Decisions of Principle, Principles of Decision

ROBERT NOZICK

THE TANNER LECTURES ON HUMAN VALUES

Delivered at

Princeton University
November 13 and 15, 1991

ROBERT NOZICK, currently Arthur Kingsley Porter Professor of Philosophy at Harvard University and past Chairperson of the Philosophy Department, was educated at Columbia College and Princeton University. He is a member of the Council of Scholars of the Library of Congress, a Fellow of the American Academy of Arts and Sciences, and a Senior Fellow of the Society of Fellows at Harvard. He is also a member of PEN, America, and the Author's Guild. He was a cultural advisor to the U.S. Delegation to the UNESCO Conference on World Cultural Policy held in Mexico City in 1982, and has made several television programs for public broadcasting, serving as host and interviewer. He is the author of numerous books and articles, including *The Examined Life* (1989) and *Philosophical Explanations* (1981), which received the Ralph Waldo Emerson Award of Phi Beta Kappa. Professor Nozick received a National Book Award for *Anarchy, State and Utopia* (1974).

1. How to Do Things with Principles

What are principles *for?* Why do we hold principles, why do we put them forth, why do we adhere to them? We could instead simply act on whim or the passion of the moment, or we could maximize our own self-interest and recommend that others do the same. Are principles then a constraint upon whim and self-interest, or is adherence to principles a way of advancing self-interest? What functions do principles serve?

Principles of action group actions, placing them under general rubrics; linked actions are then to be viewed or treated in the same way. This generality can serve different functions: intellectual, interpersonal, intrapersonal, and personal. I start with the intellectual.

Intellectual Functions

Consider judicial decision making. In one system, a judge simply decides a case so as to yield what she thinks is the best or right or preferable result in that particular case. Another system of judicial decision involves principled decision: a common law judge is to formulate a principle to fit (most or almost all) past precedents and a range of hypothetical cases and then use this principle to decide the current case. Trying to formulate an acceptable general principle is a *test* of your judgment about the par-

These lectures were delivered at Princeton University, where I was a graduate student, and I dedicate them to my teachers there: to Carl Hempel and to the memory of Gregory Vlastos. The first draft of these lectures was written at the Rockefeller Foundation Research Center at Bellagio, Italy, in the summer of 1989. I am grateful to the discussants of these lectures at Princeton, Clifford Geertz, Gilbert Harman, Susan Hurley, and Amos Tversky, and also to Scott Brewer, David Gordon, Christine Korsgaard, Bill Puka, Tim Scanlon, and Gisella Striker, for their very helpful comments and suggestions. Special thanks to Amartya Sen for the many discussions of this material we have had, inside of class and out, and to Laurance Rockefeller for his interest in and generous support of this research project.

ticular case: is there *some* adequate general principle —a principle that gives the right result in all established cases and obvious hypothetical ones —that also yields the result you want in this case? If you cannot find such a principle, reconsider what result you do want in this case.

Such a procedure is a test of a particular judgment on the assumption that any correct judgment is yielded by *some* true acceptable general principle, that true particular judgments are consequences of general principles applied to specific situations. If search fails to uncover an acceptable general principle that yields some judgment in particular, this may be because there is no such acceptable principle, in which case that particular judgment is mistaken and should be abandoned. Or perhaps you have not been astute enough to formulate the correct principle. We have no mechanical procedure to decide which is the explanation.[1]

When you find a general principle or theory that subsumes this case, a principle you would be willing to apply to other cases as well, this particular judgment receives new support. Consider empirical data points *a, b,* c, *d*. If a straight line is the simplest curve through these, this supports the prediction that another point *e*, also on that straight line, will hold. It is not an easy matter —inductive logicians have discovered —to isolate and explain how a (relatively) simple lawlike statement can group existing data points so that inferences and predictions legitimately can be made about new points. Nevertheless, we do not doubt that data can support the hypothesis that a law holds and also support a prediction that a new point will accord with that law. Similarly, the simplest principle that covers acceptable normative points *a, b, c, d* also will support an additional judgment *e* (that

[1] A weaker assumption would maintain not that *every* correct judgment is yielded by an acceptable principle but that some or most are. Still, finding an acceptable general principle that yields a particular judgment would (tend to) show that judgment was correct. However, failing to find one would not be a conclusive reason for abandoning the judgment, for it might be one of those that stands alone, no consequence of any acceptable principle.

fits this principle) as a correct normative point too. A theorist gains confidence in his particular judgment (or side in a controversy) when he can formulate a general principle or theory to fit it, especially one that is appealing on its face.[2]

Philosophers of science have tried to demarcate scientific laws from accidental generalizations. Accidental generalizations only happen to hold, or to have held, true. From such a generalization —for example, that all the coins in my pocket are dimes — one cannot infer a subjunctive statement such as: if there *were* an additional coin in my pocket now, it *would be* a dime. (Whereas from a scientific law —for instance, that all freely falling bodies fall a distance equal to $^1/_2gt^2$—we can infer that if some other object now at rest were in free fall it would travel a distance equal to $^1/_2gt^2$.) If all previous data fit a given generalization, we can plausibly infer that new data would fit it (and hence predict that new data that *will* be gathered will fit it) *only if* that generalization is of lawlike form and is a candidate for being a law. It is when data fall under a lawlike statement (or arise from several of them) that we can legitimately extrapolate to further cases. The features of a lawlike statement, those aspects that differentiate it from an accidental generalization, constitute our license to travel from given data to predictions or expectations about further data. Similarly for particular normative judgments: what licenses us to travel to a further judgment on the basis of previous ones is the previous ones' all falling under a normative general principle. The

[2] Mark Tushnet has argued that in the legal arena the requirement of principled decision constitutes no constraint upon the result a judge can reach; if the previous cases fit a principle (even an established one) whose result the judge wishes to avoid in the present case, this case always can be distinguished from the others by some feature or other. (See "Following the Rules Laid Down: A Critique of Interpretavism and Neutral Principles," *Harvard Law Review,* 96 [1983], 781–827.) However, merely to distinguish the case (at best) allows the new judgment, it does not support it. To support it, the judge would have to formulate a new principle, plausible on its face, that fits (most of) the old cases, this new one and some obvious hypothetical ones as well; that is, she would need a principled rationale for the distinction she wishes to make, and for why that distinction should make a difference. It is no easy matter to formulate acceptable principles, much less to do this as frequently as one's desires about new cases would mandate.

features of a normative principle license a subjunctive inference to a new case that steps beyond the indicative instances that happen already to have fallen under it. Principles are transmission devices for *probability* or *support;* this flows from data or cases, via the principle, to judgments and predictions about new observations or cases whose status otherwise is unknown or less certain.

What features enable principles to transmit probability? The following features have been mentioned to distinguish scientific lawlike statements (or nomic universals) from accidental generalizations.[3] Lawlike statements do not contain terms for particular individual objects, dates, or temporal periods —or if they do, these statements can be derived from more general lawlike statements that do not. Lawlike statements contain purely qualitative predicates: stating the meaning of these does not require reference to any *particular* object or spatio-temporal location. Lawlike statements have an unrestricted universality; they are not simply a finite conjunction that was established by examining all cases. Lawlike statements are supported not just by instances falling under them, but also by a linkage of indirect evidence.

These very same features might be what enables a normative principle to license the derivation of new judgments from previously accepted ones. Writers on ethics frequently say ethical principles must be formulated using general terms only, no names of particular persons, groups, or nations. This feature might enable a principle to license an inference to a new case, hence enable new normative judgments to be supported by previous ones. A generalization lacking this feature of non-particularity might be, at best, an accidental one, incapable of transferring support from some data to others. When moral principles are general and do not contain any non-qualitative predicates or particular names, rather than being a specifically *moral* aspect of the principles, this feature might link data or judgments together to support subjunc-

[3] See C. G. Hempel, *Aspects of Scientific Explanation* (New York: Free Press, 1965), pp. 264–72, and Ernest Nagel, *The Structure of Science* (New York: Harcourt, Brace and World, 1961), pp. 47–78.

tive inferences. It would be worthwhile to investigate how much of the "form" of moral principles is necessary for such linkage.

This does not mean these features are tagged onto weaker generalizations to make moral principles that perform inferential functions, any more than such features are tagged onto accidental generalizations to make scientific laws. One can hold that scientific laws and moral principles each hold true apart from any constructions we add or any uses of them we make, that their independent truth is what makes these uses possible. Nevertheless, features such as generality, containing no proper names, no positional predicates, would not be specifically *moral* features, but lawlike ones, necessary for anything to be a law, scientific or moral. In context, not specifically moral features may have moral consequences.

A person may seek principles not only to test her own judgment or give it more support but to convince others or to increase their conviction. To do this she cannot simply announce her preference for a position — she must produce reasons convincing to the others. Reasons might be very particular but also can be general considerations that apply well to a wide range of cases and point to a particular judgment in this instance. If these judgments in the other cases are ones the other person already accepts, then the general reasoning will recruit these cases as evidence and support for the judgment proposed in the present case. Principles or general theories thus have an interpersonal intellectual function, justification to another. Justification by general principles is convincing in two ways: by the face appeal of the principles and by recruiting other already accepted cases to support a proposed position in this case.[4]

In using a judge to illustrate the testing and support function of principles, I have imagined that her purpose is to arrive at the right decision about a particular case and that she treats the past

[4] Abstract principled reasoning lends support to a particular position by recruiting other accepted judgments as support, but is this only one particular mode of justification. an abstract and impersonal *male* mode?

decisions as (for the most part) right themselves. That is, I have treated a judge as structurally identical to a moral reasoner who wishes to decide what is right or permissible on this new occasion or situation and who utilizes her knowledge of what is right or permissible in other actual or hypothetical situations to formulate, test, and support a moral principle that yields a result for this situation.

Of course, a judge also is a figure in an institutional structure; principled decisions that fit past cases may have a particular point within that institution. Legal theorists tell us the doctrine of respecting precedents, *stare decisis,* can enable people to predict more exactly the legal system's future decisions and so to plan actions with some confidence about their legal consequences.[5] For this effect, the precedents need not have been decided correctly or be followed with the goal of reaching a right decision; they are followed in order to yield a result that has been expected. Second, principled decisionmaking might be desired to constrain a judge's basis for decision. To be excluded are her personal preferences or prejudices, moods of the moment, partiality for one side in a dispute, or even thought-through moral and political principles that are personal to her. It might be held that a judge's own views, preferences, or even considered views should have no more effect than anybody else's — the judge was not given that institutional position to put her own preferences into effect. A requirement that decisions be principled fittings to past precedents might be a device to *constrain* the effect of such personal factors, limiting their play or crowding them out *altogether.*

However, the analogy to science where the aim is truth and correctness casts doubt upon the last strong claim. Fitting the scientific data is a requirement — here too there is leeway, and different ways a "best fit" can be defined. But this does not uniquely

[5] I have not checked to see what empirical studies of people's decisions exist to support this *empirical* claim by the legal theorists, what alternative legal structure functioned as the control, etc.

determine one lawlike statement. An indefinite number of curves can fit any finite set of data points; more than one will be lawlike. Hence additional criteria will be necessary to select which lawlike statement to tentatively accept and use in predicting. These criteria include simplicity, analogy to supported lawlike statements in related areas,[6] fit with other accepted theories, explanatory power, theoretical fruitfulness, and perhaps ease of computation.[7] Merely requiring that a prediction fit the past data according to some lawlike statement does not uniquely determine that prediction. How likely is it, then, that merely requiring that a judge's decision in a new case fit past decisions according to some principle will suffice to uniquely determine that decision? Indeed, we find judges enjoined to utilize additional criteria, including various "formal" ones.[8] We can raise analogous issues about ethics too. Quine holds that the totality of (possible) empirical data does not uniquely determine an explanatory theory. Are correct ethical principles uniquely determined by the totality of correct judgments about particular cases, actual and hypothetical, or does underdetermination reign there? In addition to fitting particular judgments, must a moral principle also satisfy certain further criteria?

