Social Philosophy & Policy Vol. 7 Issue 1, ISSN 0265-0525

TWO-TIER MORAL CODES

By Holly M. Smith

A moral code consists of principles that assign moral status to individual actions – principles that evaluate acts as right or wrong, prohibited or obligatory, permissible or supererogatory. Many theorists have held that such principles must serve two distinct functions. On the one hand, they serve a *theoretical* function, insofar as they specify the characteristics in virtue of which acts possess their moral status. On the other hand, they serve a *practical* function, insofar as they provide an action-guide: a standard by reference to which a person can choose which acts to perform and which not. Although the theoretical and practical functions of moral principles are closely linked, it is not at all obvious that what enables a principle to fill one of these roles automatically equips it to fill the other. In this paper I shall briefly examine some of the reasons why a moral principle might fail to fill its practical role, i.e., be incapable of guiding decisions. I shall then sketch three common responses to this kind of failure, and examine in some detail the adequacy of one of the most popular of these responses.

I. Practical Viability and its Barriers

What is it for an agent to use a principle in making a decision? Let us begin by saying that an agent uses a principle as a guide for making a decision just in case the agent chooses an act out of a desire to conform to the principle, and a belief that the act does so conform. Thus, suppose Susan decides to signal a lane change because she desires to follow the highway code, and believes the highway code requires lane changes to be signaled. She has used this code to make her decision. We may say, then, that a principle is *usable* by an agent for making a decision, just in case the agent is able to use it in this sense.[1]

What barriers might there be to someone's using a principle to guide her decision? We can see at once that there could be several. For example, the principle itself may suffer from defects that prevent its practical use. A principle may be so vague that it sometimes leaves the moral status of actions indeterminate. Consider a principle which states that killing persons is wrong, but fails to clarify whether 'persons' includes early human fetuses or not. Then no one can use this principle in deciding whether or not to obtain an abortion, since she cannot tell whether or not abortions are prohibited. Obviously, this kind of vagueness in a principle not only prevents it from being used to make decisions, but also detracts from its adequacy as a theoretical account of right and wrong, since such a principle leaves the status

[1] See Holly Smith, "Making Moral Decisions," *Nous*, vol. 22 (1988), pp. 91–92, for further discussion of the kinds of usability.

of many acts indeterminate. Such a principle is flawed *both* as a theory *and* as a practical guide. Moreover, its defects as a practical guide seem to depend directly on its defects as a theoretical account of right and wrong. Clearly the appropriate response here (if any) is to revise the principle itself.

But there are many attractive moral principles having no such obvious defects *qua* theoretical accounts of right and wrong that agents are nevertheless unable to use in making decisions. In an important range of these cases it is natural to ascribe the flaw to the agent rather than to the principle, viewed as a theoretical account of right and wrong. But philosophers have often considered principles subject to such handicaps to be flawed *as action-guides*. What I have in mind here are cases where the agent suffers from one or more of a variety of *cognitive handicaps* that prevent him from making a decision by reference to the principle in question. We can distinguish, at least initially, four major types of cognitive handicaps.

(A) First, the agent may, by reason of his cognitive limitations, be unable to *understand* the principle in question: to grasp some of its crucial concepts (whether these are evaluative, formal, or empirical), or to comprehend the overall structure of the principle. For example, Donald Regan has recently proposed a principle entitled "Co-operative Utilitarianism" designed to enable consequentialist-spirited agents to achieve the best possible co-operative outcomes. Co-operative Utilitarianism is stated as follows:

> Each agent must hold himself ready to take part in co-operative effort. He must identify others who are willing and able to do their part . . . He must ascertain the behavior or dispositions to behave of the *non-co-operators* who have been identified thus far (that is, the agents who are *not* willing and able to do their part), and he must ascertain the best pattern of behavior for the co-operators in the circumstances. He must then decide whether anyone he currently regards as a co-operator has made any mistake so far. If any putative co-operator has made a mistake, then all who have made mistakes are eliminated from the class of putative co-operators, and the process of identifying the best behavior for the (reduced) class of co-operators is repeated. And so on, until it is discovered that no putative co-operator has made a mistake. At this point the inquiry shifts to the question of whether the putative co-operators are all terminating their investigations into each others' decision-making. If any putative co-operator is not terminating his investigation here but is going on to another round of checking on his fellow co-operators, then the agent in question goes on also, to be sure of catching any last minute errors the others might make. Only when the agent in question discovers that the putative co-operators are all stopping does he stop and do his part in the current best plan.[2]

A more detailed exposition of the process involved in applying CU occupies two pages. Regan himself admits that "CU is complicated"; my experience in

[2] Donald H. Regan, *Utilitarianism and Co-operation* (Oxford: Clarendon Press, 1980), pp. 165–66.

attempting to teach CU suggests that many average agents would not, and perhaps could not, understand Co-operative Utilitarianism. They could not infer what the principle required them to do. Such an agent is not *necessarily* prevented from using this principle in the sense I defined above. Perhaps he can form the desire to follow the principle when it is described in a way he can comprehend (e.g., he may form the desire to "follow Regan's new principle"). And he may come to believe on the advice of some trusted authority, such as his philosophy professor, that (e.g.) *voting for the Democratic candidate for governor* is required by the principle so conceived, even if he cannot work this out for himself.[3] But if no such authority comes to his assistance, he cannot use this principle to make any decisions. Most of the moral principles with which we are most familiar are stated in a manner lending itself to the comprehension of the average person. However, many of these may be quite beyond the cognitive capacities of mentally less well-endowed agents, who nonetheless face a variety of moral dilemmas. The fact that the most familiar moral principles can be understood by most of us may already reflect a perceived necessity to construct moral principles to fall at least within the range of normal cognitive grasp.

(B) The second kind of problem arises because an agent may not possess, or may not be able to acquire within the time allotted to her for making some decision, the *empirical information* necessary for deriving a prescription from the principle in question. For example, consider a government leader who wants to follow act-utilitarianism in deciding whether to agree to a certain disarmament treaty. Unfortunately, the leader is uncertain whether agreeing to the treaty would maximize the general happiness and so cannot assent to any empirical premise stating that one of her alternatives has the right-making characteristic specified by her moral principle. Hence she can deduce no prescription from that principle, and cannot use it in making her decision. This is true even if the decision-maker can assign definite probabilities to a given act's satisfying the principle. The government leader may think there is an eighty percent chance that agreeing to the treaty will maximize happiness, and so believe there is an eighty percent chance that agreeing to the treaty is prescribed by her principle. But this does not enable her to infer what the principle actually requires her to do. And since our definition of a moral principle's usability requires that the decision-maker be able to infer what that principle prescribes – not what it *may* prescribe, or what it *probably* prescribes – the leader is unable to use her principle in deciding what to do.[4]

(C) A third problem arises when an agent has sufficient empirical beliefs to deduce a prescription from her principle, but some of these beliefs are false, so that the derived prescription is (or would be) incorrect. For example, a juror may want to follow a deontological principle requiring adequate compensation for injured plaintiffs. The juror believes, falsely, that the plaintiff suffered damages to the extent of $100,000, but actually his damages amounted to $500,000. Hence the

[3] If the authority is reliable, the agent may even *know* that he ought to vote for the Democratic candidate.

