The Moral First Aid Manual

DANIEL C. DENNETT

THE TANNER LECTURES ON HUMAN VALUES

Delivered at
The University of Michigan

November 7 and 8, 1986

DANIEL C. DENNETT was born in Boston in 1942 and received his B.A. from Harvard and his D. Phil. from Oxford. After six years at the University of California at Irvine, he moved to Tufts University in 1971, where he is now distinguished Arts and Sciences Professor, and Director of the Center for Cognitive Studies. In addition to his books on the philosophy of mind, *Content and Consciousness* (1969), *Brainstorms* (1978), and *The Intentional Stance* (1987), he has written on free will (*Elbow Room*, 1984) and co-edited, with Douglas Hofstadter, *The Mind's I: Fantasies and Reflections on Self and Soul* (1981).

I was very pleased to be invited to give the Tanner Lecture this year, not just because of the honor of being included in this most distinguished series, and not just because of my Ann Arbor friendships, but also because of the bracing opportunity it offered to indulge in what might be called "licensed poaching." According to the letter of invitation, "The purpose of the Tanner Lectures is to advance and reflect upon the scholarly and scientific learning relating to human values and valuation."

I decided to take this seriously, and thus have been lured a little further into ethics — human values and valuation — than I belong. I was not at all unwilling to be a poacher, since I have long harbored dissatisfactions and skepticisms about what I took to be being done in ethics, in particular about the wildly unrealistic idealizations being used, and this was to be an occasion for me to express them. But poaching is a dangerous business, and as I began doing my homework I discovered that the very themes I had hoped to hold forth on have already found expression in recent work in ethics, and in just about every case were developed with subtlety and precision that went beyond my amateur ruminations. I was an ill-read floater with the *Zeitgeist*, not the pioneer I had hoped to be.

With time running out, it became less and less clear to me that I had anything, beyond some selective applause, to offer to the

NOTE: In this lecture I revise and expand on material presented in the Kathryn Fraser McKay Lecture, St. Lawrence University, September 1986, the George Brantl Lecture, Montclair St. College, February 1986, a Distinguished Lecture to the MIT Laboratory for Computer Science, March 1986, published as "Information, Technology, and the Virtues of Ignorance," *Daedalus*, Summer 1986, pp. 135–53, and a Humanities Lecture at the University of Kansas, November 1986. In addition to the discussions prompted by those lectures, I am particularly indebted to my colleagues Stephen White, Norman Daniels, and Hugo Bedau for their discussions with me, and also to Gordon Brittan, Richmond Campbell, Bo Dahlbom, Robert French, Douglas Hofstadter, Charles Karelis, Onora O'Neill, and Connie Rosati for their comments and advice.

current groundswell of reaction against the misuse of theoretical idealizations in ethics. In the end, however, I concluded that the current salutary trend has not gone far enough in certain respects, and that — unless I am deceived by the felt need to say something "new" — I may provide something of a fresh perspective on the issues. As you will soon learn, there is a certain poetic justice in my confessing that what you are about to hear is the somewhat misshapen product of time-pressured problem-solving.

1. Mill's Nautical Metaphor

A hundred and twenty-five years ago, John Stuart Mill felt called upon to respond to an annoying challenge to his *utilitarianism*: ". . . defenders of utility often find themselves called upon to reply to such objections as this — that there is not time, previous to action, for calculating and weighing the effects of any line of conduct on the general happiness." His reaction was quite fierce:

> Men really ought to leave off talking a kind of nonsense on this subject, which they would neither talk nor listen to on other matters of practical concernment. Nobody argues that the art of navigation is not founded on astronomy because sailors cannot wait to calculate the Nautical Almanac. Being rational creatures, they go to sea with it ready calculated; and all rational creatures go out upon the sea of life with their minds made up on the common questions of right and wrong, as well as on many of the far more difficult questions of wise and foolish. And this, as long as foresight is a human quality, it is to be presumed they will continue to do.
>
> [*Utilitarianism*, 1861, p. 31]

This haughty retort has found favor with many — perhaps most — ethical theorists, but in fact it papers over a crack that has been gradually widening under an onslaught of critical attention. The naïve objector was under the curious misapprehension that a system of ethical thinking *was supposed to work* and noted

that Mill's system was highly impractical — at best. This is no objection, Mill insists; utilitarianism is supposed to be practical, but not *that* practical. Its true role is as a background justifier of the foreground habits of thought of real moral reasoners. This background role for ethical theory (and not only utilitarians have sought it) has proven, however, to be ill-defined and unstable. Just how practical is a system of ethical thinking *supposed* to be? What is an ethical theory for? Tacit differences of opinion about this issue, and even a measure of false consciousness among the protagonists, have added to the inconclusiveness of the subsequent debate.

For the most part philosophers have been content to ignore the practical problems of real-time decision-making, regarding the brute fact that we are all finite and forgetful, and have to rush to judgment, as a real but irrelevant element of friction in the machinery whose blueprint they are describing. It is as if there might be two disciplines — ethics proper, which undertakes the task of calculating the principles of what the ideal agent ought to do under all circumstances — and then the less interesting, "merely practical" discipline of *Moral First Aid*, or *What to Do Until the Doctor of Philosophy Arrives*, which tells, in rough-and-ready terms, how to make "on line" decisions under time pressure.

In practice, philosophers acknowledge, we overlook important considerations — considerations that we really shouldn't overlook — and we bias our thinking in a hundred idiosyncratic — and morally indefensible — ways; but *in principle*, what we ought to do is what the ideal theory (one ideal theory or another) says we ought to do. Philosophers have then concentrated, not unwisely, on spelling out what that ideal theory is. The theoretical fruits of deliberate oversimplification through idealization are not to be denied — in philosophy or in any scientific discipline; reality in all its messy particularity is too complicated to theorize about taken straight. The issue is rather (since every idealization is a strategic choice): which idealizations might really shed some light

on the nature of morality, and which will just land us with diverting fairy tales.