There is a connection between using principles as a device for reaching correct decisions and using them to constrain the influ-

[6] All of the small number of data points we possess seem to fall on a straight line, but for all related phenomena we have found that a linear relationship does not hold. Perhaps here it is an accident of the particular data we happen to have.

[7] See the list of factors in Thomas Kuhn, "Objectivity, Value Judgment and Theory Choice," in his *The Essential Tension* (Chicago: University of Chicago Press, 1977), pp. 320–39, and W. V. Quine and Joseph Ullian, *The Web of Belief,* 2nd ed. (New York: Random House, 1978), pp. 64–82. The need for such additional criteria may not result just from the finiteness of our data. Quine has claimed that the totality of all possible observations does not uniquely select an explanatory theory. (See his "On the Reasons for Indeterminacy of Translation," *Journal of Philosophy,* 67 119701, 178–83, and "On Empirically Equivalent Systems of the World," *Erkenntnis,* 9 119751, 313–28.) It is difficult to determine the truth of this strong claim without an adequate theory of explanation and of what detailed structure the explanatory relation might involve.

[8] See P. Atiyah and R. Summers, *Form and Substance in Anglo-American Law: A Comparative Study in Legal Reasoning, Legal Theory, and Legal Institutions* (Oxford: Oxford University Press, 1987).

ence of undesired or irrelevant factors such as personal preference. We want to decide or judge a particular case by considering all and only the relevant reasons concerning it. A general principle, which forces us to look at other actual and hypothetical cases, can help test whether a reason *R* we think is relevant or conclusive in this case really is so. Would *R* be relevant or conclusive in another case? If reasons are general, we can check *R*'s force in this case by considering other cases. Moreover, deciding via a general principle can call our attention to other relevant reasons, ones we have not yet noticed in this case. Looking at another case where feature *R* does *not* have great force might lead us to notice another feature *F* that the present case has, and it is *R* and *F* together which have great force. (If we hadn't looked at the other case, we might have thought *R* alone was enough.) Including all the relevant reasons might help to ensure that *only* relevant reasons are used, *if* these fill the space and so crowd out irrelevant ones. And will we really be willing to accept the impact that an irrelevant reason imposed in this case also would have upon other cases and examples? Notice that this use of hypothetical or other actual cases to test a judgment in this case already assumes that reasons are *general.* If we assume that things happen or hold for a reason (or cause) and that reasons (or causes) are general, then a general principle, perhaps defeasible, can be formulated to capture this reason, to explain why an event the scientist studies occurs or why a particular judgment about a case is correct.[9]

Principles can guide us to a correct decision or judgment in a particular case, helping us to test our judgment and to control for personal factors that might lead us astray. The wrongness that principles are to protect us against, on this view, is individualistic —the wrong judgment in *this* case —or aggregative —the wrong judgments in *these* cases which are wrong one by one.

[9] It would be interesting to investigate how far the parallel between the structural features of reasons and of causes extends and to explain why this parallelism holds. Do reasons show parallels to phenomena of probabilistic causality?

However, judgments together might have an additional wrong, a *comparative* wrong that occurs when cases that should be decided in the same way are decided differently. It has been held to be a maxim of (formal) justice that like cases should be decided alike; this general maxim leaves open which likenesses are the relevant ones.[10] Principles might function to avoid this injustice or disparity, not simply to get each and every case decided correctly by itself but to get relevantly similar cases decided similarly. But if I see films two weeks in a row, I need not decide which ones to attend on a similar basis. These two similar decisions, then, apparently do *not* count as like cases that must be decided alike. What demarcates the domain within which the maxim of formal justice is to operate? As a moviegoer, I do not see my task in deciding which movie to attend (on either occasion) as that of reaching a *just* decision on that occasion. The issue of comparative injustice arises only in contexts that involve individual justice or injustice, however these latter contexts are marked. If case *A*, calling for a decision of justice, is decided wrongly, that is bad. If now case *B*, relevantly similar, is decided differently —that is, correctly —and if that decision introduces an additional bad into the world, not the result in case *B* itself but the comparative bad of the two cases being decided differently, and this bad stands over and above the badness involved when case *A* was decided incorrectly, then *this* context of justice is a comparative one, invoking the formal maxim of justice.[11] One function of principles, then, may be to avoid this

[10] See Herbert Hart, *The Concept of Law* (Oxford: at the Clarendon Press, 1961), pp. 155–59, and Chaim Perlman, *The Idea of Justice and the Problem of Argument* (London: Routledge, Kegan Paul, 1963).

[11] I have said that a necessary condition for invoking the formal maxim of justice is that the context is one in which a just decision is to be reached, but I have not claimed this is a sufficient condition. If there are individual decisions involving justice that do not have that comparative aspect, then a further criterion is needed to mark which contexts involving justice do invoke the formal maxim. In *Anarchy, State, and Utopia* (New York: Basic Books, 1974), chapter 7, I presented a theory of distributive justice, the entitlement theory, which explicitly was not a patterned theory and did not involve comparisons among the holdings of different people. However, that is not to say that the formal maxim would not apply to people's

particular type of injustice, ensuring that like cases will be decided alike. (Whether it would be better to decide both cases wrongly, avoiding the comparative injustice, or to decide one of them correctly, avoiding injustice in that individual case but incurring the comparative injustice, presumably will depend upon particular features of the situation and the cases.)

Interpersonal Functions

A principled person can be counted upon to adhere to his principles in the face of inducements or temptations to deviate. Not necessarily in the face of any possible temptation or of extremely great inducement —nevertheless, principles are some barrier to a person's following the desires or interests of the moment. A person's principles of action thus have an interpersonal function, in reassuring others that (usually) he will get past temptations; they also have an intrapersonal function, helping the person himself to get past temptation.

Consider, first, the interpersonal function. When (refraining from) an action is mandated by a person's principles, we can count on it more. Being able to rely to some significant extent upon his behavior, we ourselves can perform actions whose good outcome is contingent upon the principled person's specific behavior. Even were the future to bring him inducements to deviate, we can trust that he will not, and we can rely upon this in planning and executing our own actions. Otherwise we would have to behave differently, for the chance would be too great that this previous behavior would come to naught or to ill. With those personally close to us, we can rely upon their affection and continuing good motivations to produce coordinate actions; with others more distant, we rely upon their principled behavior.

---

holdings arising in accordance with the *same* general principles (of justice in acquisition, transfer, and rectification). Hence, so far as that theory goes, in addition to the injustice of someone's holdings not arising through the operation of those principles, there could be an additional comparative injustice if another's holdings had so arisen (e.g., the first is discriminated against by others who do not let those principles of justice in holdings apply to him).

Such considerations are familiar in discussions of contract law. Contracts enable a person to bind himself to carry out an action, thereby encouraging another to count upon this and thus perform an action which takes her out on a limb that would be sawed off if the first person failed to perform. Since the first person benefits from that second person's action, which would not be performed if the first person had not contractually bound himself to act, this first person is willing in advance to restrict himself to so acting in this case even should his future incentives change. For if his action was left dependent upon the vagaries of future fluctuations, the second person would not perform that complementary action which the first person now wishes her to do.

Principles constitute a form of binding; we bind ourselves to act as the principles mandate. Others can depend upon this behavior, and we too can benefit from others' so depending, for the actions they thereby become willing to do can facilitate our social ease and interactions, and our own personal projects as well.[12] *Announcing* principles is a way to incur (what economists term) reputation effects, making conditions explicit so that deviations are more easily subject to detection. These effects are most pertinent for someone who makes repeated transactions with many people, assuring others that he will act a certain way (in order to avoid diminution of a reputation that serves him in interaction).[13]

These considerations can make a person want to *seem* to others to have particular principles, but why would he actually want to

[12] Principles others can count upon our following also might deter them from certain actions rather than inducing them to cooperate. A nation or person with a principle to retaliate against certain offenses, even when that is against his immediate interests, might deter others from such offenses. Announcing such a principle increases the cost of making exceptions in order to ensure that none will be made.

[13] The U.S. government wishes to issue debt and promises not to inflate, but after the debt is taken up by others it will be optimal for the government to inflate—and the others realize this beforehand. Hence the government attempts to commit to rules for managing the currency, to be followed by an agency independent of Congress, rather than leaving itself absolute discretion. See Finn Kydland and Edward Prescott, "Rules Rather Than Discretion," *Journal of Political Economy,* 85 (1977), 473–91.

have them? For most of us, possessing principles may be the most convincing and the least effortful route to seeming to have them, but fiction and real life too abound with skilled deceivers. Suppose a person does want to have a particular principle, and not merely seem to, because this will function most convincingly for others and most easily for himself. *Can* he come to have that principle merely because of its useful interpersonal functions? Mustn't he think the principle is *correct*?

And how reassuring would I find someone's telling me that he believes his holding a principle is indeed necessary to reassure me and others? "But do you hold it," I would wonder, "and how strongly?" If his attitude toward the principle was that it was a reassurance for others, even a very necessary and extremely useful reassurance, wouldn't I wonder about his continuing adherence in the face of monetary temptations or inducements to deviate? What I would want, I think, is for the person to think the principle was *correct* and *right.* Of course, it is not enough that he think this now — his belief must be stable, not subject to overturn by the slightest counterargument or counterinducement. That's what would reassure me sufficiently so that I would run risks whose good outcome was contingent upon his good behavior. And I might be proficient at detecting genuine belief that a principle was correct and be unwilling to run cooperative risks in its absence.[14]

Believing in the correctness of his principles, then, might be a useful trait for a person to have, making possible an expanded range of interactions with others and cooperative activities. This belief could be useful, even if the notion of "correct principles" made no sense at all. For this — let us for the moment suppose — senseless belief, evidenced by the person and detected by others, would be a reliable indicator to them of his future conduct and would lead them to do trusting actions that benefit him too. (Simi-

[14] It would be useful to list and compare what bases other than his accepting principles as objectively valid there might be for reliance upon a person's actions; these might include the other functions of principles which we list whose successful performance does not depend upon a belief in the objective validity of principles.

larly, the belief that certain conduct was divinely prescribed and that all deviations would meet dire punishment might be a useful belief for people to have, whether or not it was true or made any sense at all, provided it guaranteed to others their continuing conduct.) This raises the possibility of a sociobiological explanation not of particular patterns of conduct but of the belief in an objective moral order. Believing in *correctness* might be selected for.

If people are to be assured about my future conduct, it may not be enough for me simply to announce my principles; other people may need to see, upon occasion, that I actually am adhering to these principles. Yet the principles I think most correct or adequate may be difficult for others to observe in operation; those most adequate principles might respond to subtle contextual details, nuances of history or motivation or relationship not known to others or reliably checked by them. "Justice," it is said, "must not only be done but be seen to be done." Yet what if what can be dependably seen and recognized is less complex than (fully) adequate justice requires? The interpersonal function of assuring others that justice is being done or that principles are being followed might necessitate following less subtle and nuanced principles but ones whose applications (and misapplications) can sometimes be checked by others.[15]

Thus, there can be a conflict between fine-tuning a principle to a situation and producing public confidence through the principle. The more fine-tuned the principle, the less easily can its applications be checked by others; on the other hand, beyond a point of coarsening, a principle may fail to inspire confidence, not because it cannot be checked but because *its* applications no longer

[15] David Kreps, *A Course in Microeconomic Theory* (Princeton: Princeton University Press, 1990), p. 763, reports that Robert Wilson argues that publicly held accounting firms who perform external audits of businesses, in order to assure potential investors that the auditors themselves are not suborned by the firms they are auditing, must follow established rules for auditing, rules whose application can be externally checked, even if these practices do not provide the most revealing information about the business's finances. Since the application of these established rules *can* be checked, the auditing firm is able to maintain its reputation as an independent third party.

count as desirable. It has been claimed —the matter is one of some controversy— that women's moral judgments are more finely attuned to situational details and nuances of relationship and motivation than are men's.[16] This difference, *if* indeed it holds, might be due —a statistical generalization —to women's less frequently making (or anticipating making) decisions in a nonfamilial realm where the basis or motives of decision are an object of suspicion. If in some (public) realm assurance must be given to others, anyone in that realm may need to bend (somewhat) to the dictates of what *can* provide assurance, and principles are one such device. Predictions have been made about the moral changes to be effected by women's entering in large numbers into previously male arenas — a good thing for very many reasons —but it is not certain that it will be the arenas rather than the included women who will experience the greater change.