[4] See Smith for an account of the adequacy of the most popular technique for surmounting this problem, namely supplementing moral principles with auxiliary decision-guides or "rules of thumb" designed to deliver prescriptions when agents possess probabilistic information at best.

juror's decision to vote for an award of $100,000 on the ground that this would provide adequate compensation does not in fact satisfy her deontological principle. Because I shall be referring to this problem frequently, let us label it the *Problem of Error*.

(D) A fourth problem arises when an agent possesses enough empirical information to calculate what act is prescribed by his principle, but he is intellectually unable (perhaps altogether, or perhaps just within the time available) to make the necessary calculations. For example, suppose the reaction process in a nuclear power plant starts to run out of control. The chief engineer must decide within 30 seconds whether to close down the reactor or to add extra coolant, and, if so, how much to add. Let us imagine that the engineer wants to make this decision by reference to act-utilitarianism, and that he actually has all the necessary information about the numerous consequences, and the corresponding values, of each option. However, he cannot calculate, in the available 30 seconds, which *set* of consequences has the highest overall value. This kind of case is one where the agent has the intellectual capacity to make the necessary calculations and merely lacks the necessary time. But it is clear that there are some possible (and perhaps otherwise very attractive) moral principles that ascribe rightness to an action as a function of a mathematically complex combination of characteristics that might exceed the computational ability of any human being to calculate – including human beings using powerful resources such as computers to extend their own computational abilities.

There are, then, at least four cognitive handicaps that could prevent human agents from utilizing a variety of moral principles in actual decision-making: incapacity to comprehend the principle, lack of sufficient information to apply it, erroneous empirical beliefs, and limited ability to make the requisite calculations. It is worth emphasizing that these problems may affect deontological principles as well as consequentialist ones. The difficulties I have described cut right across the consequentialism/deontologism distinction. These difficulties show that principles that might appear quite attractive as theoretical accounts of right and wrong may fail in many cases to be usable for decision-making. How are we to react to this failure?

II. Responses

The responses of moral theorists who have explicitly confronted this problem have tended to cluster into three different categories. I shall describe each kind of response briefly.

The first kind of response, which I shall call the *Replacement Response*, has been adopted by a wide variety of moral thinkers. According to these thinkers, the theoretical function of morality cannot be isolated from its practical or regulative function, in the sense that one test of a moral principle's theoretical correctness just *is* its practical usability. In David Lyons's words, these thinkers hold that moral principles must be designed to accommodate "the mistakes we make, the errors to which we are prone . . . our blockheadedness, ignorance, confusion, and stupidity."[5]

[5] David Lyons, *The Forms and Limits of Utilitarianism* (Oxford: Clarendon Press, 1965), p. 159.

Such Replacement Response theorists, on noting that the practical use of act-utilitarianism is hindered for most decisions by our lack of information about the future, have claimed that this fact provides good and sufficient reason to reject act-utilitarianism as a theoretical account of what makes acts right and wrong. Some such theorists have replaced it with a more readily usable deontological theory. Others have replaced it with "prospective" act-utilitarianism, which only requires the agent to determine which action would maximize expected utility, not which action would maximize actual utility. Others have adopted rule-utilitarianism for the same reason. Another example of the Replacement Response is provided by John Rawls, who argues that any acceptable principle of justice must be simple enough for everyone to understand, such that ascertaining which institutions satisfy the principle does not depend on information that is difficult to obtain.[6] To use a slogan, we might say that the Replacement Response attempts to narrow the gap between human decision-making capacities and the requirements of moral theory by *lowering* the theory to a level where fallible human beings can employ it.

The second kind of response, which I shall call the *Conserving Response*, claims that a moral principle's practical usability, or lack thereof, is no sign of its adequacy or inadequacy as a theoretical account of right and wrong. Conserving Response theorists tend to view moral principles on the model of scientific theories, and point out that we do not determine the truth or falsity of a scientific theory by ascertaining whether it would be easy or difficult to make predictions on the basis of that theory. Any difficulties we may experience in making predictions on the basis of a well-confirmed scientific theory should be seen as defects in *us*, not defects in the theory. Similarly, Conserving Response theorists say, if we are unable to use some normatively correct moral principle to guide our choices – because we lack sufficient empirical information, or are unable to perform the necessary calculations – that is a defect in us, not a defect in the theory. This kind of view is clearly expressed by Derek Parfit, who denies that a principle S is faulty because erroneous empirical beliefs prevent him from complying with S: "If this is the way in which S is self-defeating, this is no objection to S. S is self-defeating here only because of my incompetence in attempting to follow S. This is a fault, not in S, but in me."[7] To the extent that Conserving Response theorists are concerned with the practical use of moral principles, their slogan might be "Eliminate the gap between human decision-making capacities and the requirements of moral theory by *raising* human capacities to the level where human beings can employ the correct theory." Their advice to us is to improve ourselves by acquiring greater empirical information, increasing our ability to store information where it may be easily accessed, and employing computers to enhance our computational capacities. We should not tinker with the theory merely to disguise our own shortcomings.

[6] John Rawls, *A Theory of Justice* (Cambridge: Harvard University Press, 1971), p. 132, and "Construction and Objectivity," *The Journal of Philosophy*, vol. LXXVII (September 1980), p. 561.

[7] Derek Parfit, *Reasons and Persons* (Oxford: Clarendon Press, 1984), p. 5. Parfit applies his remarks to self-interest principles, not to patently moral principles.

The third response, which I shall call the *Moderate Response*, rejects both extremes, and claims that the best solution involves a two-pronged strategy. *First*, we are to determine which principle is the correct theoretical account of right and wrong without any reference to the practical usability of such a principle. *Second*, if that account proves impractical for making decisions, then we are to supplement it with appropriate second-level rules that are more readily applied in making decisions by human beings operating under normal constraints of information and computation. Perhaps the classic statement of the Moderate Response is found in John Stuart Mill, who used it to defend utilitarianism against the objection that "there is not time, previous to action, for calculating and weighing the effects of any line of conduct on the general happiness." Mill believed that "*whatever* we adopt as the fundamental principle of morality, we require subordinate principles to apply it by."[8] This response is a kind of halfway house between the two previous extremes, in the sense that, on the one hand, it denies that the content of the account of right and wrong must accommodate human cognitive limitations, but, on the other hand, it requires a moral theory to accommodate these limitations by *expanding* to include normatively appropriate decision-making rules as well as principles of right and wrong. Notice that the Moderate Response does not claim that we will be able to apply the correct principle of right and wrong *directly* to our decisions. Rather, the idea is that we will apply it *indirectly*, via direct application of the second-level rules to our acts. Thus, the demand that moral principles should be usable must be weakened to the demand that they should be usable at least in this indirect sense.[9]

III. Rationales for the Usability Condition

Which of these responses to cognitive deficiencies in applying moral principles is correct? Clearly, the answer to this question will largely depend on one's rationale for believing that moral principles should be usable as practical decision-guides. Here I will briefly indicate four of the most salient reasons that have been used to support the idea that moral principles must be usable.