It is easy to forget just how impractical ethical theories actually are, but we can make the truth vivid by reflecting on what is implicit in Mill's use of a metaphor drawn from the technology of his own day. The *Nautical Almanac* is a book of tables, calculated and published annually, from which one can easily and swiftly derive the exact position in the skies of the sun, the moon, the planets, and the major stars for *each second* of the forthcoming year. The precision and certainty of this annual generator of expectations was, and still is, an inspiring instance of the powers of human foresight, properly disciplined by a scientific system *and directed upon a sufficiently orderly topic*. Armed with the fruits of such a system of thought, the rational sailor can indeed venture forth confident of his ability to make properly informed real-time decisions about navigation. The practical methods devised by the astronomers actually work.

Do the utilitarians have a similar product to offer to the general public? Mill seems at first to be saying so. Today we are inured to the inflated claims made on behalf of dozens of high-tech systems—of cost–benefit analysis, computer-based expert systems, etc. — and from today's perspective we might suppose Mill to be engaging in an inspired bit of advertising: suggesting that utilitarianism can provide the moral agent with a foolproof Decision-Making Aid. ("We have done the difficult calculations for you! All you need do is just fill in the blanks in the simple formulae provided.")

Jeremy Bentham, the founder of utilitarianism, certainly aspired to just such a "felicific calculus," complete with mnemonic jingles, like the systems of practical celestial navigation that every sea captain memorized.[1]

[1] From chapter IV of Bentham's *Introduction to the Principles of Morals and Legislation*, 1789:

> *Intense, long, certain, speedy, fruitful, pure* —
> Such marks in *pleasures* and in *pains* endure.
> Such pleasures seek if *private* be thy end:
> If it be *public*, wide let them *extend*.

This myth of practicality has been part of the rhetoric of utilitarianism from the beginning, but in Mill we see already the beginning of the retreat up the ivory tower to ideality, to what is calculable "in principle" but not in practice.

Mill's idea, for instance, was that the best of the homilies and rules of thumb of everyday morality — the formulae people *actually considered* in the hectic course of their deliberations — had received (or would receive in principle) official endorsement from the full, laborious, systematic utilitarian method. The faith placed in these formulae by the average rational agent, based as it was on many lifetimes of experience accumulated in cultural memory, could be justified ("in principle") by being formally derived from the theory. But no such derivation has ever been achieved.[2]

It will help us appreciate this obvious fact about consequentialist theories such as utilitarianism if we compare them, not to the productions of the Astronomer Royal, as Mill did, but to a more contemporary technique of expectation-generation: computer-aided weather forecasting.

The current North American data-gathering grid divides the atmosphere into cells approximately thirty miles on a side and ten thousand feet in height. This yields in the neighborhood of 100,000 cells, each characterized by less than a dozen intensities: temperature, barometric pressure, wind direction and velocity, etc. How these intensities change as a function of the intensities in the neighboring cells is fairly well understood, but computing these changes in temporal increments small enough to keep some significance in the answers challenges today's largest supercomputers. Obviously, a weather prediction must be both accurate and timely;

[2] It is arguable that Robert Axelrod's *The Evolution of Cooperation* (New York: Basic Books, 1984), achieves a derivation of the Tit For Tat rule: Cooperate at the outset, punish defections with a defection, but respond to further cooperation with cooperation. As Axelrod himself points out, however, the rule's provable virtues assume conditions that are only intermittently (and controversially) realized. In particular, the "shadow of the future" must be "sufficiently great," a condition about which reasonable people might disagree indefinitely, it seems.

achieving accuracy at the cost of taking thirty-six hours to calculate a twenty-four-hour prediction is no solution.

It is not clear yet whether reliable long-range weather forecasting is possible, since the weather may prove to be too chaotic to permit *any* feasible computation. The behavior of the weather is strikingly unlike the behavior of the heavenly bodies. Suppose though, for the sake of illustration, that there were a *proven* forecasting algorithm — one that could successfully "predict" tomorrow's weather if allowed to engage in a *month* of number crunching on a bank of supercomputers. This would be scientifically very interesting, but not very useful. We can imagine taking the tour of the weather bureau and being shown the gleaming giants at their work. "How do you actually *use* the algorithm in figuring out the forecasts you are obliged to issue every day?" we ask. "Oh, we don't use the algorithm at all. We sort of eyeball the maps and the local conditions and then apply our favorite maxims. Jones is partial to 'red sky at night, sailor's delight' while I am more into aching joints and looking for the groundhog's shadow. We vote, in the end, and our track record is pretty good."

That is the way it is with ethics too — only with ethics, things are worse. At least with meteorology, there is an uncontroversial and widely accepted ideal background theory — however infeasible it might be in practical calculations. Now we can see that there are actually three ways in which Mill's metaphor is misleading. First, as just mentioned, no ethical theory enjoys the near-universal acceptance of astronomy or meteorology, in spite of vigorous campaigns by the partisans. Second, there are no feasible algorithms or decision procedures for ethics as there are for celestial navigation. Third, the informal rules of thumb people actually use have never been actually derived from a background theory, but only guessed at, in an impressionistic derivation rather like that of our imagined meteorologists.

And unlike the weather, which *may* turn out not to be a chaotic and incalculable system (but may rather asymptote on some ball-park

trends), the ethically relevant effects of our contemplated actions are bound to be incalculable unless we place *arbitrary* limits on them.

Why? Because of what might be called the Three Mile Island Effect. Was the melt-down at Three Mile Island a good thing to have happened or a bad thing? If, in planning some course of action, one encountered the melt-down as a sequel of probability *p*, what should one assign to it as a weight? Is it a strongly negative or strongly positive effect? We can't yet say, and it is not clear that *any* particular long run would give us the answer.