Another person's principles enable me to predict with reasonable (though perhaps not perfect) accuracy some aspects of his behavior and hence lead me to count upon those aspects. For that other person, though, his principles do not seem primarily to be predictive devices; only rarely do people attempt to *predict* their own future behavior— usually they just *decide* what to do. Rather, his principles play a role in producing that behavior; he *guides* his behavior by the principle. My knowing of his principles affects my estimate of the likelihood that he will behave a certain way, my estimate of the probability of his behaving that way; but *for him* the principles affect not (merely) *estimates* of the probabilities but these very probabilities themselves: for him the principles are not evidence of how he will behave but devices that help determine what he will (decide to) do.[17]

[16] See Carol Gilligan, *In a Different Voice* (Cambridge, Mass.: Harvard University Press, 1982); see also Bill Puka, "The Liberation of Caring: A Different Voice for Gilligan's 'Different Voice,'" *Hypatia,* 5 (1990), 58–82.

[17] Following the philosophical tradition, I use the term "determine" to mean fix, cause, make happen —as in "determinism" — but notice too the term's estimate/evidential/epistemological side, as in "I haven't yet determined what he's trying to do."

Personal Functions

It is because principles of behavior have a personal function, apart from issues of social interaction, that they are able to perform and achieve their interpersonal function. This interpersonal function —reassuring others of our behavior in the face of temptations, and hence leading them to choose to act coordinately with our actions —could not arise (as a solution in a coordination game) or be maintained without its basis in the personal matrix. What, then, are the personal and intrapersonal functions of principles and in what ways do they achieve these?

Principles may be one way a person can define her own *identity* —"I am a person with *these* principles" —and principles followed over an extended period are a way a person can integrate her life over time and give it more coherence. Some might say it is good to be principled because that is a way of being consistent. However, actions are not (logically) inconsistent in themselves or among themselves. An action *can* be inconsistent with a principle, and hence derivatively with the other actions that fit that principle. But if one wanted merely to avoid inconsistency, that could be done by having *no* principles at all. Principles do knit one's actions together, though. Through them, one's actions and one's life may have greater coherence, greater organic unity. That may be valuable in itself.

What does it mean to define oneself or one's identity in terms of principles? In that case, should we construe the self as a system of principles? These could include principles for transforming existing principles and for integrating new ones, thus for altering the self too in terms of principles. (Would a person's violating her principles then threaten to destroy her self?) However, continuing goals also would integrate a person's life and actions over time. Why define oneself by principles rather than goals? (Unlike principles, goals can be balanced, traded off against others, etc.) A person who doesn't *define* herself through principles nevertheless might *have* principles, not as an internal component

of her identity but as an external constraint upon the actions of a separate distinguishable identity. One thinks of the Kantian themes of self-creation and self-legislation, but if chosen goals can give self-creation, why is self-*legislation* needed? Does this role of principles depend upon controversial Kantian claims about what (and only what) gives rise to autonomous freedom?

These personal functions of principles concern one's life or identity as a whole, or at least extended parts of it. Principles also function for a person, more modestly, at the micro-level. One intrapersonal function of moral principles is connected to our commitment to them. In starting long-term projects there is the question of whether we will stick to them in the future, whether our — as some like to say — future selves will carry them out. Only if the answer is yes might it be worthwhile to begin a particular project, and beginning it might be rational only when we have some assurance it will continue. If my holding something *as a principle* now creates a greater cost for deviating from it in the future — that very same action would have less cost when it is no deviation from a principle — then a project that incorporates a current and longstanding principle will be one I am less likely to abandon; this is not because I have some additional principle to stick to my projects, but because this project embodies a principle I (probably) will continue to have. Just as principles have an interpersonal function of giving assurance to another — she can count on my behavior in planning hers — so too they have the intrapersonal function of enabling me to count on certain behavior from my future self — when he too probably will have that principle. Therefore, I now can reasonably undertake some projects that are only desirable contingent upon certain future behavior by me.

Within the process of a person's decision making, principles might function as an exclusionary or filtering device: in choice situations, do not consider as live options those actions that violate your principles. Principles thus would save decision-effort and

calculation time for a creature of "limited rationality." Yet the exclusion need not be absolute; if no sufficiently good action (above a certain level of aspiration) is found among the live options, a previously excluded action might be reconsidered.

Overcoming Temptation

The central intrapersonal function of principles I want to focus upon is getting us past temptations, hurdles, distractions, diversions. The psychologist George Ainslie has presented a theory of why we do impulsive behavior that we know is against our long-term interests and of what devices we use to cope with the temptations to such behavior.[18] Before turning to Ainslie's work, some background is useful.

We care less now about a future reward, economic and psychological data show, than we will later when that reward eventuates —we "discount" the future. The current utility to us of receiving a future reward is less than its utility will be when it occurs, and the more distant that reward, the less its current utility. This itself is an interesting phenomenon, and we may wonder about its rationality; in our plans and projects of action shouldn't we value a reward at all times as we would when it occurred? To be sure, we also want to take account of the uncertainty that we will survive until that time or that the reward will occur —each may be less than completely certain. In our present calculations, then, we wish to utilize an expected value, discounting that future reward's value by its probability, but shouldn't the utility of the reward's actually being received remain constant, no matter when the time?

Time preference — the term economists use for a utility-discounting of the future —may be evolution's way of instilling in creatures who could not perform such anticipatory probabilistic

[18] George Ainslie, "Specious Reward: A Behavioral Theory of Impulsiveness and Impulse Control," *Psychological Bulletin,* 82 (1975), 463–96; "Beyond Microeconomics," in Jon Elster (ed.), *The Multiple Self* (Cambridge: Cambridge University Press, 1986), pp. 133–75.

calculations a mechanism to roughly the same effect. Innate time preference may be a rule of thumb that approximates what calculations previously would have yielded, at least with regard to those rewards (and punishments) affecting inclusive fitness; there may have been selection for such time preference.[19] A problem arises, then, for beings with the cognitive apparatus to take explicit account of the uncertainties of a future reward's eventuating and to perform explicitly a probabilistic discounting of the future. If already installed in us is an innate time preference —evolution's attempt to perform the probabilistic discounting for us —and moreover what we explicitly discount in our probabilistic calculations is the (already discounted through time-preference) present value of the future reward, then what takes place will be a *double-discounting.* And surely that is too much. It seems that beings sophisticated enough to realize all this who perform expected-value calculations should utilize current estimates of what the utility of the future reward will be when it eventuates (which then are explicitly discounted by the probabilities) , rather than the time-preferenced current discounted values of those future rewards. Otherwise, they should skip the expected value calculations and stick with the evolutionarily instilled time preference.[20] However, if pure time preference is a rational phenomenon in itself, not *simply* an evolutionary surrogate for probabilistic discounting, but if such evolutionary shaping did take place, then the situation is more complicated.

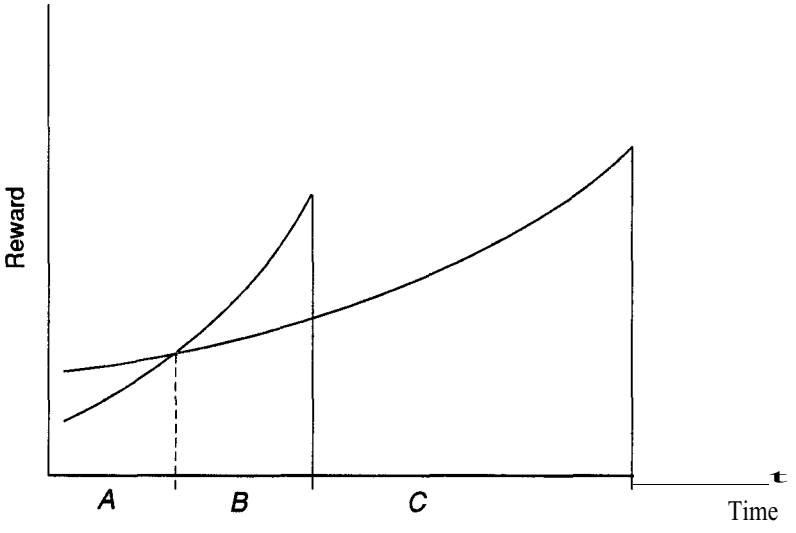
[19] Can we use information about people's current degree of time preference to make a rough estimate about the harshness and riskiness of the environment and the life history of the organisms in whom this degree of time preference first evolved? Might we use information about the general shape of the time preference curve to check theories about the domain within which selection operated (e.g., how extensive a class of kin within kin selection)?

There might be features in addition to probabilistic discounting that time preference was selected, in part, to approximate. Susan Hurley (in conversation) mentions possible change of utility due to future changes of preference.

[20] I first discussed the perils of double-discounting in "On Austrian Methodology," *Synthese,* 36 (1977), 353–92.

The curves describing the time-preferenced discounting of future rewards need not be straight lines or exponential; they may be hyperbolic.[21] Ainslie noticed that two such highly bowed curves (as the hyperbolic) can cross, and he traced out the implications of this fact, (See figure 1: the utility of a reward is measured on the y-axis; its utility for a person at a given time is measured by the height of its curve at that time. The curve slopes downward to the left because a future reward has a lesser value earlier.) Suppose there are two projects or plans of action leading to different rewards, where receiving the earlier possible reward, the smaller of the two, will thwart receiving the later larger one. A person proceeds along in time staying with the project having the highest utility at that time. In the time interval *A,* the more distant reward has the greater utility; in the time interval *B,* though,

FIGURE 1



[21] This last shape is a consequence of the "matching law" equations. See Richard Herrnstein, "Relative and Absolute Strengths of Response as a Function of Frequency of Reinforcement," *Journal of the Experimental Analysis of Behavior, 4* (1961), 267–72.

the nearer reward has the greater utility. Since the larger reward can actually be collected only at the end of the time interval *C*, the person must get through that middle period *B* without turning to the smaller reward. This presents a problem, because during that middle time interval the prospect of receiving that smaller reward *soon* has greater utility than the prospect of receiving the greater reward later.

Why assume that the person *should* try to get past that intermediate time period; why shouldn't the smaller but more immediate reward be taken?[22] What makes the two periods *A* and *C,* wherein the larger reward looms largest, the appropriate ones for deciding which choice is appropriate? During them the person will prefer acting to gain the largest reward; during period *B* she will prefer acting to gain the smaller one —that is, one that is smaller when she gains it than the other one would be when she gained *it*. Where are *we* standing when we say that avoiding the temptation is the better alternative, and why is that standpoint more appropriate than the person's standpoint within the time interval *B?*

Here is a suggestion. The time interval *B* is not the appropriate benchmark for deciding what the person ought to do because *B* is not a representative sample of her view of the matter. The time intervals *A* and C sum to a longer interval. Moreover, when we add her judgments *after* the moment the rewards are to be realized, and graph which rewards seem largest to her *then,* we find that soon after consuming the smaller reward she wishes she had not done this, but after consuming the larger reward (at the end of the time interval *C*) , she continues to prefer having chosen that larger reward. I suggest that, often, what makes resisting the temptation and taking the larger reward the preferred option is that this is the person's preference for a majority of the time: *it* is her (reasonably) stable preference, the other is her preference at

[22] I thank Amartya Sen for raising this question.

a nonrepresentative moment.[23] (Leaving aside any after-the-fact preferences, if the time interval *B* lasted for longer than the intervals *A* and *C,* would it be clear *in that case* that the temptation should be resisted?) Temptations should not always be resisted, only when the desire for the larger reward (including the preference after the fact) is the person's preference for the larger amount of time. This criterion is meant to be defeasible, not conclusive. It does have the virtue of staying close to a person's preferences (though it is not wedded to a particular local preference) in contrast to saying that it simply *is* in the person's interests to resist the temptation or that the relevant criterion is — and resisting temptation serves — the maximization of utility over a lifetime.[24]

Ainslie describes various devices for getting oneself past that intermediate period of temptation. These include taking an action during interval *A* that makes it impossible to pursue the smaller reward during *B* (e.g., Odysseus tying himself to the mast) ; taking an action during interval *A* (e.g., making a bet with another person) that adds a penalty if you take the smaller reward, thereby altering *its* utility during interval *B;* taking steps during *A* to avoid noticing or dwelling upon the virtues of the smaller reward during *B.*[25] And — our current topic — formulating a personal general principle of behavior.