First, it may be argued that the very concept of morality requires that moral principles be usable for action-guiding purposes. Sometimes this is expressed as Prichard does, when he claims that ordinary thought holds that there can be no particular duty that is not recognized as such by the person obliged to do it.[10] Thus Prichard would hold that the juror has no duty to compensate the plaintiff with $500,000, since she is unaware this is her duty. Other moralists, of whom Hare might provide an example, hold that it is part of the meaning of moral terms, such as 'ought', that their function is to help guide choices. Clearly, to fulfill this role it is necessary that the principles governing the application of these terms be usable.

[8] John Stuart Mill, *Utilitarianism* (Indianapolis: The Bobbs-Merrill Company, Inc., 1957), pp. 30, 32. My emphasis. It is not wholly clear that Mill had in mind by "subordinate principles" precisely what I do here.

[9] Certain moral codes, such as utilitarianism, are often criticized on the ground that they demand too much of mere human beings by way of motivation: they require us to perform acts involving so much sacrifice of our own interests that no one could possibly be motivated to adhere to such principles. This is a criticism about the "strains of commitment." Notice that the same three responses that I have just outlined to problems of cognitive deficiency could also be proposed as responses to problems of motivational deficiency.

[10] H.A. Prichard, "Duty and Ignorance of Fact," in H.A. Prichard, *Moral Obligation and Duty and Interest* (London: Oxford University Press, 1968), pp. 18–39.

Still other moralists have held that "she ought to do A" implies "she can do A" in the sense "she would do A if she wanted to." This, in turn, implies that the agent *knows how* to perform the act in question. Moralities subject to this constraint could not be subject to some of the cognitive impediments to usability I have described.

A second justification offered in favor of usability has been stated most persuasively by Bernard Williams. He points out that there seems to be a special form of injustice created by moral principles which cannot be universally used. Suppose it turns out, for example, that certain moral principles cannot be used as widely by the dull or the poorly informed as by the highly intelligent and well-educated. Such a morality would violate the ideal that the successful moral life be available to *everyone*. Williams, who traces this ideal back to Kant, claims that it has the ultimate form of justice at its heart and embodies something basic to our ideas of morality.[11]

A third justification offered in favor of usability in moral principles holds that the function of a moral code is to enhance social welfare. Warnock speaks for many when he states that "the 'general object' of morality ... is to contribute to betterment – or non-deterioration – of the human predicament."[12] The usual idea here is that a moral code is to serve as a kind of informal analogue to a legal code, constraining behavior in ways that make every member of society better off. The connection between serving this function and being usable is thought to be roughly this: moral rules must be designed so that (a) they can be successfully followed, and (b) when they are successfully followed, they will increase social utility through actions that avoid violent conflict, enhance social cooperation, and so forth. Rules that cannot be followed cannot be guaranteed to lead to such desirable results, since the acts, and the consequences of those acts, resulting from misapplications of such rules are unpredictable or even pernicious.

Finally, a fourth justification for usability states that the function of morality is not to produce valuable social consequences, but rather to produce the best possible *pattern of actions* (where desirable actions are specified by the theoretical criterion provided by the morality). If certain actions are right, then it is a good thing if they are performed. Typically, but not always, moralists who take this line defend a deontological moral code. They argue that the ideal pattern of actions can only be achieved if the moral principles are usable without hitch or error by the individuals subject to them. Otherwise misapplications, or failures of application, will lead to morally inferior acts.

In this paper I shall take on a very limited task: what I shall examine is the extent to which a particular version of the Moderate Response constitutes a successful solution to the Problem of Error – the problem raised by the fact that decision-makers often have false empirical beliefs that would lead them to derive incorrect prescriptions from otherwise attractive moral principles. In considering which response is the best solution to the general problem of cognitive deficiency, I believe it is crucial to distinguish the various *kinds* of cognitive deficiencies that can hinder decision-making. Since a given response may provide a satisfactory solution

[11] Bernard Williams, "Moral Luck," in Bernard Williams, *Moral Luck* (Cambridge: Cambridge University Press, 1981), p. 21.

[12] C.J. Warnock, *The Object of Morality* (London: Methuen and Co., Ltd., 1971), p. 26.

to one kind of cognitive handicap but not to another, it is important not to be misled about the appropriateness of a response by conflating the various kinds of shortcomings it might be invoked to circumvent.

IV. Two-Tier Responses

As a response to the Problem of Error, the Moderate Response consists in advocating what I shall call a *two-tier system*. In this system, the first tier consists in principles that provide the correct theoretical account of right-making characteristics. Let us call the set of these principles *M*. The second tier consists in rules that are to be used for actual decision-making. Let us call these principles *M**. Since people are often mistaken about the empirical nature of their prospective acts, they often err as to which acts are required by M, and so in attempting to follow M sometimes perform acts that it *pro*scribes. M*, on the other hand, is so constructed that agents who attempt to apply it will do what M itself prescribes. Two-tier systems of this sort are most familiar when the first tier is consequentialist and the second deontological. Sidgwick's view that commonsense morality ought to be used by most people in decision-making, even though utilitarianism is the correct account of right and wrong, is a salient example of this kind of proposal. However, any combination of deontological and consequentialist tiers is possible. For example, someone who believed the first tier should be deontological in character might recognize that the correct principles would be misapplied by many people, owing to their mistaken beliefs about the world. Such a theorist could advocate at the second level a set of principles less subject to erroneous application: these could either be simpler deontological principles or even simple consequentialist ones (referring only to easily ascertained effects of actions). Thus a deontologist might believe that in certain extreme circumstances, the use of torture by military officials is justified; but he might also believe that the likelihood of such officials' incorrectly believing themselves to be in these circumstances is so great that it would be better if they settled the issue of torture by reference to the simple rule "Never use torture."

For purposes of assessing the adequacy of the two-tier approach, one must begin by considering an *ideal* M* – i.e., one such that attempts to follow it would *always* lead the decision-maker to do what M prescribes.[13] There may seem little hope of

[13] Notice, however, that there seems no reason to demand that M* itself avoid the Problem of Error. That is, agents may make mistaken applications of M*, so long as their doing so does not lead them to violate M itself.

As Eric Mack pointed out in a discussion of this paper, there may be a difficult equilibrium problem in constructing coextensive pairs of M and M*, at least in cases where M is consequentialist. What concrete actions a consequentialist M requires depends on the specific historical context, which includes the nature of the moral code believed by the general population. Thus if the population believes code C, M may require agent S to perform act A (since it would lead C-believers to pursue certain courses of actions), while if the population believes code C', M may require agent S instead to perform act B (since it would lead C'-believers to pursue different courses of action than they would have had they believed in C). Thus to identify the relevant M*, we cannot simply start with M and ask what code would be coextensive with it; instead we have to start with M and a possible concrete historical context, including general belief in a given code, and ask whether that code is coextensive with M under those conditions. If not, we look at a different possible historical context and ask the parallel question, until finally we have found a matching pair. This may not be an easy task.