Compare the problem facing us here with the problems confronting the designers of computer chess programs. One might suppose that the way to respond to the problem of real time pressure for ethical decision-making techniques is the way one responds to time pressure in chess: heuristic search-pruning techniques. But there is no checkmate in life, no point at which we get a definitive result, positive or negative, from which we can calculate, by retrograde analysis, the actual values of the alternatives that lay along the path taken. How deep should one look before settling on a weight for a position? In chess, what looks positive from ply 5 may look disastrous from ply 7. There are ways of tuning one's heuristic search procedures to minimize (but not definitively) the problem of misevaluating anticipated moves. Is the anticipated capture a strongly positive future to be aimed at, or the beginning of a brilliant sacrifice for your opponent? A *principle of quiescence* will help to resolve that issue: always look a few moves beyond any flurry of exchanges to see what the board looks like when it quiets down. But in real life, there is no counterpart principle that deserves reliance.

Three Mile Island has been followed by quite a long intervening period of consolidation and quiescence, but we *still* have no idea whether it is to be counted among the good things that have happened or the bad.

The suspicion that there is no stable and persuasive resolution to such impasses has long lain beneath the troubled surface of

criticism to consequentialism, which looks to many skeptics like a thinly veiled version of the classically vacuous stock market advice: buy low and sell high — a great idea in principle, but systematically useless as advice to follow.[3]

So not only have utilitarians never made an actual practice of determining their specific moral choices by calculating the expected utilities of (all) the alternatives (there not being time, as our original objector noted), they have never achieved stable "off-line" *derivations* of partial results — "landmarks and direction posts," as Mill puts it — to be exploited on the fly by those who must cope with "matters of practical concernment."

What, then, of the utilitarians' chief rivals, the various sorts of Kantians? Their rhetoric has likewise paid tribute to practicality — largely via their indictments of the impracticality of the utilitarians.[4]

What, though, do the Kantians put in the place of the unworkable consequentialist calculations? Kantian decision-making typi-

[3] Judith Jarvis Thomson has objected that neither "buy low and sell high" nor its consequentialist counterpart, "do more good than harm," is strictly vacuous; both presuppose something about ultimate goals, since the former would be bad advice to one who sought to lose money, and the latter would not appeal to the ultimate interests of all morally-minded folk. I agree. The latter competes, for instance, with the advice the Pirate King gives to Frederick, the self-styled "slave of duty," in *Pirates of Penzance*: "Aye me lad, always do your duty — and chance the consequences!" Neither slogan is *quite* vacuous.

[4] A Kantian who presses the charge of practical imponderability against utilitarianism with particular vigor and clarity is Onora O'Neill in "The Perplexities of Famine Relief," in *Matters of Life and Death*, Tom Regan, ed. (New York: Random House, 1980). She shows how two utilitarians, Garrett Hardin and Peter Singer, armed with the same information, arrive at opposite counsels: we should take drastic steps to prevent short-sighted efforts to feed famine victims (Hardin), or we should take drastic steps to provide food for today's famine victims (Singer). For a more detailed consideration, see her *Faces of Hunger* (Boston: Allen and Unwin, 1986). An independent critic is Bernard Williams, who claims in *Utilitarianism For and Against*, p. 137, that utilitarianism makes

> enormous demands on supposed empirical information, about peoples' preferences, and that information is not only largely unavailable, but shrouded in conceptual difficulty; but that is seen in the light of a technical or practical difficulty, and utilitarianism appeals to a frame of mind in which technical difficulty, even insuperable technical difficulty, is preferable to moral unclarity, no doubt because it is less alarming. (That frame of mind is in fact deeply foolish . . . .)

cally reveals that rather different idealizations — departures from reality in other directions — are doing all the work. For instance, unless some *deus ex machina* is handy to whisper in one's ear, it is far from clear just how one is to figure out how to limit the scope of the "maxims" of one's contemplated actions before putting them to the litmus test of the Categorical Imperative. There seems to be an inexhaustible supply of candidate maxims.

Certainly the quaint Benthamite hope of a fill-in-the-blanks decision procedure for ethical problems is as foreign to the spirit of modern Kantians as it is to sophisticated utilitarians. All philosophers can agree, it seems, that real moral thinking takes insight and imagination, and is not to be achieved by any mindless application of formulae.[5]

This is not meant to be a shocking indictment, but just a reminder of something quite obvious: no remotely compelling system of ethics has ever been made *computationally tractable*, even indirectly, for real-world moral problems. So even though there has been no dearth of utilitarian (and Kantian, and contractarian, etc.) *arguments* in favor of particular policies, institutions, practices, and acts, these have all been heavily hedged with *ceteris paribus* clauses and plausibility claims about their idealizing assumptions. These hedges are designed to overcome the combinatorial explosion of calculation that threatens if one actually attempts — as theory says one must — to *consider all things*. And as arguments — not derivations — they have all been controversial (which is not to say that none of them could be sound in the last analysis).

If there is a *Moral Almanac* actually in use, then, it is less like the *Nautical Almanac* than it is like *The Old Farmer's Almanac* — an unsystematic collection of wise sayings, informal precepts, tra-

[5] As Mill himself puts it, still in high dudgeon, "There is no difficulty in proving any ethical standard whatever to work ill if we suppose universal idiocy conjoined with it" (*Utilitarianism*, p. 31). This bit of rhetoric is somewhat at war with his earlier analogy, since one of the legitimate claims of the systems of practical navigation was that just about any idiot could master them.

ditional policies, snatches of taboo, and the like, a vade mecum vaguely approved of by the experts — who, after all, rely on it themselves — but so far lacking credentials.

There is not now any ethical theory that stands in the same relation to practical ethical decision-making as astronomy stands to practical navigation. That is hardly controversial. The hope of achieving such a theory has not been entirely abandoned, however. Theorists have attempted to salvage a close relation to practice by creating various brands of *indirect* utilitarianism, for instance, which factor in rules, dispositions, habits, institutions, and the like to "govern" our actual practices after receiving their credentials from the idealized theory. The general form of such licensing is an argument to show why and how, after all, *it is rational* or *it is optimizing* or *it is better* for agents to adopt (or just follow) these rules, inculcate (or just have) these dispositions, given some facts about the actual predicaments of such agents. This is still very much in the spirit of the "all things considered" tradition; for the proponents can claim that among *all* the things to be considered are certain crucial facts about agents' actual circumstances that were simply overlooked in the earlier idealizations. It is rather as if the earlier astronomers had neglected to notice the need for navigators to apply a variable "height-of-eye correction" to their sextant readings.