A general principle of behavior groups actions; it classifies a particular act along with others — for example, "never eat snacks between meals," "never smoke another cigarette." (I do not, for present purposes, make any distinction between principles and

[23] There also is the phenomenon of *regret,* a lowering of current utility due to looking back upon currently undesired past action. Having a tendency toward regret might help one somewhat to get over the temptation during *B,* since during *B* you can anticipate the lowered utility level during *C* and also afterward if you take the smaller closer reward now. But will this anticipation feed back sufficiently into the overall utilities during *B* to affect the choice made then?

[24] For a critical discussion of the single goal of maximizing the total utility over a lifetime, see my *The Examined Life* (New York: Simon and Schuster, 1989), pp. 100–102.

[25] See also Jon Elster, *Ulysses and the Sirens* (Cambridge: Cambridge University Press, 1979).

rules.) We might try to represent the effect of this principled grouping of action within utility theory and decision theory as follows. By classifying actions together as of type *T,* and by treating them similarly, a principle links the utilities of these *T*-actions (or the utilities of their outcomes). It would be too strong to say that because of the principle all *T*-actions must have the same utility; there may be other types and principles that one particular *T*-action falls under while another *T*-action does not, so their utilities may diverge. What a principle sets up is a *correlation* between the utilities of the various actions falling under it. Stating this at the level of preference, when acts of type *T* are ranked with other actions in a preference ordering, there will be a correlation between the rank orders of the *T*-acts. However, if this correlation were the only effect that adopting or accepting principles had on the utilities of the actions falling under them, then principles would not be of help in getting us past temptations.

The mark of a principle (''never eat snacks between meals," "never smoke another cigarette") is that it ties the decision whether to do an immediate particular act (eating *this* snack, smoking *this* cigarette) to the whole class of actions of which the principle makes it part. This act now stands for the whole class. By adopting the principle, it is as if you have made the following true: if you do this one particular action in the class, you will do them all. Now the stakes are higher. Tying the utility of this act of snacking to the disutility of all those acts of snacking in the future may help you to get through the period *B* of temptation; the utility for you now of this particular snack is altered. This snack comes to stand for all the snacks, and at this early point the current utility of being thin or healthy later far outweighs the current utility of those distant pleasures of eating; the current disutility of poor health or a poor figure becomes a feature of the currently contemplated particular act of snacking.[26]

[26] In its focus upon a whole group of actions of a certain kind in the personal realm, this may remind some readers of rule utilitarianism in the public realm. How-

But why assume the person will formulate a principle during time period *A* rather than during period *B?* Why won't the person take the snack this time and formulate a principle to always snack or, more generally, a principle to always give in to immediate temptation? But formulating and accepting such a principle (alongside the action of taking the snack now) will not itself bring reward immediately or maximize reward over time. It does generally reduce delay in reward, but during period *B,* facing one particular temptation, do I want to *always* reduce delay for any and every reward? No, for though I am in that *B* period with respect to one particular reward, with regard to many other (pairs of) rewards I am in the *A* period (or the *C* period). With regard to these other more distant pairs of lesser and greater, I do not now want always to take the more immediate one, even though I do now wish to take one *particular* reward (which I am in the *B* period of) that is more immediate. It is because temptations are spread out over time that, at any *one* time, we are in more *A* (or *C)* periods than *B* periods. Hence we would not accept a principle always to succumb to temptation.[27]

By adopting a principle we make one action stand for many others and thereby we change the utility or disutility of this particular action. This alteration of utilities is due to exercising our power and ability to make one action *stand for* or *symbolize* others.

---

ever, our question is how the acceptance of a general principle affects the choice of a particular action that, in the absence of the principle, would not have maximal utility. The comparable question would be how someone with act utilitarian desires who (somehow) decides upon a rule utilitarian principle can manage to put it into effect in particular choice situations.

[27] The proponent of succumbing to temptation may reply, "You are saying that we don't *want* always to succumb to temptation. But you say a principle is the device to get us past what may be our current desire. So perhaps we need a principle to get us past the desire not to always succumb to temptation." Leaving aside the skirting of paradox, a principle is (most easily) adopted during a time period *t* when a contrary desire is stronger than the temptation is during *t*. (The temptation will reach full strength later than *t.)* And there will not be a period when the desire *always* to succumb is not weaker than a contrary desire. (Or if such a temporary period did arise, any principle adopted then soon would be overturned on the basis of a later desire that wasn't just momentary.)

Violating the principle this one time does not necessitate that we always shall violate it; having this snack does not necessitate that we shall become continual snackers. Before we adopted the principle it was not true that doing the act this one time would involve doing it always. Adopting the principle forges that connection, so that the penalty for violating the principle this time becomes the disutility of violating it always. It would be instructive to investigate *how* precisely we are able to do this.

The fact that we can, though, has important consequences. We can so alter utilities (by adopting a principle and making one act stand for others), but we cannot do this too frequently and make it stick. If we violate a particular principle we have adopted, we have no reason to think the next occasion will be any different than this one. If each occasion is the same, and we do it this time, won't we do it on such occasions always? Unless we can distinguish this occasion from the later ones, and also have reasons for believing that this distinction will carry weight with us *later* so that we won't indulge once again by formulating another distinction which again we won't adhere to still later, then doing the action this time will lead us to expect we shall continue to repeat it. (To formulate a distinction that allows this one act yet excludes future repetitions is to formulate yet another principle; we must have more reason to think we shall adhere to that one than to this, or the reformulating will give no credibility to our abstention in the future.) Doing the act this one time, in *this* situation, means we shall continue to do it in the future. Isn't this enough to alter the utility now of doing it this one time, attaching to this particular act now the disutility of all its future repetitions?

We expect that if we do it this one time, we also shall do it repeatedly in the future, but does our doing it this once actually *affect* the future; does it *make* it more likely that we repeat the action? Or does it simply affect our *estimate* of how likely that repetition is? There are two situations to consider. When no principle was adopted previously that excludes the action, doing the

action now may have a minor effect on the probability of repetition in accordance with the psychologist's "law of effect": positive reinforcement of an action raises its probability of occurrence in the future. And the estimate of the probability of repetition may be raised somewhat if this action is added to a number of similar ones in the past. When a principle was adopted previously, acting in violation of the principle will raise an observer's estimate and the agent's own estimate too of how likely she is to repeat this particular act. Also, it makes it more likely that she will. The principle has broken down, one bar to the action has been removed; moreover, realizing this may produce discouragement and make the agent less likely to exert effort to avoid the action in the future. (Notice that an action that affects *her estimate* of the probability of similar future actions may then produce discouragement and thereby affect the actual probability of repetition.) Formulating a principle that would constitute an additional bar to the actions it excludes is a way of actually tying the effects of all to the effects of any (previous) one. The more one has invested in a principle, the more effort previously put behind adhering to it, the greater the cost in violating it now. (For how likely is it that you will continue to adhere to another one if you couldn't manage to stick to this one despite so much effort?) Moreover, adhering to the principle this time is a type of action subject to the law of effect; its being positively reinforced makes it more probable that adherence to that principle will occur in the future.

The effects of violating a principle may be more general still, for the probability or credibility of your successfully utilizing *any* principles at all in *any* arena (when faced with a temptation as strong as the one which caused you to succumb this time) may be affected. To be sure, you may try to demarcate and limit the damage to this *one area* but this presents the same problem — one level up — as limiting the damage *within* this area to just *this one* violative action. Deontological principles may have the greatest weight when their violation directly threatens *any* and all prin-

cipled action in the future: if I violate *this* principle (in this circumstance), how can I believe I will succeed in adhering to any (desirable) principle ever again? Someone might try, in an excess of Kantian zeal, to increase the potential effect of spreading disaster by formulating a (meta-) principle never to violate any principle. But while getting any violation to stand for all might lessen the probability of any given one, the actual consequences of the slightest violation would get dangerously magnified. This is not to say that one violation of a principle, because one act stands for all, discharges a principle so one then can violate it freely and with impunity. One act has the disutility of all, but then so does the next, even if that first act was done. This disutility can be escaped by dropping the principle, not by violating it; however, one then faces the very disutility that adopting the principle was designed to avoid.

Since adopting a principle itself .is an action that affects the probability linkages among other actions, some care is appropriate in choosing which principles to adopt. One must consider not only the possible benefits of adherence, but the probability of its violation and what future effects that violation would have. It might be better to adopt a less good principle (when followed) but one easier to adhere to, especially since that principle may not always be available as a credible fall-back if one fails to adhere to the more stringent one. (Also, one wants to adopt a principle sharp enough to clearly mark its violations, so one's future self cannot easily fudge the issue of whether the principle is being followed.) No doubt, a theory of the optimal choice of principles could be formulated, taking such considerations into account.[28]

A principle speaks of all the actions in a group and it makes each present act stand for all. To perform its functions, it must

[28] The promulgation of a principle also affects how third parties will carry it out; a designer of principles will take account of how others might distort or abuse them. For a related point about how social theorists such as Marx and Freud should have taken precautions against vulgarization, see my *The Examined Life,* p. 284.

speak of *all* the actions of a certain kind. We do not have principles that say: most *P*s should be *Q*s; or 15% of *P*s have to be *Q*s. Sometimes, though, all we need is to do something some or most of the time (e.g., skipping desserts most evenings, paying most of our bills each month). The way we achieve this through principles is nevertheless to formulate a statement that speaks of "all," "each," or "every," yet is coextensive with the mix we desire. *Each* month, pay most of your bills; *every* week, skip desserts most evenings. A teacher —not myself —whose principle it is not to give very many *A*s grades *every* class on a curve. Thereby, each week or month or class comes to stand for all. Thus, we can explain why principles concern all the members of a class, not just some, (A norm could concern itself with *n%,* where *n* is not 0 or 100, but a principle cannot.) A principle has certain functions, and to perform these one instance must stand for or symbolize all. The observed "all"-character of principles thus provides support for our view of the functions principles have and the ways they perform them.[29]

Principles may seem crude devices for accomplishing our goals; their universal coverage —giving up *all* desserts, *all* diversions until the task is done —may be more than is necessary to reach the goal. The leeway in what the "all" covers (desserts, weeks) mitigates this somewhat, narrowing the overkill of principles. Still, some will remain. If there were a clear threshold of *n* repetitions of an action, past which the consequences of continuing that action thwart the goal but before which the goal still can be reached, wouldn't a rational person perform the action precisely *n* times and then stop? (A more complicated statement is needed if each repetition increases the difficulty of reaching the goal.) No principle would be needed to exclude the n + 1th action, since that action itself would have bad consequences on balance. This might

[29] An alternative explanation of principles incorporating "all" might propose that principles codify reasons and that reasons are universal (though defeasible); hence principles are too. But why is it that reasons are not "for the most part" but instead are "universal but defeasible," even though the percentages may be the same?

be a theory of (approximately) when the person decides to stop smoking (or gaining weight, etc.), and hence of when she decides to institute a principle. Yet, given temptation, it is a principle that needs to be instituted *then*.