In this paper I am confining my attention to first-tier moral codes (i.e., candidates for M) that are *purely behavioral*: that is, they prescribe actions characterized solely in behavioral terms, not actions partly characterized in terms of the agent's beliefs, intentions, or other motivational states. Without this restriction it would be difficult or impossible to construct a coextensional M*, at least if that required the agent to have the same mental state as that required by M, as well as to perform the same bit of behavior required by M.

identifying such an M* for any M which we seriously believe might be an adequate theoretical account of moral rightness. However, I believe this pessimism is incorrect: ideal M*s can be identified. I will discuss how to do so later on. Given that there are ideal M*s, and hence ideal versions of the two-tier approach, we should focus on assessing them first, and only subsequently turn to an assessment of non-ideal versions of the two-tier strategy. Only in this way will we be able to see the advantages and disadvantages *intrinsic* to this kind of solution. We must beware of prematurely rejecting a solution because of difficulties or objections that arise only for non-ideal versions of the solution, the ones falling short of the objective: these difficulties may arise from the failure of the version to meet its objective, not from any inherent flaw in the strategy itself. Having assessed ideal versions, we may then turn to non-ideal versions. This paper, however, must confine itself to consideration of the ideal versions alone.

Two-tier solutions appear in various forms. The first major distinction among such forms concerns the extent of knowledge about the structure of the moral system permitted to the decision-maker who is to use the second-tier rules. One version, which we might call the *Enlightened Decision-Maker* version, allows this agent full awareness of the relation between M and M*. A second version, and the one I shall discuss here, keeps the agent in the dark about the status of M*. In this *Benighted Decision-Maker* version, the decision-maker falsely believes that M* is the correct theoretical account of rightness and wrongness, and never learns that the real role of M* is to secure conformity to M. On some variants of the Benighted Decision-Maker version, a coterie of enlightened persons retains knowledge of the true roles of M and M*. The elite itself uses M both for theoretical assessments and for decision-making purposes. Sidgwick labels such an arrangement an *Esoteric Morality*; it shall be the main focus of our discussion here.[14] On other variants, which Parfit calls *Self-Effacing*, even the enlightened see that it would be best if *they* no longer believed (and tried to follow) M, and so (perhaps by hypnosis) replace their belief in M with a belief in M* as the true theoretical account of rightness and wrongness. They then apply M* in their decision-making, and always wind up doing what M itself prescribes. No one remains who recognizes that M, not M*, is the correct theory about what makes actions right or wrong.[15]

Esoteric Morality solutions to the Problem of Error have had many detractors. Many of these detractors have viewed it as a solution required only by consequentialist, and particularly utilitarian, moralities, and therefore as constituting a disadvantage of such moralities as compared to others. As we have seen, it is a mistake to suppose that consequentialist moralities are the only ones vulnerable to the Problem of Error; consequently, it is a mistake to believe that consequentialist moralities are the only ones for which this problem might be remedied by the Esoteric Morality solution. If this solution is a poor solution, it should be avoided by any morality; and if a better solution cannot be found, then all moral systems are disadvantaged equally by the necessity of employing it. Let us

[14] Henry Sidgwick, *The Methods of Ethics*, 7th ed. (Chicago: The University of Chicago Press, 1962), pp. 489-90.

[15] Parfit, sec. 17.

begin our assessment of Esoteric Morality solutions by examining some of the standard objections that have been lodged against them. I shall argue that these objections have significantly less force than is commonly thought. I shall then argue that these solutions must be rejected for a wholly different type of reason.[16]

V. The Objection from Manipulation

Bernard Williams attacks the Esoteric Morality solution to systems with a utilitarian first tier. He envisions a society in which the secretly-utilitarian rulers encourage and maintain a non-utilitarian morality on the part of the general populace. This situation, he claims, would be inherently manipulative, because the rulers must be unresponsive to non-utilitarian demands made on them, and maintain their political position by means other than responsiveness to public demands. These means are likely to involve coercion or severe political restrictions.[17]

But Williams's complaint has little force against an ideal two-tier system. For the record, we should note that the secret utilitarian elite in an Esoteric Morality need not be the political rulers at all. They may have no particular power over the views or activities of others, but simply realize that the general populace produces more utility by following their nonutilitarian morality than they would by attempting to follow utilitarianism. The elite sit back and watch the situation with approval; but they may have no power to change it even if they wished to. Such an elite can hardly be charged with manipulation of the sort Williams describes.

But the more important point is that even a scenario in which the political rulers *do* form the utilitarian elite fails to be manipulative in Williams's sense. Williams claims that because of the difference in moral beliefs between the general populace and the rulers, the populace will demand that the government act in certain ways which the government must refuse to do. For example, the populace might demand, but the government resist, the setting of equitable (but non-utility-maximizing) taxes, the upholding of treaties (when violating them would better promote utility), or the punishment of criminals in a manner commensurate with

[16] In this paper I will focus primarily on the capacity of M* to secure the same pattern of action as M. Of course, on many views, M and M* would need to be compared on other grounds. For example, M* might be more costly overall to social welfare than M because it would be so difficult to teach; or M* might actually secure fewer right actions than M because even though people would be infallible in applying it, it would be far less capable of eliciting allegiance than M, and so produce less actual compliance. For the most part I shall leave these issues aside.

It is worth pointing out here, however, that a kind of two-tier morality (with a version of utilitarianism as the first tier, and a set of deontological rules as the second tier) has sometimes been proposed as a technique for avoiding *normative* objections to act-utilitarianism. Thus it is claimed that act-utilitarianism erroneously requires (for example) a sheriff to convict and punish an innocent person in order to avert race riots. This counter-intuitive result, it is said, can be averted by a system of rules prohibiting punishment of the innocent. Such a system allegedly could be justified on general utilitarian grounds, even though it would not prescribe every utility-maximizing individual act. This type of rationale for a two-tier system is not compatible with the kind of rationale I am exploring. The rationales explored in this paper assess a second-tier rule as better insofar as the acts it prescribes *match* those prescribed by the first-tier principle, while the normative-objection rationale only succeeds if the second-tier rules sometimes deliver prescriptions that *diverge from* those of the first-tier principle. I am grateful to Julia Annas for calling this point to my attention.

[17] Bernard Williams, "A Critique of Utilitarianism," in J.J.C. Smart and Bernard Williams, *Utilitarianism: For and Against* (Cambridge: Cambridge University Press, 1973), pp. 138–39.

the evil of their crimes (rather than in a manner calculated to maximize deterrence effect). But in the Esoteric Morality solution under consideration, the alternative morality M* to which the populace is committed is *coextensional* with the correct morality M (in this case, utilitarianism): all acts and policies required by the one are also required by the other. So, in general, the policies demanded by the populace will be the very same policies desired by the utilitarian elite.[18] There will be no more need in such a society for the government to resist its citizens' demands than is normally the case. No unusual engines of coercion will be necessary to ensure that the elite's policies are carried out. (Of course, insofar as utilitarianism requires certain forms of manipulation, both the governors and the populace will demand such coercion. But this is a straightforward consequence of utilitarianism as such, not a consequence of embedding utilitarianism within a two-tier system.) Manipulation in Williams's special sense is certainly not required by the presence of an ideal two-tier system: it only becomes necessary insofar as the system falls short of ideal correspondence between M and M*. Failure, not success, of the two-tier system gives rise to manipulation.