While I have learned a great deal from the recent work in this spirit, and even more from the critics of that work (I have in mind such authors as Gauthier, Gibbard, Hare, Parfit, Slote, and Williams), I will try to show that it still idealizes away from the heart of the problems. And while I think that compelling arguments can be (and have been) given to show that the hope of an astronomical foundation — an "Archimedean point," in Bernard Williams' phrase — should be abandoned as confused, I will simply assume in what follows that this is so.

Suppose, then, that we try to write the *Moral First Aid Manual*, but suppose further that we should write it with no

expectation that the Doctor of Philosophy will ever arrive with the Ultimate Right Answer.[6]

I should say at the outset that the job I envision, if done right, would involve systematic empirical studies and experiments by psychologists in addition to my informal and anecdotal explorations, and formal analyses of the task domains and the useful heuristics for them (of the sort sometimes produced by people in artificial intelligence), in addition to my intuitive guesswork. I am not ready to do this work, but — as philosophers are wont — I am ready to talk about why it would be *interesting* work for somebody to do. (And I should also add that previous Tanner Lecturers in Ann Arbor include two of the real pioneers in this endeavor: Herbert Simon and Thomas Schelling.)

2. JUDGING THE COMPETITION

To get a better sense of the difficulties that contribute to actual moral reasoning, let us give ourselves a smallish moral problem and see what we do with it. While a few of its details are exotic, the problem I am setting exemplifies a familiar structure.

Your Philosophy Department has been chosen to administer a munificent bequest: a Twelve-year Fellowship to be awarded in open competition to the most promising graduate student in philosophy in the country. You duly announce the award and its conditions in the *Journal of Philosophy*, and then to your dismay

[6] I will mention, but pass over, two other well-known reactions to the recognition that there is no hope of astronomy-like foundations for ethics: the defeatist banalities of "situation ethics" and other dreary relativisms of laziness on the one hand, and the useful — but, I think, only marginally useful — retreat to an "ethics of virtue" on the other. It is all very well to say, more or less with Aristotle, that if we *concentrate* our theoretical attentions on Virtue, the process of decision-making will take care of itself (since the Virtuous Person will know how to make morally wise decisions without any need to consult a Manual). This just passes the buck; how, exactly, is the paragon of Virtue supposed to do this? This "design" question remains achingly open — it is both theoretically and practically interesting, since few of us take ourselves to be beyond improvement in this regard — even if we agree (as we should not, in fact) that the ideally virtuous agent needs no help from our designers.

you receive, by the deadline, 250,000 legal entries, complete with lengthy dossiers, samples of written work and testimonials. A quick calculation convinces you that living up to your obligation to evaluate all the material of all the candidates by the deadline for announcing the award would not only prevent the Department from performing its primary teaching mission, but — given the costs of administration and hiring additional qualified evaluators — bankrupt the award fund itself, so that all the labor of evaluation would be wasted; no one would gain.

What to do? If only you had anticipated the demand, you could have imposed tighter eligibility conditions, but it is too late for that: every one of the 250,000 candidates has, we will suppose, a right to equal consideration, and in agreeing to administer the competition you have undertaken the obligation to select the best candidate.[7]

When I have put this problem to colleagues, I find that after a brief exploratory period, they tend to home in on one version or another of a mixed strategy, such as:

> choose a small number of *easily checked* and *not entirely unsymptomatic* criteria of excellence — such as grade point average, number of philosophy courses completed, weight of the dossier (eliminating the too light and the too heavy) — and use this to make a first cut. Conduct a lottery with the remaining candidates, cutting the pool down randomly to some manageably small number of finalists — say 50 or 100 — whose dossiers will be carefully screened by a committee, which will then vote on the winner.

There is no doubt that this procedure is very unlikely to find the best candidate. Odds are, in fact, that more than a few of the losers, if given a day in court, could convince a jury that they

[7] I don't mean to beg any questions with this formulation in terms of rights and obligations. If it makes a difference to you, recast the setting of the problem in terms of the overall disutility of violating the conditions set forth in your announcement of the competition. My point is that you would find yourself in a bind, whatever your ethical persuasion.

were *obviously* superior to the elected winner. But, you might want to retort, that's just tough; you did the best you could. It is quite possible, of course, that you would lose the lawsuit, but you might still feel, rightly, that you could have arrived at no better decision at the time.

My example is meant to illustrate, enlarged and in slow motion, the ubiquitous features of real-time decision-making. First, there is the simple physical impossibility of "considering all things" in the allotted time. Note that "all things" doesn't have to mean *everything* or even *everybody in the world*, but just "everything in 250,000 readily available dossiers." You have all the information you need "at your fingertips"; there need be no talk of conducting further investigations. Second, there is the ruthless and peremptory use of some distinctly second-rate cut rules. No one thinks *grade point average* is a remotely foolproof indicator of promise, though it is probably somewhat superior to *weight of dossier*, and clearly superior to *number of letters in surname*. There is something of a trade-off between ease of application and reliability, and if no one can *quickly* think of any easily applied criteria that one can have *some* faith in, it would be better to eliminate the first cut step and proceed straight to the lottery for all candidates. Third, the lottery illustrates a partial abdication of control, giving up on a part of the task and letting something else — nature or chance — take over for awhile, while still assuming responsibility for the result. (That is the scary part.) Fourth, there is the phase where you try to salvage something presentable from the output of that wild process; having *over*-simplified your task, you count on a meta-level process of self-monitoring to correct or renormalize or improve your final product to some degree.[8] Fifth, there is the endless vulnerability to second-guessing and

[8] See my "A Route to Intelligence: Oversimplify and Self-monitor" (CCM-85-4, Center for Cognitive Studies, Tufts), forthcoming in J. Khalfa, ed., *Can Intelligence Be Explained?* (Oxford University Press); and "Designing Intelligence" (CCM-86-4, Center for Cognitive Studies, Tufts), British Association for Advancement of Science, September 2, 1986.

hindsight wisdom about what you should have done — but done is done. You let the result stand and go on to other things.