Sunk Costs

One method Ainslie mentions for getting past the tempting time interval *B* is this: *commit* yourself during the earlier interval *A* to seeking the larger reward during *C* and during *B*. One mode of such commitment is, during *A*, to invest many resources in the (future) pursuit of that larger reward. If I think it would be good for me to see many plays or attend many concerts this year, and I know that when the evening of the performance arrives I will frequently not feel like rousing myself at that moment to go out, then I can buy tickets to many of these events in advance, even though I know that tickets still will be available at the box office on the evening of the performance. Since I will not want to waste the tickets I have bought, to waste the money already spent on them, I will attend more performances than I would if I left the decisions about attendance to each evening. True, I may not use *all* of these tickets—lethargy may triumph on some evenings—yet I will attend more frequently than if no tickets had been purchased in advance. Knowing all this, I purchase the tickets in advance in order to drive myself to attend.

Economists present a doctrine that all decision making should pay attention only to the (present and) future consequences of various alternative actions. The costs of past investments in these courses of action already have been incurred. While existing resources may affect the consequences of the various actions now open before me —already possessing the ticket I can attend the performance making no additional future payment —and hence be taken into account through these consequences, the mere fact that costs already have been borne to further a certain project should not carry any weight at all as a person makes a decision.

These costs, "sunk costs" as the economists term them, are a thing of the past; what matters now is only the future stream of benefits. Thus, sitting at home this evening, if I now would prefer staying home to going out and attending a performance (for no monetary payment), then the evening at home has higher utility for me than traveling out and attending the performance; therefore I should stay at home. It should make no difference that I already have spent money on the ticket for the performance — so runs the economists' doctrine that sunk costs should be ignored.[30]

This may be a correct rule for the maximization of monetary profits, but it is not an appropriate general principle of decision, for familiar reasons. We do *not* treat our past commitments to others as of no account except insofar as they affect our future returns, as when breaking a commitment may affect others' trust in us and hence our ability to achieve other future benefits; and we do *not* treat the past effort we have devoted to ongoing projects of work or of life as of no account (except insofar as this makes their continuance more likely to bring benefits than other freshly started projects would). Such projects help define our sense of ourselves and of our life.[31]

The particular issue we have been discussing indicates yet another defect in the doctrine of ignoring sunk costs as a general principle of decision. The fact that we do not ignore sunk costs provides one way to get past the temptation during the *B* time interval to choose the smaller but more immediate reward. Earlier, during the time interval *A* when we can clearly see the benefits of the larger but more distant reward, we can sink resources and effort

[30] People frequently do not adhere to the doctrine of ignoring sunk costs, as indicated by their decisions when presented with hypothetical choices. On this, see H. R. Arkes and C. Blumer, "The Psychology of Sunk Cost," *Organizational Behavior and Human Decision Processer,* 35 (1985), 124–40. Arkes and Blumer see the people who deviate from the doctrine in the ticket-example as being irrational.

[31] See the Bernard Williams essay in J. J. C. Smart and Bernard Williams, *Utilitarianism: For and Against* (Cambridge: Cambridge University Press, 1973); "Persons, Character and Morality," in Amelie Rorty (ed.), *The Identities of Persons* (Berkeley: University of California Press, 1976).

into achieving that reward, knowing that when the time *B* of temptation comes, the fact that we do *not* want (and will not want) to have wasted those resources will count for us as a reason against choosing the smaller reward, adding to its disutility. If I know I will be tempted some evening in the future by the smaller immediate reward of comfort (not having to go out into the rain, etc.), yet I also know that now and afterward too I will be happy to have attended all those performances, then I can buy the tickets now, in advance, to spur myself to forgo staying home when that evening arrives.

Everyone sees succumbing to the smaller reward during the time interval *B* as a problem, an irrationality or an undesirable short-sightedness. The person herself sees it that way —beforehand and later, if not right then —and we see it thus too as we think about it. The economist also sees another type of behavior, the honoring of sunk costs, as irrational and undesirable. But we now see that this latter behavior, anticipated in advance, can be used to limit and check the first type of undesirable behavior (viz., succumbing to the smaller but nearer reward). We can knowingly utilize our tendency to take sunk costs seriously as a means of increasing our *future* rewards. If this tendency is irrational, it can be rationally utilized to check and overcome another irrationality. If someone offered us a pill that henceforth would make us people who *never* honored sunk costs, we might be ill-advised to accept it; this would deprive us of a valuable tool for getting past temptations of the (future) moment. (Might such a tendency to honor sunk costs, which can be adaptive, have been selected for in the evolutionary process?) Since taking sunk costs into account sometimes is desirable (so the economists' general condemnation is mistaken), and sometimes is not, the desirability of taking such a pill would depend upon the comparative numbers of, and stakes within, these two types of situations.

Earlier, I mentioned that the more effort one has put behind adherence to a principle designed to get past temptations of the

moment, the greater is the cost in violating it now. It is unlikely that you will manage to stick to another principle if you could not stick to this one despite so much previous effort. Realizing this gives you much reason to hold onto this one — it's the one life-raft in sight — and therefore gives great weight to not violating it in the face of this particular temptation. Groupings of action (in order to avoid immediate temptation) that we have succeeded in following thereby gain a further tenacity. Notice that this involves a sunk cost phenomenon. My reasoning behind sticking to *this* principle, and its associated grouping, involved saying that, if I could not stick to it despite so much previous effort, how could I hope to stick to another? It is only if I *am* someone who honors sunk costs that I will be able to make this argument; only one who thus honors sunk costs would have a reason to adhere now to this current principle for bypassing temptation, rather than succumbing this one time and then formulating a different principle, which too will succumb when its time comes, perhaps on its very first test. It is sunk costs that makes *this* principle the place to take a stand. (Do not argue that these are future-regarding considerations about the future consequences of the two different courses of action — sticking to the present policy vs. succumbing to the temptation and then formulating a new policy — and hence that the person who does not honor sunk costs can go through the same line of reasoning; it is only because of the known tendency to honor sunk costs that one course of action will have, and can be seen to have, significantly different consequences than the other. Otherwise, why think it is less likely that I will adhere to the new principle after violating the old one than that I will continue to adhere to the old principle if I don't violate it now?) Might the known phenomenon of our honoring sunk costs play some role in why we adhere to principles we have just adopted? We now know that if we can manage to adhere to this principle for some time, the fact that we will have invested in it will provide us in the future, as

honorers of sunk costs, with reasons to continue to adhere to that principle then —and that may give us some reason now.[22]

To these functions performed by our honoring sunk costs, the economist might reply that, for an otherwise perfectly rational person, honoring sunk costs is not desirable at all; only someone with some *other* irrationality should indulge in it. However, this is not so evident, even leaving aside what was mentioned earlier: commitments made to other persons and past investment in our projects of work and life. For it might be interpersonally useful to have a means of convincing others that we shall stick to projects or aims even in the face of threats that seem to make this adherence work to our future disadvantage —as a way of discouraging their making such threats or carrying them out.[33] This might be useful even if you have no other tendency to irrational behavior, and the others you are trying to convince have none either.[34] However, the theme of countering or fencing in one irrationality with another is worth marking. Can some other things that we think irrational —perhaps weakness of will, self-deception, or fallacies of reasoning —consciously be put to use to thwart or limit still other irrationalities or otherwise undesirable happenings? (And could a total package of such somewhat counterbalancing apparent irrational tendencies even work together better than the *total* package of apparent —when separately considered —rational tendencies?)

[32] I owe this suggestion to Susan Hurley, who also asks, in reference to and in parallel to our earlier question about whether we can rely upon someone adhering to a principle when his only reasons for holding it are the benefits to him of our so relying, whether someone can expect to honor costs he has sunk if he will not think he earlier had some independent reason to sink them, a reason other than to get himself to honor them later.

[33] See Thomas Schelling, "The Art of Commitment," in his *Arms and Influence* (New Haven: Yale University Press, 1966), pp. 35–91. See also Schelling's discussion of "the rationality of irrationality."

[34] It also might be a useful trait, especially for the young, to be optimistic about the chances of success of possible projects — otherwise no new and daring things would be tried — yet also to tend to stick to ongoing projects in which significant investment has been made, for otherwise at the first serious difficulty one might turn to another untried project one is still (overly) optimistic about.

Let me mention one other technique a person might use to carry herself over that tempting time interval *B* where the smaller reward looms so large. She might consider what action she would recommend to another person in that very situation, someone whose well-being she cares about —a child or a friend, for example —and then adopt that advice for herself. Distancing oneself from the situation, looking at the diagram impersonally instead of simply looking ahead from one time point, might be a way to defuse the allure of the otherwise nearer (but ultimately smaller) reward. This procedure requires an ability to look at a situation you are in impersonally and to think the same principle of choice should apply to yourself as to others, that you should do the very same action another should do in that situation. A strong predisposition to such an impartial attitude would be extremely useful in surmounting the *B* interval of the crossed curves, hence in maximizing a person's total reward. And this very disposition constitutes one component of ethical judgment: applying the same principles to one's own behavior as to others.

There is one function of principles I have not yet mentioned: *drawing the line.* Principles mark a boundary beyond which we will not step —"this is where I draw the line!" —and we think, "If I don't draw it here, where *will* I draw it?" There may be no other obvious place in a gradient of situations, no obvious place within acceptable territory. (Or there may be another acceptable place, but we feel we will not *succeed* in drawing the line there.) This is connected to the earlier mentioned function of principles, getting one past the temptation of the moment, but in this case it is not temptation but rather the *reasoning* of the moment that needs getting past. If I reach *that* point, I will reason that there is no special reason to stop just then, so I had better stop much beforehand, where there *is* a clear line and a *special* one.[35]

[35] Thomas Schelling's theory of coordination games might usefully utilize this notion of specialness. In attempting to coordinate with another, I am searching for an action we both will think is special (yet also desirable), and both will realize we

This, I think, is what enables principles to define a person. *"These* are the lines I have drawn." It is these lines that limn/ delineate him. They are his outer boundaries. A person in very fortunate circumstances, then, who knows he won't actually get taken very far along any undesirable gradient, may not *have to* draw any specific lines. Thus, in this sense he may not be as well defined as someone in less fortunate circumstances.

Symbolic Utility

We have said that by adopting the principle, doing the particular short-sighted action this one time in this situation now *means* we shall continue to do it in the future. This act *stands for* all the others the principle also excludes; doing this one *symbolizes* doing the rest. Is this fact of *meaning, standing for,* and *symbolizing* constituted by the intertwining of the two strands of connection between doing the act now and repeating it in the future that we already have discussed: the way doing it now affects your estimate of the probability of doing it again and the way doing it now alters the very probability of doing it in the future? Or is symbolizing a further fact, not exhausted by these two strands but one that itself affects the utility of alternative actions and outcomes? Symbolizing, I believe, is a further important strand, one that an adequate decision theory must treat explicitly.

Freudian theory explains the occurrence or persistence of neurotic actions or symptoms in terms of the symbolic meaning of these actions or symptoms. Producing evident bad consequences and apparently irrational, these actions and symptoms have an unobvious symbolic significance; they symbolize something else, call it M. Yet the mere having of such symbolic meaning alone cannot explain the occurrence or persistence of an action or symptom. We have to add that what these actions and symptoms symbolize —

---

both think it special —not simply striking but special. When there are ten alternatives, nine of them extremely striking, the special one might be the one that isn't — at least at the first level —striking at all.

that is, *M* — itself has some utility or value (or, in the case of avoidance, disutility or negative value) for the person, and moreover that this utility of the *M* which is symbolized is imputed back to the action or symptom, thereby giving *it* greater utility than it appeared to have. Only thus can it explain why it was chosen or manifested. Freudian theory must hold not only that actions and outcomes can symbolize still further events for a person, but that they can draw upon themselves the emotional meaning (and utility values) of these other events. Having a symbolic meaning, the actions are treated as having the utility of what they symbolically mean; a neurotic symptom is adhered to with a tenacity appropriate to what it stands for. (I am not aware of a clear statement in the Freudian literature of this equation or of the weaker claim that *some* of the utility of what is symbolized is imputed back to the symbol, despite some such version's being presupposed, I believe, in some Freudian explanations.) Disproportionate emotional responses to an actual event or occasion may indicate their standing for other events or occasions to which the emotions are more suite.[36]

For the symbolic action to get done, *it* must somehow come to have a higher utility than the other actions available to the agent.[37] I have suggested it happens this way: the action (or one of its outcomes) symbolizes a certain situation, and the utility of this symbolized situation is imputed back, through the symbolic connection, to the action itself. Notice that standard decision theory also believes in an imputation back of utility, along a (probabilistic) causal connection. By virtue of producing a particular situation for sure, an action comes to have, to have imputed to it, the utility

[36] Once an action or outcome comes to symbolize others, its presence may get taken as evidence for the others or as causes of them, but this is a result of the symbolizing, and not its original fabric (although this evidential or causal role may then reinforce the strength of the symbolic connection).