VI. Violating the Publicity Condition

We have seen that Esoteric Moralities will not, by themselves, lead to coercive political manipulation. When M and M* are coextensional, M does not require implementing governmental policies that contradict the will of the populace. The will of the populace and the will of the moral elite are the same. But it may be claimed that even ideal Esoteric Moralities will inevitably involve another evil, namely *deceit*. For the elite are required to conceal their own morality, and thus to violate what is now called, following John Rawls, the "Publicity Condition." The Publicity Condition, whose foremost advocate is Rawls, states that moral principles are invalid unless they can be publicly advocated without being self-defeating.[19] Rawls advocates the Publicity Condition on the following ground. He states that the principles of justice chosen by contractors in the Original Position are to serve as a public conception of justice: one which is acknowledged by all parties to a dispute, and which can be publicly appealed to by anyone in order to settle interpersonal conflicts within society. Thus the contractors in the Original Position will only choose principles in the understanding that they will be public knowledge.[20]

However, Rawls's assumption about the role of principles of justice does not provide a conclusive argument against an ideal Esoteric Morality. In a society with an ideal Esoteric Morality, the general populace *does* possess a common and publicized moral theory – namely M* – to which members of the general populace can appeal in order to resolve interpersonal disputes among themselves. Likewise

[18] More accurately, the two moralities are coextensional except for the cases in which it is the populace's *misapplication* of M* which would lead them to do or want what M requires. But in these cases what the populace mistakenly thinks required by their theory is what is actually required by the rulers' theory M, so there will be no conflict between the populace and the rulers on the moral character of the policies in questions.

[19] Rawls, *Theory of Justice*, p. 133. Rawls traces the history of the condition to Kant.

[20] *ibid.* See also the Dewey lectures (*Journal of Philosophy*, vol. LXXVII (September 1980)), where this idea is developed in more detail.

the elite possesses a common and publicized (among themselves) moral theory – namely M – to which they can appeal in order to resolve interpersonal disputes within their own ranks.[21] Of course, M and M* are different. But this presents no problem even when disputes arise between the elite and the general populace, since the elite, knowing of the correspondence between M and M*, will be perfectly content to abide by the prescriptions generated by M*. Hence, none of the bad effects Rawls envisions arising from non-public moralities will actually arise under the Esoteric Morality we are considering.[22]

Of course, this does not obviate the fact that the elite must deceive the general populace about the moral views they themselves hold: unless unusual conditions obtain, the elite must tell the general populace, falsely, that they believe M*. Such a situation is sometimes objected to on the ground that it involves serious psychological strains on the deceiving parties, who must always guard their statements and never reveal, perhaps even to those individuals who are personally closest to them, the nature of their true values. Of course, these strains will be much alleviated by the fact that M and M* prescribe the same acts, so that the elite are never required to assert they are in favor of acts that, in reality, they abhor. To the extent that these strains are a problem (and this will depend on the precise social arrangements and degree of interaction between the elite and the general populace), then they must be counted as among the costs of a two-tier system. If the rationale for a two-tier system is to increase social welfare above what it would be if everyone were to believe and attempt to follow M, then these costs must be taken into account in determining whether social welfare really *would* be higher under a two-tier system: perhaps the loss would be large enough to outweigh the gains secured through universal compliance with M. In such a case, the two-tier system would fail to secure its objective and should not be adopted, at least if other possible solutions to the Problem of Error would be less costly. It must be said, however, that

[21] There may be some disputes between members of the elite that must be carried out in full view of the general populace. In such cases, the elite cannot overtly appeal to M. However, they will be content to appeal to M* itself, since they know it generates the same prescriptions as M. Complexities might arise if the case in question is one in which the general populace would, through some erroneous factual belief not shared by the elite, derive an "incorrect" prescription from M* – a prescription that actually accords with what M itself prescribes (see note 13). In such a case, the elite would have to feign the same factual beliefs as the general populace.

[22] Avoiding these bad effects may not be as simple as the text suggests. So far I have spoken as though both M and M* governed the actions of both the elite and the general population. Technically, however, M* need only govern the actions of the general population (since they are the only ones subject to the Problem of Error). Nonetheless, if M* failed to address the activities of the elite, it would be difficult to persuade the general population that such an incomplete M* was the genuine theoretical account of right and wrong. Hence M* must probably be constructed to govern the activities of all. Now, it is logically possible that the actions required by M for the two groups differ. For example, it might turn out, according to M, that the general population ought never to lie, while it is permissible for the elite to lie under circuirstances C (which never arise for the general population). Hence M* might be constructed to contain two components, M* (GP) which forbids the general population to lie, and M*(E) which permits the elite to lie under circumstances C. But it would probably be more psychologically effective to construct a coextensional M* which permitted lying to *anyone* so long as they found themselves in circumstances C. Thus, the general population would know that they, too, could lie if they ever were in circumstances C. (But suppose 'circumstances C' = 'being an elite when the general population needs to be misled about the true moral code in order to avoid the Problem of Error'. An M* containing a clause referring to such a C would certainly tend to undermine the system as a Benighted Agent solution.)

this seems unlikely. The size of the elite is likely to be small, so that the number of those subjected to these psychological strains will not be large. Hence the cost to them is likely to be outweighed by the benefits secured through universal compliance with M.

Suppose, on the other hand, that a two-tier system is not advocated because of its benefits for society, but rather because it is held that the concept of morality requires usability for decision-making, or that justice requires that everyone should be able to lead the successful moral life, or that it is a good thing that right acts be done. Then what? Placed against these claims, the fact that the elite must suffer some psychological strains to support the required system seems of small or uncertain importance. To promote justice, for example, it is often necessary to make great sacrifices of self-interest; what the elite are required to do to maintain an Esoteric Morality is certainly no more costly than what justice might require from them in other contexts. The objection that a two-tiered system may generate psychological strains seems of little consequence in the context of these other rationales in favor of this solution.