The decision process just described is an instance of the fundamental pattern first explicitly analyzed by Herbert Simon, who named it "satisficing."[9] Notice how the pattern repeats itself, rather like a fractal curve, as we trace down through the sub-decisions, the sub-sub-decisions, and so forth until the process becomes invisible. At the departmental meeting called to consider how to deal with this dilemma, (1) everyone is bursting with suggestions — more than can be sensibly discussed in the two hours allotted, so (2) the chairman becomes somewhat peremptory, deciding not to recognize several members who might well, of course, have some very good ideas, and then (3) after a brief free-for-all "discussion" in which — for all anyone can tell — timing, volume, and timbre may count for more than content, (4) the chairman attempts to summarize by picking a few highlights that somehow strike him as the operative points, and the strengths and weaknesses of these are debated in a rather more orderly way, and then a vote is taken. After the meeting, (5) there are those who still think that better cut rules could have been chosen, that the department could have afforded the time to evaluate 200 finalists (or only 20), etc., but done is done. They have learned the important lesson of how to live with the sub-optimal decision-making of their colleagues, so after a few minutes or hours of luxuriating in clever hindsight, they drop it.

"But *should* I drop it?" you ask yourself, just as you asked yourself the same question in the midst of the free-for-all when the chairman wouldn't call on you. Your head was teeming at that moment (1) with reasons why you should insist on being heard, competing with reasons why you should go along with your colleagues quietly, and all this was competing with your attempts to follow what others were saying, and so forth — more informa-

[9] *Models of Man*, 1957; "Theories of Decision-Making in Economics and Behavioral Science," *The American Economic Review* XLIX (1959), pp. 253–83.

tion at your fingertips than you could handle, so (2) you swiftly, arbitrarily, and unthinkingly blocked off some of it — running the risk of ignoring the most important considerations, and then (3) you gave up trying to *control* your thoughts; you relinquished meta-control and let your thoughts lead wherever they might for awhile. After a bit you somehow (4) resumed control, attempted some ordering and improving of the materials spewed up by the free-for-all, and made the decision to drop it — suffering (5) instant pangs of dubiety and toying with regret, but, because you are wise, you shrugged these off as well.

And how, precisely, did you go about dismissing that evanescent and unarticulated micro-wonder ("should I have dropped it?")? Here the processes become invisible to the naked eye of introspection, but if we look at cognitive science models of "decision-making" and "problem-solving" *within* such swift, unconscious processes as perception and language comprehension, we see further tempting analogues of our phases in the various models of heuristic search and problem-solving.[10]

My suggestion, then, is that time-pressured decision-making is like that *all the way down*. Satisficing extends even back behind the fixed biological design of the decision-making agent, to the design "decisions" that Mother Nature — the process of natural selection — settled for when designing us and other organisms. There may be dividing lines to be drawn somehow between biological, psychological, and cultural manifestations of this structure, but not only are the structures — and their powers and vulnerabilities — basically the same; the particular contents of "deliberation" are probably not locked into any one level in the overall process but can migrate. Under suitable provocation, for instance, one can dredge up some virtually subliminal consideration and elevate

[10] The suggestion of temporal ordering in the five phases is not essential, of course. The arbitrary pruning of randomly explored search trees, the triggering of decision by a partial and non-optimal evaluation of results, and the suppression of second-guessing need not follow the sequence in time I outline in the initial example.

it for self-conscious formulation and appreciation — it becomes an "intuition" — and then express it so that others can consider it as well. Moving in the other direction, a reason for action perennially mentioned and debated in committee can eventually "go without saying" — at least out loud — but continue to shape the thinking, both of the group and the individuals, from some more subliminal base (or bases) of operations in the process. And as Donald Campbell and Richard Dawkins have argued, cultural institutions can sometimes be interpreted as compensations or corrections of the "decisions" made by natural selection.[11]

3. The Panglossian Slide

The fundamentality of satisficing — the fact that it is the *basic* structure of all real decision-making, moral, prudential, economic, or even evolutionary — gives birth to a familiar and troubling slipperiness of claim that bedevils theory in several quarters. To begin with, notice that merely claiming that this structure is basic is not necessarily saying that it is best, but that conclusion is certainly invited — and inviting. We began this exploration, remember, by looking at a moral *problem* and trying to *solve* it: the problem of designing a *good* (justified, defensible, sound) candidate evaluation process. Suppose we decide that the system we designed is about as good as it could be, given the constraints. A group of roughly rational agents — us — decide that this is the right way to design the process, and we have reasons for choosing the features we did.

Given this genealogy, we might muster the *chutzpah* to declare that this is optimal design — the best of all possible designs. This apparent arrogance might have been imputed to me as soon as I set the problem, for did I not propose to examine how *anyone*

[11] Donald Campbell, "On the Conflicts Between Biology and Social Evolution and Between Psychology and Moral Tradition," in *American Psychologist* (December 1975), pp. 1103–26; Richard Dawkins, *The Selfish Gene* (Oxford: Oxford University Press, 1976), ch. 11.

ought to make moral decisions by examining how *we in fact* make a particular moral decision? Who are we to set the pace?[12]

Optimality claims have a way of evaporating, however; it takes no *chutzpah* at all to make the modest admission that this was the best solution *we* could come up with, given our limitations.[13]

I call this the Panglossian slide, after Voltaire's optimistic Dr. Pangloss, who claimed that this is the best of all possible worlds, only to find that the pessimist agreed with him: no better world was possible, alas. The Panglossian slide is ubiquitous. In philosophy it regularly appears in debates about what is rational, in epistemology and in ethics. When it comes to defining *knowledge* as opposed to mere true belief, is "good enough" ever good enough? When it comes to doing the right thing, is it ever right to live by a rule you know to be sub-optimizing or non-maximizing? Can it be rational to opt on occasion for irrationality? This same question reappears in the interpretation of experiments in psychology, sometimes provoking quite hostile debates (for instance, between L. Jonathan Cohen and Amos Tversky about whether human irrationality can be experimentally demonstrated).[14] In

[12] Well, who else should we trust? If we can't rely on our own good judgment, it seems we can't get started: "Thus, what and how we do think is evidence for the principles of rationality, what and how we ought to think. This itself is a methodological principle of rationality; call it the *Factunorm Principle*. We are (implicitly) accepting the Factunorm Principle whenever we try to determine what or how we ought to think. For we must, in that very attempt, think. And unless we can think that what and how we do think there is correct — and thus is evidence for what and how we ought to think — we cannot determine what or how we ought to think." R. Wertheimer, "Philosophy on Humanity," in R. L. Perkins, ed., *Abortion: Pro and Con* (Cambridge, Mass.: Schenkman, 1974), pp. 110–11. See also Nelson Goodman, *Fact, Fiction and Forecast*, 2d ed., 1965, p. 63.