[37] So a maximizing decision theory would assume. There are other forms of normative decision theory, such as Herbert Simon's "satisficing" theory, but this too would require the action that is done to have, or have imputed to it, a utility above the (shifting) level of aspiration.

of that situation; by virtue of probabilistically producing certain situations, an action comes to have, to have imputed to it, their utilities in the form of an expected utility. What the current view adds is that utility can flow back, be imputed back, not only along causal connections but along symbolic ones.

One mark that it is an action's symbolic connection to an outcome that plays a central role in the decision to do it, rather than the apparently causal connection —I am thinking of cases where the agent does not think the action is itself intrinsically desirable or valuable —is the persistence of the action in the face of strong evidence that it does not actually have the presumed causal consequence; sometimes a person will even refuse to look at or countenance this evidence or other evidence about harmful consequences of the action or policy. (On these grounds, one might claim that certain antidrug enforcement measures *symbolize* reducing the amount of drug use and that minimum wage laws *symbolize* helping the poor.) A reformer who wishes to avoid such harmful consequences may find it necessary to propose another policy (without such consequences) that equally effectively symbolizes acting toward or reaching the goal; simply halting the current action would deprive people of its symbolic utility, something they are unwilling to let happen.

Of course, *according* a particular symbolic meaning to an action A has causal consequences of its own, as it affects which actions we perform, and a purely consequentialist theory can say something about that. It can speak of whether giving such symbolic meaning (or, later, refraining from extinguishing that symbolic meaning) is itself a causally optimal action. However, this will be different than a purely consequentialist (nonsymbolic) theory of the action A itself, and it does not imply that we must assess the according or tolerating of symbolic meaning solely by its causal consequences.

Since symbolic actions often are *expressive* actions, another view of them would be this: the symbolic connection of an action

to a situation enables the action to be expressive of some attitude, belief, value, emotion, or whatever. Expressiveness, not utility, is what flows back. What flows back along the symbolic connection to the action is (the possibility of) expressing some particular attitude, belief, value, emotion (etc.). Expressing this then has high utility for the person, and so she performs the symbolic action.[38]

There may not seem to be much difference between these two ways of structuring our understanding of a symbolic action's being chosen. Each will give a different explanation of why a symbolic act is not done. For the first, wherein utility is imputed back to the action along the symbolizing connection, this presents a puzzle. Presumably the symbolizing connection always holds, so that an action of handwashing always symbolizes removing guilt or whatever. Since this situation symbolized, being guilt-free, presumably always has high utility, if utility is imputed back, why won't the action of handwashing always have maximal utility, so that the person will always be doing it? (Apparently, this does happen with some compulsive hand-washers, but not with all, and not with all actions done because of their symbolic meaning.) The expressiveness theory says the possibility of expressing some attitude toward being guilt-free is always present, as a result of the ever-present symbolic connection, but the utility of expressing this may vary from context to context, depending upon how recently one has expressed it, what one's other needs and desires are, and so forth. The utility of expressing that attitude or emotion competes with other utilities. The utility imputation theory will describe this differently. The absolute or relative utility of the symbolized situation can fluctuate for the person; the utility of being guilt-free can actually become less if the person has recently taken steps to alleviate guilt —there now (temporarily) is less to deal with; or the utility of being guilt-free can remain constant while

[38] Not that it need always be expressiveness that flows back along the symbolic connection. Perhaps other things may, and these will give rise to new characteristics of the action which themselves have high utility for the agent. The point is that utility is not what flows back.

the utility of other competing goods, such as eating, temporarily rises to become greater than the utility of removing guilt. Each of these structures for understanding symbolic expressiveness will have some utility fluctuate —a slightly different one. What I want to emphasize now is the *importance* of this symbolic meaning, however it is precisely structured.

When utility is imputed to an action or outcome in accordance with its symbolic meaning— that is, when the utility of an action or outcome is equated with the utility of what it symbolically means —we are apt to think this irrational. When this symbolic meaning involves repressed childhood desires and fears, or certain current unconscious ones, this may well result in behavior doomed to be frustrating, unsatisfying, or tormenting. Yet mightn't symbolic meanings based upon unconscious desires also add gratifying reverberations to consciously desired goods? In any case, not all symbolic meanings will be rooted in Freudian material. Many of these others too, however, will look strange to someone outside that network of meanings: recall the dire consequences some people bear in order to avoid "losing face," the deaths people risked and sometimes met in duels to "maintain honor" or in exploits to "prove manhood." Yet we should not too quickly conclude that it would be better to live without any symbolic meanings at all or better never to impute utilities in accordance with symbolic meanings.

Ethical principles codify how to behave toward others in a way that is appropriate to their value and to our fellow-feelings with them. Holding and following ethical principles, in addition to the particular purposes this serves, also has a symbolic meaning for us. Treating people (and value in general) with respect and responsiveness puts us "on the side of" that value, perhaps allying us with everything else on its side, and symbolizes our intertwining with this. (Does it symbolize this to a greater extent than it actually intertwines us, or does a welcomed symbolic connection constitute an actual intertwining?) Kant felt that in acting morally

we act as a member of the kingdom of ends, a free and rational legislator. The moral action doesn't *cause* us to become a (permanent) member of that kingdom — it is what we would do as a member, it is an instance of what would be done under such circumstances, and hence it symbolizes doing it under those circumstances. The moral acts get grouped with other possible events and actions and come to stand for and mean them. Thereby being ethical acquires a symbolic utility commensurate with the utility these other things it stands for actually have. (This depends, then, upon these further things actually having utility for the person — a contingency Kant would be loath to rely upon.) There are a variety of things an ethical action might symbolically mean to someone: being a rational creature that gives itself laws; being a law-making member of a kingdom of ends; being an equal source and recognizer of worth and personality; being a rational, disinterested, unselfish person; being caring; living in accordance with nature; responding to what is valuable; recognizing someone else as a creature of God. The utility of these grand things, symbolically expressed and instantiated by the action, becomes incorporated into that action's (symbolic) utility. Thus, these symbolic meanings become part of one's reason for acting ethically. Being ethical is among our most effective ways of symbolizing (a connection to) what we value most highly.

A large part of the richness of our lives consists in symbolic meanings and their expression, the symbolic meanings our culture attributes to things or the ones we ourselves bestow.[39] It is unclear, in any case, what it would be to live without any symbolic meanings, to have no part of the magnitude of our desires depend upon such meanings. What then would we desire? Simply ma-

[39] Notice that symbolic meanings might not all be good ones, just as desires or preferences might not be either. The point is that a theory of rationality need not *exclude* symbolic meanings. However, these do not guarantee good or desirable content. For that, one would need to develop a theory of which symbolic meanings and which preferences and desires were admissible, using that to constrain which particular meanings and desires could be fed into the more formal theory of rationality.

teria1 comfort, physical security, and sensual pleasure? And would no part of how much we desired these be due to the way they might symbolize maternal love and caring? Simply wealth and power? And would no part of how much we desired this be due to the way these might symbolize release from childhood dependdence or success in competition with a parent, and no part either be due to the symbolic meanings of what wealth and power might bring? Simply the innate unconditioned reinforcers evolution has instilled and installed in us, and other things only insofar as they are effective means to these? These had served to make our ancestors more effective progenitors or protectors of related genes. Should we choose this as our only purpose? And if we valued it highly, might we not value also whatever symbolized being an effective progenitor? "No, not if that conflicted with actually being one, and in any case one should value only actually bearing or protecting progeny and relatives, and the effective means to this that evolution has marked out, namely, the unconditioned reinforcers, and also the means to *these.*" (Notice, though, that evolution's having instilled desires that serve to maximize inclusive fitness does not mean that it has instilled the desire to be maximally inclusively fit. Males now are not, I presume, beating at the doors of artificial insemination clinics in order to become sperm donors, even though that would serve to increase their inclusive fitness.) But why is actually leading to something so much better than symbolizing it that symbolization shouldn't count at all? "Because that's the bottom line, what actually occurs; all the rest is talk." But why is this bottom line better than all other lines?

In any case, if we are symbolic creatures —and anthropology attests to the universal nature of this trait —then presumably evolution made us so; therefore the attractive pleasures of symbolization, and symbolic satisfactions too, are as solidly based as the other innate reinforcers. Perhaps a capacity for symbolization served to strengthen other desires or to maintain them through periods of deprivation in reinforcement by their actual objects.

Whatever the evolutionary explanation, though, this capacity, like other cognitive capacities, is not mired in its original adaptive function; it can be employed in other valuable ways, just as mathematical capacities can be employed to explore abstract number theory and theories of infinity, although this was not the function for which they were evolutionarily selected. Once the capacity for symbolic utility exists, it may enable us, for example, to achieve in some sense —that is, symbolically —what is causally or conceptually impossible, thereby gaining utility from that, and also enable us to separate good features from bad ones they actually are linked with, gaining only the former through something that symbolizes only them.

This is not to deny the dangers opened by symbolic meanings and symbolic utilities. Conflicts may quickly come to involve symbolic meanings that, by escalating the importance of the issues, induce violence. The dangers to be specially avoided concern situation where the causal consequences of an action are extremely negative yet the positive symbolic meaning is so great that the action is done nevertheless. (Recall the examples of compulsive handwashing and drug prohibition.) A rational person would seek an (almost) equally satisfying symbolic alternative that does not have such dire actual consequences. (However, this does not imply that symbolic meanings always should be subordinate to, and come lexicographically after, causally produced outcomes.) Sometimes a symbolic connection will be thought better than a causal one; if an outcome —such as harming someone in revenge —is desired but seen as bad, it may be better for a person to achieve this symbolically than to inflict actual damage.[40] It would be nice to discover a general structural criterion about the kinds of links that establish symbolic meanings that can distinguish the good symbolic meanings from the bad, but perhaps we must simply be vigilant in

[40] So should we distinguish cases where the goal is x and someone acts symbolically to achieve x from cases where the goal is a symbolic connection to x and someone acts instrumentally to achieve that?

certain kinds of situations —conflict is one —to isolate and exclude particular symbolic meanings. It may help that many undesirable symbolic meanings are not in equilibrium under knowledge of their causes; if we knew what gave rise to these meanings, or the role they are playing in our current actions, we would not want to act upon them.[41] Some symbolic meanings do withstand these tests, though (e.g., the symbolic meaning of a romantic gesture to the person you love). Perhaps the crucial thing is to stay aware of when meanings and connections are symbolic ones, keeping separate track of these and not treating them (unknowingly) as causally real. This would help with the many Freudian symbolic meanings which, when they enter into conscious deliberation as symbolic, lose their power and impact.[42] (Years ago, this might have helped with those people who devoted their lives to the pursuit of wealth as a "status symbol" but now, in the United States, we find people who knowingly and openly pursue status. Or might they be pursuing status as a wealth symbol?)

Symbolic meaning also is a component of particular ethical decisions. It has been argued that the symbolic meaning of efforts to save a known currently threatened person —a trapped miner, for instance— or of refusing to make those efforts affects our decision in allocating resources to current efforts to save versus accident-prevention measures. (This issue has been termed one of "actual vs. statistical lives".)[43] It also has been argued that the symbolic meaning of feeding someone, giving sustenance, enters into the discussion of the ways in which the lives of direly ill people permissibly may be terminated —turning off their artificial

[41] For a discussion of acts in equilibrium, see my *Philosophical Explanations* (Cambridge, Mass.: Harvard University Press, 1981), pp. 348–52.

