differ, and full information pertains to advisability.[21]

Perhaps rationality in action is best construed, in the spirit of Brandt's proposal, as acting in full and vivid awareness of whatever information one has. Whether or not that is the best way to fix things up, the definition as he gives it seems best taken as an analysis of advisability, or rationality in light of full information. How does it fare if so taken? Brandt himself offers the proposal not as an analysis of our actual usage but as a reform. Nevertheless, we may ask whether the proposed reform misses anything worth preserving in our unreformed use of the term.

Now we ordinarily suppose that there is more to rationality than Brandt's proposal captures—as his own examples show. Whether an intrinsic desire is rational, on Brandt's account, is a matter of whether it would be extinguished in cognitive psychotherapy. In consequence, Brandt points out, any intrinsic desire that is so deeply ingrained in early childhood as to be inextinguishable by cognitive psychotherapy will count as rational on his definition (113). Suppose, for instance, a

21. Harman makes this point (1982, 127). Brandt does speak of "subjective" and "objective" rationality (1979, 72–73); my question is whether "objective rationality" is rationality at all. Brandt's "objective rationality" differs somewhat from what I am calling "advisability": it is rationality not in light of all the facts, known and unknown, but in light of everything currently known by experts.

person neurotically wants to keep his hands as germ-free as possible, and so washes his hands a number of times each hour. The desire, to be sure, might extinguish with vivid and repeated awareness of certain facts: that keeping his hands perfectly germ-free will not promote health, that incessant washing may indeed threaten his health, that the hand-washing interferes with all the other activities that he finds of value, and that if he did not care about germs so much, he would find nothing missing in his life on that account. Suppose, though, he has confronted all these facts, repeatedly and vividly, and he nevertheless says "I realize all that. But I just don't want those creepy-crawly things on my hands—and least of all do I want to be a person who would be willing to tolerate them on his hands." Then the person is rational in his preferences, on Brandt's account.

Brandt, I take it, does not regard this conclusion as a virtue of his account. It is an unsought consequence, which Brandt accepts only because, in his opinion, any account that yields the opposite answer will have defects far worse. What he claims, then, is not that his account does everything we should want, but that no alternative will do better on the whole. What I have said is thus no refutation of Brandt: it simply shows a respect in which we might hope to do better.

What we should note, though, is that the funny cases—the cases where Brandt's account labels crazy acts rational—have a systematic import. The word 'rational', in the sense we are after, has an automatically recommending force. Brandt indeed spends a whole chapter trying to show that his account preserves this recommending force (chap. 8). Rationality in Brandt's sense, however, need not be a recommendation. Suppose, for instance, I think that full and vivid realization of all the relevant facts would evoke a debilitating neurosis—a neurosis I have kept from controlling my life only by avoiding vivid confrontation with certain facts. Perhaps with a more vivid realization of what peoples' innards are like, I would want to stay away from dinner parties and do all my eating alone—although then I would feel lonely and isolated. Suppose on that ground I accept the claim "If you were fully rational in Brandt's sense, you would not ever eat with other people." Would that have recommending force?

A civil servant who firmly rejects all offers of bribes might fear that if he dwelt vividly on all that he is forgoing, he would yield to temptation. That, roughly, is to fear that his determination not to take bribes is irrational in Brandt's sense. Cognitive psychotherapy, to be sure, would involve vivid awareness of the social consequences of bribery

and its personal dangers. If the personal danger is minimal, though, the civil servant may well suspect that vivid realization of the social consequences of bribery would little avail against vivid realization of the pleasures accepting bribes would open to him. Moreover, once temptations had done their work, he might well be glad they had. Suppose all this is so. Is he being irrational when out of moral conviction he avoids contemplating the temptations of his position? Is it irrational for him to refuse a bribe? On Brandt's account, his moral compunctions against losing his moral compunctions do not count if they would not survive cognitive psychotherapy. Does that recommend dwelling on the temptations? Does that recommend bribery? The honest civil servant might well think not.

The same kind of story might be told for an egoist. Suppose an egoist achieves happiness by keeping his mind off the joys and sorrows of others, but he thinks that if he fully realized what suffering he could alleviate and what joy he could spread by a life of self-sacrifice in the service of others, he would sadly forsake his life of self-centered enjoyment, and take on an irksome burden of service to humanity. Must he find that a recommendation for self-sacrifice? Why is it not instead a reason to shield himself from the facts? Can he not say without linguistic error, "It's crazy to dwell on the effects one can have on the lives of others. For if you do, then next thing you know, you will be making all sorts of sacrifices for others at the expense of your own enjoyment. Why deliberately take a path that leads to sorrow?" Now perhaps human psychology is not as I have described it, and a life of service to others is more personally rewarding than a life of ignorant, selfish enjoyments. That, however, is not the situation this egoist thinks he faces. He is convinced that the life he would want to lead if he fully realized what was at stake for others is painful in comparison to the life he will lead with his egoism fostered by ignorance. If, thinking that, he is also convinced that the altruistic life is the one he would choose with full knowledge, why should that commend the altruistic life to him?

These examples have a common structure. Often we suppose that we are reliable formers of rational desires, if only we get the facts straight and have them clearly and vividly in view. The examples I have given are cases in which a person thinks himself an unreliable transformer of vivid realizations into rational desires, and so avoids dwelling on the facts he thinks would lead him astray. On a full-information account like Brandt's, this talk of reliability has no

substance: that we are reliable transformers of vivid realizations into rational desires is analytic—true just in virtue of what 'rational' means. Brandt himself does not claim this for any ordinary sense of 'rational', for he thinks in ordinary use the term has no clear meaning (6–7). In all the examples, though, there is a common element that an account like Brandt's misses: the protagonist endorses a system of ends he thinks would not survive a vivid, repeated confrontation with the facts. In that sense, he thinks his desires would no longer be rational if he underwent such a confrontation. In the argument that follows I pursue the element of endorsement that full-information accounts leave out. In that element, if I am right, lies the specially normative aspect of the term.

2 • Nature and Judgment

If we try to paint normative life as a part of nature, crucial parts keep looking off shape. Reasons in the picture look not quite like genuine reasons. Meanings are hardly recognizable. For reasons and for meanings both, naturalistic depiction seems wrong, or if somehow it can be got to work, artistry is needed. We have reason to do various things, or so we think, but if we try to brush that fact into a picture of nature, we risk a botched likeness; bald descriptions do not straight off yield reasons. The normative things we say have meaning, but how can we include these meanings, and others too, in a picture of nature? The job may not be impossible, but it will require invention.

The task is to put in one picture apparent facts of three kinds:

1. Naturalistic facts: us as a part of nature, our acts and thoughts and feelings as they might be understood in a natural science
2. Normative facts: what it is rational to do and to think and to feel, what it *makes sense* to do or think or feel
3. Facts of meaning: what our words mean in general, and in particular what our normative claims mean—what it means to say it is rational to do or think or feel such-and-such

In my own picture, all strict facts will be naturalistic. Facts of meaning will come out as genuine facts, and so as naturalistic. Apparent normative facts will come out, strictly, as no real facts at all; instead there will be facts of what we are doing when we make normative judgments. It does make sense to do some things and not others, but that will not be part of a systematic picture of nature. Our thinking about these things will.

In this chapter, I talk about the kind of naturalistic picture I shall be painting, and then I go on to talk about meanings. To do what I plan