[13] Compare that with the claim: "Mother Nature isn't perfect, but she does the best she can." Is that a Panglossian statement or not? See my "Intentional Systems in Cognitive Ethology: The 'Panglossian Paradigm' Defended," *Behavioral and Brain Sciences* 6 (1983), pp. 343–90.

[14] L. J. Cohen, "Can Human Irrationality Be Experimentally Demonstrated?" *Behavioral and Brain Sciences* 4 (1981), pp. 317–70; see also "Continuing Commentary," *Behavioral and Brain Sciences* 6 (1983), pp. 487–517. Commentary by Tversky and others and replies by Cohen are included in these references.

biology it appears in the debate between the adaptationists, who make use of optimality assumptions, and their opponents, who claim — mistakenly — to be innocent of such ideology.[15]

The mistake that is sometimes made is to suppose that there is or must be a single (best or highest) perspective from which to assess ideal rationality. Does the ideally rational agent have the all-too-human problem of not being able to remember certain crucial considerations when they would be most telling, most effective in resolving a quandary? If we stipulate, as a theoretical simplification, that our imagined ideal agent is immune to such disorders, then we don't get to ask the question of what the ideal way might be to cope with them.

The *Moral First Aid Manual* should thus be considered not merely as a grubby compromise with practicality, but itself just as pure an ideal vision as any other in ethics: if you like, it is the book the ideally rational agent would write as his own vade mecum, written in the light of his perfect self-knowledge about his many limitations.

Any such exercise presupposes that certain features — the "limitations" — are fixed, and other features are malleable; the latter are to be adjusted so best to accommodate the former. But one can always change the perspective and ask about one of the presumably malleable features whether it is not, in fact, fixed in one position — a constraint to be accommodated. And one can ask about each of the fixed features whether it is something one would want to tamper with in any event; perhaps it is for the best as it is. Addressing that question requires one to consider still

[15] See note 13. A recent clear expression of the claim — which also clearly reveals its confusion, in my opinion — is Stephen Jay Gould, "Cardboard Darwinism," *New York Review of Books*, Sept. 25, 1986, pp. 47–52. Gould's long-term fascination with what he calls the "paradox" inherent in the unavoidable mixture of teleology and tinkering — *bricolage* or *satisficing* versus a God's-eye view of what is best — provides philosophers with a valuable guide to the pitfalls encountered in these issues. See also my "Evolution, Error, and Intentionality," forthcoming in my collection *The Intentional Stance*, from Bradford Books/MIT Press.

further ulterior features as fixed, in order to assess the wisdom of the feature under review. There is no Archimedean point here, either; if we suppose the readers of the *Moral First Aid Manual* are *complete* idiots, our task is impossible, while if we suppose they are saints our task is too easy to shed any light.[16]

4. SOME SUGGESTIONS FOR THE MORAL FIRST AID MANUAL

If *The Moral First Aid Manual* is to be optimally (or at least pretty well) addressed to a time-pressured decision-maker, it may help us design it if we slow the process down once more and look at what makes for good decision-making at the departmental level. First, of course, you want to have good colleagues: people who can be relied upon to come up with the right sorts of considerations right away, without wasting precious time on irrelevancies.[17]

This crucial component is often idealized away in ethical discussions via the introduction of what amounts to a Master of Cere-

[16] This comes out graphically in the slippery assumptions about rationality in theoretical discussions of the Prisoner's Dilemma; there is no problem if you are entitled to assume that the players are saints; saints always cooperate, after all. Near-sighted jerks always defect, so they are hopeless. What does "the ideally rational" player do? Perhaps, as some say, he sees the rationality in adopting the meta-strategy of turning himself into a less than ideally ratipnal player — in order to cope with the less than ideally rational players he knows he is apt to face. But then in what sense *is* that new player less than ideally rational? It is a mistake to suppose this instability can be made to go away if we just think carefully enough about what ideal rationality is. That is the *truly* Panglossian fallacy. (Cf. the reflections along these lines in A. Gibbard, "Moral Judgment and the Acceptance of Norms," Nicholas Sturgeon, "Moral Judgment and Norms," and A. Gibbard, "Reply to Sturgeon," in *Ethics* 96 [October 1985], pp. 5–41.)

[17] This "translates" readily into the discussions among utilitarians, familiar since Mill's day, of the value of inculcating *good habits of thought*. But the hunch that there is any straightforward way of getting such habits to *work* is of a piece with the complacent assumptions among *epistemologists* that hid the Frame Problem from them. See my "Cognitive Wheels: An Introduction to the Frame Problem of AI," in Christopher Hookway, ed., *Minds, Machines and Evolution: Philosophical Studies* (Cambridge: Cambridge University Press, 1984).

monies who handily *provides a frame*, by telling the agent exactly what the available options are.

"Your two choices *this* hour, Mr. Dennett, are

(A) stuff envelopes for Oxfam OR

(B) go to a movie!

What do you choose?"

Well, what happened to the option of watching the evening news, thereby informing myself of national and international problems, or answering some long-overdue letters, or spending the hour chatting with my daughter, or . . . ?