[42] I thank Bernard Williams for mentioning this example. Williams also points out that some symbolic meanings involve a fantasy that is strictly impossible to realize; and it is unclear how utilities are to be assigned to impossible situations. I would not want to preclude, however, that even incoherent situations might have high utility for us.

[43] See Charles Fried, *An Anatomy of Values* (Cambridge, Mass.: Harvard University Press, 1970), pp. 207–18.

respirator but not halting their food and starving them to death.[44] The political philosophy presented in *Anarchy, State, and Utopia* ignored the importance to us of joint and official serious symbolic statement and expression of our social ties and concern and hence (I have written) is inadequate.[45]

We live in a rich symbolic world, partly cultural and partly of our own individual creation, and we thereby escape or expand the limits of our situations, not simply through fantasies but in actions, with the meanings these have. We impute to actions and events utilities coordinate with what they symbolize, and we strive to realize (or avoid) them as we would strive for what they stand for.[46] A broader decision theory is needed, then, to incorporate such symbolic connections and to detail the new structuring these introduce.

Among social scientists, anthropologists have paid the most attention to the symbolic meanings of actions, rituals, and cultural forms and practices and their importance in the ongoing life of a group.[47] So elaborate is their work that it is somewhat embarrassing to introduce a relatively crude and undifferentiated notion of symbolic meaning. Still, this notion has its uses, not served by nuanced and textured discussions that do not easily connect with formal structures. By incorporating an action's symbolic meaning, its symbolic utility, into (normative) decision theory, we might link theories of rational choice more closely to anthropology's concerns. There are two directions in which such a linkage might go. The first, the upward direction, explains social patterns and struc-

[44] See Ronald Carson, "The Symbolic Significance of Giving to Eat and Drink," in Joanne Lynn (ed.), *By No Extraordinary Means: The Choice to Forgo Life-sustaining Food and Water* (Bloomington: Indiana University Press, 1986), pp. 84–88.

[45] See my *The Examined Life,* pp. 286–92.

[46] For a discussion of how some advertising of products utilizes this phenomenon, see my *The Examined Life,* pp. 121–22.

[47] See Raymond Firth, *Symbols: Public and Private* (New York: Cornell University Press, 1973); Clifford Geertz, "Deep Play: Notes on the Balinese Cockfight," in his *The Interpretation of Cultures* (New York: Basic Books, 1973).

tures in terms of individual choice behavior that incorporates symbolic utility. This, the methodological individualist and reductionist direction, is not the one I am proposing here.[48] The second, the downward direction, explains how the patterns of social meanings anthropologists delineate have an impact within the actions and behavior of individuals, that is, through their decisions which give some weight to symbolic utility. (Some anthropologists, as a matter of professional pride, seem not to be concerned with how the cultural meanings they delineate are mediated in individual behavior.)

How does the symbolic utility of an action (or of an outcome) work? What is the nature of the symbolic connection or chain of connections? And in what way does utility, or the possibility of expressiveness, flow through this chain from the situations symbolized to the actions (or outcomes) that do the symbolizing? Notice first that symbolic meaning goes beyond the way in which the adoption of principles makes some actions stand for others. There, an action stood for other things of the same type — other actions — or for a whole group of these, while symbolic meaning can connect an action with things other than (a group of) actions — for instance, with being a certain sort of person, with the realization of a certain state of affairs.

Some useful and suggestive categories have been provided by Nelson Goodman.[49] According to Goodman, *A denotes B* when *A* refers to *B; A exemplifies P* when *A* refers to *P* and *A* is an instance of *P,* that is, is denoted by *P* (either literally or metaphori-

[48] Indeed, given the extent to which symbolic meaning is socially created, maintained, and coordinated, as well as limited by social factors, we might find here a limit to methodological individualist explanations — an important one given the effects and consequences of such meanings. For a symbolic utility might be social not only in being socially shaped, and in being shared, that is, the same for many people in the society, but also in being viewed *as* shared — that being intrinsic to its having that symbolic utility. It is not clear how methodologically individualist explanations might cope with the intricacies involved, In any case, it is not clear what a methodologically individualist account of language would look like.

[49] Nelson Goodman, *Languages of Art* (Indianapolis: Bobbs-Merrill, 1968), pp. 45–95.

cally) ; *A expresses P* when *A* refers to *P* and *A* has the property *P* figuratively or metaphorically (so that *P* figuratively denotes *A*), and *A* functions as an aesthetic symbol in exemplifying *P*. These relations can be chained together. *A alludes to B* when *A* denotes some C and that *C* exemplifies *B*, or when *A* exemplifies some *C* and that *C* denotes *B*. Even longer chains are possible,[50] some of whose links will be figurative or metaphorical. These chains, and others, can connect an action to further and larger situations or conditions, the ones it can symbolically represent or allude to (etc.), and the utility of these larger situations then provides the action itself with a *symbolic utility* that enters into decisions about it. These chains need not be very long; when *A* is in the literal extension of a term *P* and *B* is in that term's metaphorical extension, *A* might have *B* as part of its symbolic meaning. Sometimes an action may symbolically mean something by being our best instantiated realization of that thing, the best we can do.[51]

In what particular way is the symbolic utility (or expressiveness) of an action determined by the utility of that larger situation the chain connects the action to, and by the nature of the chain itself? Do shorter chains transmit more utility/expressiveness from the larger situation to the action itself; is utility/expressiveness lost, the more linkages there are; do different kinds of linkages transmit differing proportions of (or possibilities of expressing) the larger situation's utility? (I am assuming that the symbolic utility of an action cannot be greater than the utility of the larger situation it is connected to by the chain and that it can be less.) Do only some symbolic connections induce the imputation of utility back, and what determines which ones these are? These questions all arise about situations of choice under certainty;

[50] Catherine Elgin, *With Reference to Reference* (Indianapolis: Hackett Publ. Co., 1983), p. 143, discusses a particular chain with five links.

[51] Can the symbolic utility of an action be viewed as an *interpretation* of that action, a way of seeing oneself or it a certain way, so that the various modes of interpretive linkage, and full theories of interpretation itself, might enter into the specification of symbolic utility?

further issues arise about choice under risk or uncertainty. Is there a probabilistic discounting along some particular chains; do some kinds of larger situations, even when they are not certain to occur, transmit their full utility back to the action which might yield them? And, of course, the very fact that an action has particular risks or uncertainties associated with it may itself give it a particular symbolic meaning and utility, perhaps connected with being a daring and courageous person or a foolhardy one. Sometimes, though, the presence of probabilities rather than certainty may remove a symbolic meaning altogether. It is *not* the case that a half or a one-tenth chance of realizing a certain goal always itself has half or one-tenth the symbolic utility of that goal itself—it need not symbolize that goal, even partially. Here is another reason why symbolic utilities must be treated as a separate component of a theory of decision and not simply incorporated within existing (causal and evidential) decision theories. For such symbolic utilities do not obey an expected value formula. We might attempt to understand and explain *certain* of the observed deviations from an expected value formula and from the associated axioms of decision theory, by attributing these to the presence of symbolic utilities. I have in mind here the Allais paradox, the certainty effect, certain deviations from Savage's Sure Thing principle, and so forth. There is a symbolic utility to us of *certainty* itself. The difference between 0.9 and 1.0 is greater than that between 0.8 and 0.9, though this difference between differences disappears when each is embedded in larger otherwise identical probabilistic gambles—this disappearance marks the difference as symbolic.[52] A

[52] Double-digit inflation has the symbolic meaning of inflation out of control, so there is more concern about a rise from 9% to 10% than from 16% to 17%; if we counted in base eleven the (symbolic) line would be fixed elsewhere. In *Anarchy, State, and Utopia,* I commented on the symbolic meaning of *eliminating* a problem completely, so that there is a greater difference between reducing the number of instances of an evil from one to zero than there is in reducing the number from two to one. There I referred to this as a mark of an ideologue (p. 266); it is better seen as a mark of symbolic meaning.

Notice that the certainty effect, when it occurs, requires measuring utility by a slightly different procedure than the usual one. In the usual procedure, two out-

detailed theory of symbolic utility awaits development. What we can do now is mark a place for it within the structure of a more general theory of decision, a place I shall say more about in the next lecture.

Teleological Devices

Principles help you to discover the truth, by transmitting evidential support or probability from some cases to others. Principles also help you to overcome temptation by transmitting utility from some actions to others. Principles are transmission devices for probability and for utility.[53]

---

comes x and z are assigned utility numbers ordered in accordance with the preference between them, and the utility of any third thing y is found in accordance with the archimedean condition. This condition says that when x is preferred to y and y is preferred to z, then there is a unique probability p (between zero and one exclusive) such that the person is indifferent between y for sure and an option consisting of a probability p of x and a probability $(1\text{-}p)$ of z. When the person is fully satisfying all the Von Neumann–Morgenstern conditions there will be no problem, but when the certainty effect occurs, that intermediate certain option y will be assigned a misleading utility. A better procedure might be to measure utility without considering any certain outcomes, by embedding all of the preceding within canonical probability mixtures, for instance, with probability 1/2 The person then would be asked to find the probability p such that he is indifferent between a 1/2 chance of nothing and a 1/2 chance of y, and a 1/2 chance of nothing and a 1/2 chance of (a probability p of x and a probability $1\text{-}p$ of z). Thereby we control for the certainty effect. Of course, such a procedure can work only if it is not sensitive to the particular probability, in this example 1/2, within the canonical probability mixture. It would have to be the case that the same results would be gotten with a wide variety of probabilities within the canonical mixture, perhaps with all but those within epsilon of 0 and 1.

[53] Must all principles transmit only one or the other of these, or can some principles transmit both? Should we speculate that there is *one* thing which all principles transmit, namely $p_i \ X \ u_i$, probability *and* utility? There is no single term within decision theory to denote this weighted sum, $p_i \ X \ u_i$, despite their very frequent travel together *as a unit*. Indeed, formal theories have to institute very particular procedures to disentangle them, procedures that frequently assume they have been successfully disentangled in specific cases and then utilize devices to extend this to situations in general. We might learn something interesting by treating probability and utility as part of one integrated quantity —call it importance —and not separating these components too soon, by investigating what conditions this integrated quantity satisfies. (But isn't there an asymmetry at the beginning between the components, in that importance can be embedded in *probability* mixtures? Do we need to investigate the corresponding possibilities of utility mixtures, which may magnify or diminish the constituent importances? And might a temporal factor be included in the combination to begin with, only later to be abstracted out as a com-

Principles have various functions and effects: intellectual, intrapersonal, personal, and interpersonal. This is not to say they have these effects in every possible situation. A temperature regulatory mechanism will work only within a certain range of temperature; beyond that range it will not be able to bring temperature back and, depending upon its material, it may even itself melt or freeze. Why didn't evolution give us better regulatory mechanisms for body temperature? Given the small probability of such extreme cases' arising, that would be too costly in terms of energy and attendant sacrifice in other functions. A mechanism can perform its function pretty well, well enough, even if it won't work for some of the situations that might arise—similarly for principles.

In order to justify a principle, you specify its functions, and show that it effectively performs that function, and does this more effectively than others would given the costs, constraints, and so forth. We also can ask about the desirability of that function. Why should *anything* do that? A justification will show (or assume) that the function is desirable and does not interfere with other more desirable functions. Fully specified, a justification of a principle *P* is a decision-theoretic structure, with the principle *P* occupying the place of an action, competing with specific alternatives, having certain probabilities of reaching certain goals with certain desirabilities, and so on. (Our earlier discussion of factors that would be considered by a theory of the optimal choice of principles would fit into this decision-theoretic, teleological framework.)

A principle can be designed to cope with certain situations or to protect against *particular* dangers, such as giving in to temptations of the moment, favoring one's own interests, believing what one wants to be true. Hence, someone who doesn't face those dangers might not have need for *those* principles. And there might be devices other than principles to cope with such dangers.

---

ponent? Is time-preference primarily a matter concerning probability or utility, or does temporal distance itself constitute a diminution in importance? Does the extension of a utility in time —not its displacement in time — magnify its importance?)