Still a third objection may be raised because of the deceit itself: it may be thought that deceiving others about one's moral views is inherently immoral, and any solution to the Problem of Error that requires it must be rejected for this reason. The first thing we should notice about this objection is that it is hardly theory-neutral. Anyone who is convinced that the correct theoretical account of right-making properties, i.e., M, is consequentialist will not agree that deception is inherently immoral. Hence this person will have no moral reason to reject a system that involves deception about their moral views by the elite. On the other hand, someone who believes that M is deontological may believe that M itself includes a prohibition against deception. Such a person will have *moral* reason to reject an Esoteric Morality in which the elite must deceive the general populace. The question for such a deontologist will be now to weigh M's prohibition of deception against whatever rationale he accepts in favor of the widespread usability of morality. Suppose, for example, he believes the dictum that " 'ought' entails 'can' " requires morality to be usable by everyone, and at the same time rejects the Replacement Solution, according to which it is a constraint on the correctness of any theoretical account of rightness that it be directly usable by everyone for making decisions. A two-tier solution seems his only recourse, even if it requires forms of deception prohibited by what he believes to be the correct theoretical account of rightness. What we have here is a conflict between *ethical* reasons against a social arrangement and *meta-ethical* reasons in favor of it. It is a difficult to know how such conflicts might be resolved, but it is far from clear that the ethical considerations must always outweigh the meta-ethical ones. It is noteworthy that in many deontological codes the prohibition against deceit is only a *prima facie* prohibition at most, often outweighed by conflicting considerations that may point the other direction. By contrast, meta-ethical considerations are rarely thought of as merely *prima facie*. We cannot conclude that the prohibition against deceit, even for a deontologist, shows that Esoteric Moralities should not be accepted as the best solution to the Problem of Error. And for the consequentialist, the objection will have no weight at all.

VII. The Objection from Cognitive Manipulation

We have surveyed the objection to an Esoteric Morality that it requires the elite to violate the Publicity Condition, i.e., to disguise their own moral beliefs. However, it is one thing to conceal one's own moral beliefs; it is quite another to dupe others into affirming false moral beliefs. Nevertheless, it may be objected, this is precisely what the elite must do in order to bring about belief in the incorrect M* by the general population. Thus, for example, the elite must deceive the general populace into believing that an action is wrong because it involves stealing, whereas in fact it is wrong because it fails to maximize happiness. Someone who concedes that it is sometimes necessary to conceal one's own moral beliefs may well find it far more objectionable to induce incorrect moral views in others.

Once again, we should note that Esoteric Moralities need not involve systems in which a political or educational elite manipulates the moral views of the general population. The moral elite may have no power over the views of the general populace; they may merely note with approval that the population's adherence to M* leads them to do precisely what M demands. Such an elite cannot be charged with manipulation of beliefs.[23]

But what about a two-tier system in which the moral elite *does* influence the moral views of the general population by leading them to believe in M*? How objectionable would such a system be? Clearly, any serious objection to such a system will start from the premise that the moral elite is *deceiving* the general population, and the judgement that deception is wrong. As we have just seen, anyone who believes in a consequentialist M will not be moved by this objection. Moreover, even someone who believes in a deontological M that includes a prohibition against deception may find that this prohibition is outweighed by the importance of rendering morality universally usable. But whether or not it is outweighed will depend on how wrong the deception is, and this in turn surely depends on how grave a harm or disadvantage deception is to the deceived party. A relatively harmless lie must be far less evil than a lie that significantly harms its victim. In this case the harm is whatever harm is constituted by having false moral beliefs. Thus this objection, to seriously worry a deontologist, must show or assume that having false moral beliefs is a grave harm or disadvantage to the general population in an Esoteric Morality.

The question of whether holding incorrect moral views is a harm or disadvantage to the person who holds them is a difficult one. It cannot be answered definitively here, but it is worth sketching some of the considerations. Normally this question is not hard to answer, because false moral beliefs lead to wrong actions, and we may feel that it is clear enough what is wrong about wrong actions. Thus there is, so to speak, an action-oriented explanation for why false moral beliefs are to be avoided. But in the scenario we are envisioning, the false moral

[23] An interesting proposal, somewhat along these lines, has been suggested by Nigel Smith (see "Enchanted Forest," *Natural History*, vol. 92 (August 1983), pp. 14–20). Smith recounts the (patently false) superstitious beliefs that prevent the rural populations of the Amazon basin from destroying the jungle ecological system, and recommends "tapping" these folk beliefs in order to strengthen official conservation efforts.

beliefs lead to right actions, so we cannot criticize them on those grounds. (And, indeed, true moral beliefs would lead to wrong actions.) So the most natural ground for thinking that possession of false moral views is unavailable in this context.

But there may be other grounds for objection. According to a Kantian tradition, lying to a person is wrong because it interferes with her effective exercise of her rational agency.[24] This tradition has been picked up in medical ethics contexts, where it is said, for example, that a patient's *autonomy* requires the physician to disclose all relevant information before the patient makes any decision about a proposed course of treatment. But what is "rational agency" or "autonomy"? These are notoriously vague concepts. They may be defined in such a way that a person only counts as rational, or autonomous, if the person makes decisions on the basis of true beliefs. On such definitions, of course misleading a person about morality will undermine her rational agency and autonomy. But why should we accept *these* definitions of "rationality" and "autonomy"? Why must these characteristics depend on true belief? In the context we are considering, this question becomes critical. For it cannot be claimed that here the person possessing the false belief labors under a deficiency that will lead her to make the wrong decision, or a decision she would regret if she had true beliefs. On the contrary, if she had true beliefs in this sphere, she would make the wrong decision. Her effective decision-making is not undermined by her false views. Pursuit of this line of thought seems to lead to a dead end: we must first establish why true (moral) beliefs are important, before we can argue that rationality and autonomy, properly understood, require them.

Why, generally speaking, are true beliefs valuable? Epistemologists who discuss this issue (with respect to empirical beliefs) typically cite the importance of true belief for successful action.[25] But we are dealing with a case where, by hypothesis, false (moral) belief would lead to successful action, and true belief would lead to unsuccessful action.[26] So that rationale is unavailable to us. It might plausibly be claimed that we simply *want* to have true moral opinions (just as we might want to know if a deceased spouse had been unfaithful to us) quite apart from the usefulness of such opinions in making correct moral decisions. We want to grasp the nature of moral reality, even if we would act just as well not knowing its nature. But would we want to know this if we realized that such knowledge would lead us to act badly? At the very least this would require a difficult balancing judgment as to whether action or knowledge was most important.

A venerable tradition within philosophy maintains that one's fundamental moral beliefs constitute part of one's character.[27] Thus, someone who believes that

[24] See, for example, Barbara Herman, "The Practice of Moral Judgment," *The Journal of Philosophy*, vol. LXXXII, no. 8 (August 1985), p. 431.

[25] See, for example, Robert Nozick, *Philosophical Explanations* (Cambridge: Harvard University Press, 1981), p. 284; see also pp. 323–26. But see Alvin Goldman, *Epistemology and Cognition* (Cambridge: Harvard University Press, 1986), p. 98.

[26] Nozick, p. 321, speculates that the only way an "action can track an evaluative fact is via . . . the person's knowledge of the fact." But our case is one in which there is a counterfactual connection between the evaluative facts (specified by M) and their M* counterparts. So a person's belief in M* would enable her actions to "track" the genuine evaluative facts identified by M.