We need to have "alert," "wise" habits of thought — colleagues who will regularly, if not infallibly, draw our attention in directions we will not regret in hindsight. There is no point having more than one colleague if they are clones of each other, all wanting to raise the same consideration, so we may suppose them to be specialists, each somewhat narrow-minded and preoccupied with protecting a certain set of interests.

Now how shall we avert a cacophony of colleagues? We need some *conversation-stoppers*. In addition to our timely and appropriate generators of considerations, we need consideration-generator-squelchers. We need some ploys that will arbitrarily terminate reflections and disquisitions by our colleagues, and cut off debate independently of the specific content of current debate. Why not just a *magic word?* Magic words work fine as control-shifters in artificial intelligence programs, but we're talking about controlling intelligent colleagues here, and they are not apt to be susceptible to magic words, as if they were under post-hypnotic suggestion. That is, good colleagues will be reflective and rational, and open-minded within the limits imposed by their specialist narrow-mindedness. (They could take their motto from the philosophical journal *Nous*: *Nihil philosophicum a nobis alienum putamus* [We deem nothing philosophical to be foreign to us].)

They need to be hit with something that will appeal to their rationality, while discouraging further reflection.

It will not do at all for these people to be *endlessly* philosophizing, endlessly calling us back to first principles and demanding a justification for these apparently (and actually) quite arbitrary principles. What could possibly protect an arbitrary and somewhat second-rate conversation-stopper from such relentless scrutiny? A meta-policy that forbids discussion and reconsideration of the conversation-stoppers? But, our colleagues would want to ask, is *that* a wise policy? Can it be justified? It will not always yield the best results, surely and . . . and so forth.

This is a matter of delicate balance, with pitfalls on both sides. On one side, we must avoid the error of thinking that the solution is *more rationality*, more rules, more justifications, for there is no end to that demand. Any policy *may* be questioned, so unless we provide for some brute and a-rational termination of the issue, we will design a decision process that spirals fruitlessly to infinity. On the other side, no mere brute fact about the way we are built is — or should be — entirely beyond the reach of being undone by further reflection.[18] Although such fixed and hence sphexish (Hofstadter's term; for a discussion see *Elbow Room*, p. 11ff) features of our lives are unavoidable — indeed even sometimes essential — elements in our competence, no one of them is exempt from rational assessment, and we can always at least imagine what it would be like for the feature to be otherwise.

One cannot expect there to be a single stable solution to such a design problem, but rather a variety of uncertain and temporary equilibria, with the conversation-stoppers tending to accrete pearly layers of supporting dogma which themselves cannot withstand

[18] Stephen White, in "Self-Deception and Responsibility for the Self," forthcoming in a volume on self-deception, edited by A. Rorty, discusses Strawson's well-known attempt ("Freedom and Resentment," Proc. Brit. Acad., 1962) to terminate the demand for a justification of "our reactive attitudes" in a brute fact about our way of life about which "we have no choice." He shows that this conversation-stopper cannot resist a further demand for justification (which White provides in an ingeniously indirect way).

extended scrutiny, but which do actually serve on occasion, blessedly, to deflect and terminate consideration.

Here are some promising examples:

"But that would do more harm than good."

"But that would be murder."

"But that would be to break a promise."

"But that would be to use someone merely as a means."

"But that would violate a person's *right*."

Bentham once rudely dismissed the doctrine of "natural and imprescriptible rights" as "nonsense upon stilts" and we might now reply that perhaps he was right; perhaps talk of rights *is* nonsense upon stilts, but *good* nonsense — and good only because it is on stilts, only because it happens to have the "political" power to keep rising above the meta-reflections, not indefinitely, but usually "high enough," to reassert itself as a compelling — that is, conversation-stopping — "first principle."

It might seem then that "rule worship" of a certain kind is a good thing, at least for agents designed like us. It is good not because there is a certain rule, or set of rules, which is provably the best, or which always yields the right answer, but because having rules works — somewhat — and not having rules doesn't work at all.

But this cannot be all there is to it — unless we really mean "worship," i.e., a-rational allegiance, because just *having* rules, or *endorsing* or *accepting* rules is no design solution at all. Stephen White has suggested to me that a good bumper-sticker slogan for act utilitarians would be "Rules don't punish people; people do!" He goes on to point out that we can reinterpret this slogan as a reminder about the mistake he calls the Nominalist Fallacy: *there is nothing magical, or even forcing, about the mere* presence of a rule — or other intellectual property, such as a proposition — in a rational agent.[19] Having the rules, all the

[19] More secure than a-rational allegiance, White argues ("Self-Deception"), is a "self-supporting disposition" which flows from a set of "noninstrumental desires

information, and even good intentions does not suffice, by itself, to guarantee the right action; the agent must find all the right stuff and use it, even in the face of contrary rational challenges designed to penetrate the conviction.

Having, and recognizing the force of, rules is not enough, and sometimes the agent is better off with less. Douglas Hofstadter draws attention to a phenomenon he calls "reverberant doubt," which is stipulated out of existence in most idealized theoretical discussions. In what Hofstadter calls Wolf's Dilemma, an "obvious" non-dilemma is turned into a serious dilemma by nothing but the passage of time and the possibility of reverberant doubt.

> Imagine that twenty people are selected from your high school graduation class, you among them. You don't know which others have been selected, . . . All you know is that they are all connected to a central computer. Each of you is in a little cubicle, seated on a chair and facing one button on an otherwise blank wall. You are given ten minutes to decide whether or not to push your button. At the end of that time, a light will go on for ten seconds, and while it is on, you may either push or refrain from pushing. All the responses will then go to the central computer, and one minute later, they will result in consequences. Fortunately, the consequences can only be good. If you pushed your button, you will get $100, no strings attached. . . . If *nobody* pushed their button, then *everybody* will get $1,000. But if there was even a single button-pusher, the refrainers will get nothing at all.[20]

Obviously, you do not push the button, right? But what if just one person were just a little bit overcautious or dubious and began wondering whether this was obvious after all? Everyone should allow that this is an outside chance, and everyone should recog-

---

in Ideal Reflective Equilibrium" — a disposition which no rational criticism can challenge, because the subject has no desire to which a critical appeal can be directed.