[27] Gilbert Ryle, "Forgetting the Difference Between Right and Wrong," ed. A.I. Melden, *Essays in Moral Philosophy* (Seattle: University of Washington Press, 1958), pp. 147–59.

stealing is fundamentally wrong has a different character from someone who believes that failing to maximize happiness is fundamentally wrong – even though it may be true (unbeknownst to the person) that stealing always fails to maximize happiness. A second, equally venerable tradition holds that having a good moral character is morally important in itself, quite apart from the acts it leads one to perform.[28] On this tradition, a society of people, each of whom believes the false M*, would be a *worse* society than another society in which everyone believes the true M – even though the acts performed in the two societies were identical. And even if the acts performed in the first society were better, the whole situation (encompassing both the pattern of action and the people's moral characters) might be morally worse. To decide whether it was actually worse would take a careful weighing of the relative value of good character versus right acts.

But we need to look more carefully at the claim that someone acting from incorrect moral theory has an inferior moral character. Is it really true that adherents of M* would have worse moral characters than adherents of M? It might be objected (as perhaps Kant would) that although the M* adherents are mistaken as to the *content* of their duty, it is nonetheless true that if their fundamental motive is to do their duty because it is their duty, then they are morally on a par with M adherents, whose fundamental motive is also to do their duty. But many people find this claim implausible. It entails, for example, that if Hitler genuinely believed that it was morally required to eliminate the Jews, then his moral character cannot be worse than the moral character of, say, Mother Theresa. And this seems to many people incredible. In their view, the nature of a person's moral character includes not only his desire to fulfil his duty, but also the content that he perceives that duty to have. Mother Theresa doing her perceived duty is a better person than Hitler doing his perceived duty.

But matters are more complicated than the Hitler example may make them appear. For one thing, even if one's moral character depends on the content of one's moral beliefs, it may also depend on their psychological history and status. Thus, it might be said that we cannot view a person as having bad character, even if his moral views are evil, if his coming to have those views (and maintaining them) was reasonable. So a racist who imbibes his racism from plausible authority figures, who never interacts with members of the downgraded race, and never has any opportunity to interact with them or to discover their real characteristics might be said not to have a bad character. Precisely this scenario may obtain for adherents to M*. We can imagine that they come to hold their views in the same reasonable way that adherents of M come to hold their views: e.g., both are taught their views by friends, family, and religious institutions; in both cases, plausible justifications for the respective theories are provided. So neither party believes his theory on unjustified grounds. Moreover, if M is correct, and M* is coextensional with it in this world, then the features picked out by M* as right-making are unlikely to be morally monstrous ones (in the assessment of M). At worst, they are likely to be

[28] Immanuel Kant, *Foundations of the Metaphysics of Morals* (Indianapolis: Bobbs-Merrill Company, Inc., 1959).

mildly evil, and at best morally neutral or even mildly good. Otherwise the coextensionality between and M and M* would not obtain. So the errors embodied in M* are not likely to leap out at any right-minded person, and we cannot fault its adherents on the ground that they have maintained their false view in the face of overwhelming evidence against it. Moreover, if the features of actions identified by the adherents of M* are morally unobjectionable (as assessed by M), then we cannot reasonably judge the characters of the M* adherents as being morally monstrous.

It seems, then, as though this objection to Esoteric Moralities is, at best, a mild one. No compelling argument that true moral beliefs are of overriding value has been put forward; certainly no argument has established that having true moral beliefs outweighs the evil of the wrong actions such beliefs would sometimes lead to in the kind of case we are considering. We saw that the moral character of a person who believes M* *may* be inferior to the moral character of someone who believes the correct M; but it is controversial whether or not this is so at all, since moral character may simply turn on one's desire to do one's duty, the difference in their characters is likely to be small in any event, and, in any case, there is certainly no guarantee that the loss in good moral character of a populace believing in M* would not be counterbalanced by the gain in right actions performed by them. The objection to Esoteric Moralities that focuses on the fact that a populace under such a Morality must have false moral beliefs is a troubling one, but under close scrutiny it doesn't have the power we might originally have expected. And from this we can conclude that the objection that Esoteric Moralities require the elite to deceive the general populace about the content of morality is at best a weak one; this particular form of deceit has not been shown to be very serious.

VIII. Problems of Implementation

We have now surveyed three of the standard objections that are raised to Esoteric Moralities as solutions to the Problem of Error. We have found that, at least for *ideal* Esoteric Moralities, these objections have far less force than is commonly supposed. Several miss their mark altogether, and others depend on a comparative weighing of the importance of different desiderata (e.g., avoiding wrong acts vs. avoiding poor moral character) that may or may not show Esoteric Moralities to be, on balance, undesirable. Let us turn our attention, then, to another kind of problem that more decisively undermines this solution.

Let us start by asking whether it is really possible to implement an Esoteric Morality. The first issue here is whether or not there exist second-level rules of the kind we have been discussing. For each candidate moral system M, vulnerable to the Problem of Error, does there exist a corresponding set of rules M* which is such that attempting to follow M* would lead each decision-maker to do what M prescribes? Proponents of Esoteric Moralities tend to speak as though there are laws of nature connecting the morally significant act-types identified by M and those identified by M*. Such a law of nature might state that, for example, every act of telling a lie (wrong according to M*) is also an act of failing to maximize utility (wrong according to M). The act-types identified as morally significant by M*

must be ones with respect to which members of the general populace are *infallible* – these agents must make no mistake about which acts would be of these types, even in cases where they would make mistakes about which acts are of the corresponding M-identified types.[29] But it seems extremely unlikely that, for any M of genuine interest, there exist simple correlations between the occurrence of M-identified act-types and any act-types that could serve in a corresponding M* theory. Lying does not always involve failing to maximize utility, and even if it did, agents are not always infallible in their beliefs about whether a prospective act would be a case of lying. Parallel things are true of every act-type that might be mentioned as a candidate for an M* theory. There seems no prospect of finding any laws of nature that will do the trick here, and so no prospect of finding an ideal system M*. Presumably it is precisely this assumption that has led proponents of Esoteric Moralities to focus their attention solely on non-ideal variants of this solution.

However, we should not give up so easily. Absence of the requisite nomological connections does not establish that no appropriate system M* exists. For each action that is of a type identified as significant by M is simultaneously of *many* other types.[30] In every case, at least one of these other types is one which the agent could unerringly ascribe to the act.[31] Suppose, in a particular case, that the act prescribed by M is also of (M-irrelevant) type T, the agent correctly believes that she has an act of type T available to her, and she knows how to perform this act. Then an instruction in this case to perform an act of type T would lead the agent to perform the act prescribed by M. For example, although the agent may not know which of her prospective acts is of the M-prescribed type *fully compensate an injured plaintiff*, that very act is also of the type *vote for an award of $500,000*, and she does know how to perform this act. If she wanted to vote for an award of $500,000, she would do so, and in doing so she would in fact carry out the demand of her moral code. Of course, which act-type correlates in this way with the M-prescribed type, and is also such that the agent knows how to perform it, will vary from case to case, depending on the circumstances and the agent's beliefs. Thus there will be no simple rules to substitute for M. But the rules of M* may take the form of an extended list of prescriptions to perform individual actions. Each prescribed action would be described in terms of an act-type having the feature that if the agent tried to do an action of that type, he would perform the act actually required by M in those circumstances. Such a list might appear as follows: at ten o'clock, empty the

[29] More accurately: the act-types must either be ones with respect to which the agents are infallible, or else such that the agent who wants to perform an act of that type will in fact perform the act prescribed by M itself.

[30] In an alternate terminology: any act of an M-significant type is on the same act-tree with many acts of different types.

[31] I assume here that if the agent is able to perform the act at all, then there is some description of it under which the agent's desiring to perform it would lead to his performance of that act. This may be too strong. There might be cases in which no *correct* description of the act would elicit its performance. (Consider the familiar finger game in which the fingers of both hands are entangled in such a way that one becomes confused as to which fingers belong to which hand. In these circumstances, wanting to *straighten the first finger of one's left hand* will elicit straightening the first finger of one's *right* hand, but no accurate description of this act will elicit it.) I shall ignore such cases in the discussion in the text; they imply that a thorough list might need to include misdescriptions of the actions to be performed.

dishwasher; at quarter past ten, pay one's bills; at eleven o'clock, balance one's checkbook; and so forth. Presented with such a list, the agent could follow it and so do everything required of him by M – even though he might not believe any of these acts to (for example) maximize utility and so would not perform them if he were instructed instead to act so as to maximize utility. Such lists could be relativized to each agent. An agent armed with a suitably designed list of this sort, and morally motivated, would perform each of the actions prescribed by M.

So appropriate systems M*, of a peculiar kind, do exist.[32] But their mere existence does not show that the Esoteric Moralities can provide a viable solution to the Problem of Error. There are several insurmountable obstacles to success. First, although for each M an appropriate M* exists, there is no reason to believe that anyone knows, or could find out, what the content of any appropriate M* is. Certainly the decision-maker herself cannot determine what the content of the appropriate M* is, for the decision-maker could only determine this if she knew what M requires in each particular case. But by hypothesis, the decision-maker would (at least sometimes) err if she tried to determine the prescription of M for each particular case. So no one who needs an M* theory to avoid the Problem of Error can construct that theory for herself. Of course, in any case where the decision-maker would err, some *other* individual might know which act was to be done, and might know under what description of it the decision-maker would be led to perform the correct act. But there is no reason to believe that, for *every* decision the decision-maker must make, there is someone who would know this. Nor is there reason to suppose that a person possessing this knowledge is always in a position to instruct the decision-maker as to what to do. There is even less reason to think that there is any one individual, or small group of individuals, who have this kind of knowledge about every act and every decision-maker subject to the Problem of Error, and who have instructional access to all these decision-makers. But this is what would be needed for an Esoteric Morality to work. Indeed, the kind of rules that M* requires – an extended list of acts – is not simple enough to be learnable in advance by any person of normal intelligence. Hence the moral elite would have to operate literally as guardian angels, hovering constantly about and advising the decision-makers from moment to moment as to what to do. Thus the difficulties in actually implementing this solution appear overwhelming.

How critical is the "implementation" problem? If the rules of M* were designed (at least in part) to serve a theoretical function – i.e., to provide an account of right-making characteristics – the difficulty I have just described would not be devastating. The rules still *would* provide this account, even if no one could determine what the content of the rules was.

But the rules of M* are only designed to serve a *practical* function: to enable decision-makers to act as M commands. The difficulty we have just seen shows that they cannot do this. There are two distinct, although related, ways to describe this failure. On the one hand, we can say that what we wanted was a practical solution to a practical problem. An analogy here might be our needing a solution to a practical

[32] But see note 13.

problem such as taking last year's license plates off our car. To get the plates off we need a Phillips screwdriver. It is no solution to the problem to be told that a Phillips screwdriver exists somewhere in the house: to get the plates off, we need the right screwdriver actually in hand. If we cannot find it, we cannot solve our problem. Similarly, if our practical problem is to bring it about that agents with mistaken factual beliefs nonetheless do what M prescribes, it is no help to be told that a certain M* exists which is such that if these agents attempted to follow it, they would do what M demands. To get the agents to do what M demands, we need the right M* actually in hand. Since neither we nor anyone else can identify this M*, we have *not solved our problem.*

A second way to see the nature of this difficulty is the following. The Problem of Error arises because many agents have mistaken factual beliefs that mislead them when they attempt to follow M. To avoid this problem, it is suggested that we adopt an alternative moral system, in which M is supplemented by a suitable second-tier M*. It is demonstrated that an M* exists which is such that if these agents attempted to follow it, they would actually do what M demands. Under the new M-M* system, the agents suffer no Problem of Error that would lead them to misapply the rules they are to use in making decisions, namely M*. However, the cognitive deficiency that hindered them relative to the simple M system now simply reappears in M-M* at another level: they have traded one Problem of Error for another. The same empirical misinformation that plagues their application of M, now prevents them from seeing that M* is the correct code by which to guide their actions. They can apply M*, but they cannot see that it, rather than some alternative, is justified as an action-guide. Their false empirical beliefs have been converted into false beliefs about the moral status of M*. But the end result is that the two-tier solution has not improved their overall situation – it has only altered the location at which their deficiency impedes successful practical reasoning.

IX. A Further Problem

It is now clear that ideal Esoteric Moralities – ones that avoid the Problem of Error altogether – are possible in principle, but technically infeasible, given the limitations of human knowledge and memory. Esoteric Moralities fail because they assume greater knowledge among the moral elite than this class possesses or could ever possess. But suppose, contrary to fact, that some elite group *did* have the necessary information: they could construct the "rules" of M* (i.e., ad hoc lists of prescribed actions), and could communicate those rules to ordinary decision-makers. It appears that such a group would, by that very fact, possess an alternative method of securing general compliance with M. Such a group would *also* be in a position simply to convey their empirical information to ordinary decision-makers, and allow them to derive the correct prescriptions from M itself. An elite group with sufficient information to construct the rules of M*, and suitable instructional access to ordinary decision-makers, would also have sufficient information and access to teach these decision-makers (perhaps moment by moment) the facts they need to know in order to apply M itself unerringly. In other words, such a group would be in a position to implement a *Conserving Response* solution to the Problem

of Error: conserving because it avoids the Problem by retaining M in both a theoretical and practical role, and enhancing the knowledge of decision-makers to the point where M itself is no longer vulnerable to the Problem. Indeed, insofar as the standard objections to Esoteric Moralities have weight, the Conserving Response would be superior to the Esoteric Morality solution, since it would be weakened by none of these flaws. For example, it would require no coercive manipulation of moral beliefs, nor any false moral beliefs at all on the part of the general population. So even if an Esoteric Morality could successfully be implemented, a superior Conserving Response solution would by that very fact be available. The upshot is this: ideal Esoteric Moralities, forms of Moderate Response solutions to the Problem of Error, are unworkable. If they could work, an equivalent Conserving Response solution would *ipso facto* also be available – and, insofar as the standard objections to Esoteric Moralities have force, the Conserving Response would be superior. We have fair reason, then, to reject ideal Esoteric Moralities as a solution to the Problem of Error.

Philosophy, University of Arizona