[20] "Dilemmas for Superrational Thinkers, Leading Up to a Luring Lottery," in Hofstadter, *Metamagical Themas* (New York: Basic Books, 1984), pp. 739–55.

nize that everyone should allow this. As Hofstadter notes, it is a situation "in which the tiniest flicker of a doubt has become amplified into the gravest avalanche of doubt. . . . And one of the annoying things about it is that the brighter you are, the more quickly and clearly you see what there is to fear. A bunch of amiable slowpokes might well be more likely to unanimously refrain and get the big payoff than a bunch of razor-sharp logicians who all think perversely recursively reverberantly" (p. 753).[21]

Faced with a world in which such predicaments are not unknown, we can recognize the appeal of a little old-time religion, some unquestioning dogmatism that will render agents impervious to the subtle invasions of hyperrationality. Creating something rather like that dispositional state is indeed one of the goals of the *Moral First Aid Manual*, which, while we imagine it to be framed as *advice* to a rational, heeding audience, can also be viewed as not having achieved its end unless it has the effect of changing the "operating system" — not merely the "data," not merely the contents of belief or acceptance — of the agents it addresses. For it to succeed in such a special task, it will have to address its target audiences with pinpoint accuracy.

There might, then, be several different *Moral First Aid Manuals*, each effective for a different type of audience. This opens up a disagreeable prospect to philosophers, for two reasons. First,

[21] Robert Axelrod has pointed out to me that what Hofstadter calls Wolf's Dilemma is formally identical to Jean-Jacques Rousseau's Parable of the Stag Hunt (in the *Discourse on the Origin and Foundations of Inequality Among Men*). There is a good discussion of the Stag Hunt in Russell Hardin, *Collective Action* (Baltimore and London: Johns Hopkins University Press, 1982), pp. 167ff, from which it appears that Hofstadter's point has been missed in previous discussions, such as those of David Lewis, in *Convention* (Cambridge: Harvard University Press, 1969), and Kenneth Waltz, in *Man, the State, and War* (New York: Columbia University Press, 1965). On Lewis's discussion, "the problem of social cooperation [in Stag Hunt circumstances] does not seem intractable" — because Lewis ignores the prospect of imperfection in the agents, and while Waltz focuses on the possibility that agents with "limited time horizons" will turn such occasions into Prisoners' Dilemmas, there is no appreciation of the further point — Hofstadter's — that even if one supposes that such imperfections are extremely unlikely, the tiniest doubt — even a *groundless* doubt — can undo the stability of the solution.

it suggests, contrary to their austere academic tastes, that there is reason to pay more attention to rhetoric and other only partly or impurely rational means of persuasion; the ideally rational *audience* to whom the ethicist may presume to address his or her reflections is yet another dubiously fruitful idealization. And more important, it suggests that what Williams calls the ideal of "transparency" of a society — "the working of its ethical institutions should not depend on members of the community misunderstanding how they work" — is an ideal that may be politically inaccessible to us.[22] Recoil as we may from elitist mythmaking, and such systematically disingenuous doctrines as the view Williams calls "Government House utilitarianism," [23] we may find — this is an open empirical possibility after all — that we will be extremely lucky to find any rational and transparent route from who we are now to who we would like to be.

Rethinking the *practical* design of a moral agent, via the process of writing various versions of the *Moral First Aid Manual*, might nevertheless allow us to make sense of some of the phenomena traditional ethical theories wave their hands about. For one thing, we might begin to understand our current moral position — by that I mean yours and mine, at this very moment. Here we are, devoting an hour to my meta-meta-meta-reflection on values and valuation. Is this time well spent? Shouldn't we all be out raising money for Oxfam or picketing the Pentagon or writing letters to our senators and representatives about various matters? Did you consciously decide, on the basis of calculations, that the time was ripe for a little sabbatical from real-world engagement, a period "off line" for maintenance and inventory control? Or was your process of decision — if that is not too grand a name for it — much more a matter of your *not* tampering with some current "default" principles that virtually ensure that you

[22] *Ethics and the Limits of Philosophy*, p. 101. Williams notes that this is the ideal Rawls calls "publicity" in *A Theory of Justice*.

[23] Ibid., p. 108.

will ignore all but the most galvanizing potential interruptions to your rather narrow, personal lives?

If so, is that itself a lamentable feature, or something we finite beings could not conceivably do without? Consider a traditional bench test which most systems of ethics can pass with aplomb: solving the problem of what you should do if you are walking along, minding your own business, and you hear a cry for help from a drowning man. That is the easy problem, a conveniently delimited, already well-*framed* local decision. The hard problem is: how do we get there from here? How can we *justifiably* find a route from our actual predicament to that relatively happy and straightforwardly decidable predicament? Our prior problem, it seems, is that every day, while trying desperately to mind our own business, we hear a thousand cries for help, complete with volumes of information on how we might oblige. How on earth could anyone prioritize that cacophony? Not by any systematic process of considering all things, weighing expected utilities, and attempting to maximize. Nor by any systematic generation and testing of Kantian maxims — there are too many to consider.

Yet we do get there from here. Few of us are paralyzed by such indecision for long stretches of time. By and large, we must solve this decision problem by allowing an utterly "indefensible" set of defaults to shield our attention from all but our current projects. Disruptions of those defaults can only occur by a process that is bound to be helter-skelter heuristics, with arbitrary and unexamined conversation-stoppers bearing most of the weight.

That arena of competition encourages escalations, of course. With our strictly limited capacity for attention, the problem faced by others who want us to consider their favorite consideration is essentially a problem of advertising — of attracting the attention of the well-intentioned. This is the same problem whether we view it in the wide-scale arena of politics, or in the close-up arena of personal deliberation. The role of the traditional formulae

of ethical discussion as directors of attention, or as shapers of habits of moral imagination, is thus a subject for further scrutiny.

For better or for worse, your attention got attracted to my considerations for more than my share of time. I am grateful for it, and hope it proves to have been time well spent.