(Might a person cope with favoring her own interests not only through principles, but through empathic interaction with others and imaginative, full projection into their situations?)

We might ask whether the device of general principles itself has its *own* biases or defects. Putting things in terms of decision theory enables us to see principles as devices (that are supposed) to have certain effects — their functions — and hence not only to compare some principles with others, but also to compare principles with other devices. Some goals might be impossible or very difficult for principles to reach, while other means might reach *those* goals more easily.

If one important goal is living together without a conflict so intense that it tears apart and destroys valuable social institutions, then when contending parties strongly put forward incompatible principles there may be no way to resolve that conflict by getting the parties to agree to any third principle, much less to either of the original two. What may be needed is some compromise — but compromise is just what principles are not supposed to do! Hence a leader of an institution or a country may simply try to keep things going, to work out some arrangement to damp down people's fury so that institutional life can continue. To be sure, there may be a principle that recommends doing this, a principle to be applied to all situations of serious principled conflict that threatens to rend and make dysfunctional valuable institutions. However, the particular content of the compromise may simply be determined by what the contending forces, given their respective powers, can manage to live with. That compromise need not itself be determined by principle in the sense that its details are taken to set a precedent for other similar situations. This is not to recommend that political and institutional leaders be unprincipled. Perhaps they are to be principled in their decisions and actions unless in those rare situations where the above-stated principle mandating (unprincipled) compromise comes into effect. (However, looking at the structure of the United States government, there

seems to be a different division: some types of decision, those made by the judiciary, are held to require principles, while the details of other decisions, those of the chief executive and legislature, generally are left to the play of various forces, with some oversight by the judiciary to ensure that certain general principles are not violated.) The only point I wish to make here is that the teleological device of principles may not be suited to each and every purpose.

Another reason for thinking that principles of action have a teleological function is this. An actual case, for instance Nazi Germany, may thoroughly refute a principle *P* that would countenance or allow that. But why wasn't the hypothetical example enough? In 1911 couldn't one say: principle *P* would allow or even in certain circumstances would require (something like) Nazi Germany. Therefore, *P* is false, unacceptable, evil.

However, if principles are only supposed to cover the cases that will, would, and could arise, then before the fact, if it is thought such a case is impossible (that the situation, motivations, etc., that would lead to it couldn't arise or succeed it might not be considered a *relevant* counterexample to that or any principle. But once it is discovered that human nature *can* do that —because it *did* —then the principle *P* which countenances it is refuted.

The consequences of people accepting and acting upon a principle can discredit the principle. "They acted on the principle *P,* and look at the horrendous situation to which that led." Someone else might say that they took the principle too far or took it in a wrong direction — that the principle itself didn't *require* what they did. Nevertheless, the principle *P* is discredited. When everyone is revolted by the earlier consequences of following *P,* it is difficult for someone to say, "Let's follow *P* again, but this time in the right way." Why? Is it because *P* so easily *lent itself* to that way of acting, even when it didn't require it? That's what accepting *P* leads to when people like people actually are follow

it.[54] If a principle is a device for having certain effects, it is a device for having those effects *when it is followed,* so what actually happens when it is followed, not just what it *says,* is relevant in assessing that principle as a teleological device.

But aren't principles also basic truths, which aid our understanding by subsuming instances (à la Hempel) and hence explaining why they hold? Here, again, principles might be considered to be as devices with an *epistemological* function (viz., to produce understanding), and so even here we can ask (decision-theoretically) whether there are other routes to understanding, whether these others are better suited for some contexts or subjects, and so forth.

But mightn't principles be what makes the particular truths true, what gives rise to them —in which case the primacy of principles would be *ontological?* If this does not simply repeat the epistemological function —we understand the particular truths best through principles —and if "giving rise to" is not a temporal relation, and if "makes it true" is not a causal relation, then it is not clear exactly what the ontological thesis claims. Still, this last would not make of principles solely a teleological device, and in any case we need not deny that the formulation of principles (of mathematics, of natural phenomena, of psychology) can bring coherence to these phenomena and depth to our understanding, whether the relation between the phenomena and the principles be ontological, epistemological, or some mixture. Hence, there is a further intellectual function to principles other than the one we began with —the transmission of support and probability — namely, to deepen and unify and make explicit our understanding of what the principles concern. (This will produce tighter relations of support and probability; might these *constitute* rather than result from the increased understanding?) The formulation

[54] See my *The Examined Life,* sec. on "The Ideal and the Actual." This also opens the possibility that people who don't want *P* to be followed to a certain result could arrange to have *P* followed to another monstrous result, thereby discrediting it.

of moral principles, thus, could deepen our understanding of moral action or moral facts and phenomena. Here, though, moral principles would have no different a status than physical or psychological ones that describe phenomena but which there is no evident reason to act *on*. It might be said that although correct moral principles hold true —in that they *ought* to be followed —the only way to get them realized, that is, to be true of our actual behavior, is to try to follow them, to act *on them*. This is an empirical claim, one that would require evidence. Perhaps the principles we are able to formulate and follow are so far off what correct moral principles —more complex moral truths — would require that we would better conform to the latter by following routes other than trying to act on principle. It is, after all, an empirical question. In any case, that makes acting on principle, once again, a teleological device.

The Kantian tradition tends to hold that principles function to guide the deliberation and action of self-conscious reflective creatures; hence principles have a theoretical and a practical function. We are creatures who do not act automatically, without any guidance. We could imagine having automatic guidance — would that make principles completely otiose for us? —or, more to the point, acting in a way that doesn't utilize guidance, for instance, at random. (Would acting completely at random suffice to free us from the domain of causality, the function Kant reserves for principles?) Doesn't this show that the purpose of principles is to guide us to something, whatever that is, that we wouldn't reach by acting at random? And doesn't that leave principles as teleological devices? However, Kant also would hold that principles are an expression of our rational nature, constitutive of rationality. To think or act rationally just is to conform to (certain kinds of) principles. Hence it would be a mistake to look only for the extrinsic *functions* that principles serve. If principles are something only a rational agent can formulate and utilize, and if being rational is something we value, then following principles can sym-

bolize and express our rationality. Principles thus might have high utility for us, not because of what their use leads to, but because of what it symbolizes and expresses. To that extent, principles would not be solely teleological devices. But there would remain the question of why we would so value our rational nature, and the acting on principles and reasons which expresses this, if our rational nature serves no further purpose. Why does the buck stop there?

Why are principles so intimately connected with rationality? And why do we value rationality? To speak of something, an action or belief, as rational is to assess the reasons for which it was done or held (and also the way in which the person took account of the reasons *against* doing or believing that). If reasons are, by their nature, general, and if principles capture the notion of acting *for* such general reasons —so that the person is committed to acting thus in other relevantly similar circumstances also —then to act or think rationally you must do so in accordance with principles. But why should we believe or act rationally? One answer would be that we *are* rational, we have the capacity to act rationally, and we value what we are.[55] But if we are to step beyond simple self-praise, mustn't we invoke the functions served by believing or acting rationally? And why must reasons be general? Compare them with their most similar nongeneral relatives. To explain why we should utilize reasons rather than these alternatives, we must again invoke the functions of reasons. Thus, the question turns from one about principles to one about rationality. What are reasons for? What is the function of rationality? Is rationality itself wholly teleological, wholly instrumental?

[55] For some critical reflections on the view that we are free when our actions are determined self-consciously by a law of reason, which is a principle constitutive of our essential nature, see my *Philosophical Explanations,* pp. 353–55.

II. Decision-Value

Newcomb's Problem

Newcomb's Problem is well known and I shall just describe it briefly here.[56] A being in whose power to correctly predict your choices you have great confidence is going to predict your choice in the following situation. There are two boxes, B1 and B2; box B1 contains $1,000 and box B2 contains either $1,000,000 ($M) or nothing. You have a choice between two actions: (1) taking what is in both boxes; (2) taking only what is in the second box. Furthermore, you know and the being knows you know (etc.) that if the being predicts you will take what is in both boxes, he does not put the $M in the second box; if the being predicts you will take only what is in the second box he does put the $M in the second box. First the being makes his prediction, then he puts the $M in the second box or not, according to his prediction, then you make your choice.

The problem is not only to decide what to do, but also to understand precisely what is wrong with one of the two powerful arguments that conflict. The first argument is this: if you take what is in both boxes, the being almost certainly will have predicted this and will not have put the $M in the second box and so you will almost certainly get only $1,000, whereas if you take only what is in the second box, the being almost certainly will have predicted that and will have put the $M into the second box and so you will almost certainly get $M. Therefore, you should take only what is in the second box. The second argument is this: the being already has made his prediction and has already either put the $M into the second box, or has not. The $M is either already sitting in the second box, or it is not, and which situation obtains is already

[56] The problem was thought of by William Newcomb, a physicist, told to me by a mutual friend, and (with Newcomb's permission) first presented and discussed in Robert Nozick, "Newcomb's Problem and Two Principles of Choice," in N. Rescher et al. (eds.), *Essays in Honor of C. G. Hempel* (Dordrecht, Holland: Reidl, 1969), pp. 114–46.

fixed and determined. If the being has already put the $M in the second box, then if you take what is in both boxes you will get $M + 1,000, whereas if you take only what is in the second box you will get just $M; if the being has not put the $M in the second box, then if you take what is in both boxes you will get $1,000, whereas if you take only what is in the second box you will get no money at all. In either case, whether the $M has been placed in there or not, you will receive more money, $1,000 more, by taking what is in both boxes. (Taking what is in both boxes, as it is said, *dominates* taking only what is in the second.) Therefore, you should take what is in both boxes.

Since 1969 when I first presented and discussed this problem, there has been much detailed investigation and illuminating theorizing about it.[57] In my initial essay, I distinguished those conditional probabilities that mark an action's *influencing* or *affecting* which state obtains from mere conditional probabilities that mark no such influence, and I suggested that when it conflicts with the dominance principle the principle of maximizing conditional expected utility should not be invoked if its conditional probabilities were of the second (nonaffecting, noninfluencing) sort. I supported this by intuitive examples. (These, because of an attempt to incorporate a certain reflexivity, are somewhat more complicated than examples others discussed afterward.) Linked genetic predispositions to a disease and to a career choice should not, I argued, lead someone to avoid one career since this raises the estimate of her chances of getting the disease — whether she actually does have that genetic makeup or will actually get the disease is not *influenced* or *affected* by the career choice. It did not occur to me to utilize this theme for the full and systematic development of competing versions of decision theory, causal and evidential,

[57] For a selection of articles until 1985, and a bibliographical listing of others, see Richmond Campbell and Lanning Sowden, *Paradoxes of Rationality and Cooperation: Prisoner's Dilemma and Newcomb's Problem* (Vancouver: University of British Columbia Press, 1985).

with their differing versions of the expected utility principle and even their differing versions of the dominance principle.[58]

The traditional principle of maximizing expected utility treats the expected utility of an action *A*, EU(*A*), as the weighted sum of the utilities of its (exclusive) possible outcomes, weighted by their probabilities which sum to 1.

$$\mathrm{EU}(A) = \mathrm{prob}(0\mathrm{l}) \times \mathrm{u}(0\mathrm{1}) + \mathrm{prob}(0\mathrm{2}) \times \mathrm{u}(0\mathrm{2}) + \ldots + \mathrm{prob}(0n) \times u(0n), = \mathrm{SUM}\ (i=1, \ldots, n)\ \mathrm{prob}(0i) \times U(0\mathrm{i}).$$

A more adequate principle, noticing that the outcomes need not be probabilistically independent of the actions, specifies the expected utility as weighted not by the simple probabilities of the outcomes, but by the conditional probabilities of the outcomes given the actions—call this the evidentially expected utility of *A*, EEU(*A*).[59]

$$\mathrm{EEU}(A) = \mathrm{prob}(O\mathrm{l}/A) \times u(0\mathrm{1}) + \mathrm{prob}(O\mathrm{2}/A) \times u(0\mathrm{2}) + \ldots + \mathrm{prob}(On/A) \times u(On), = \mathrm{SUM}\ (i=1, \ldots, n)\ \mathrm{prob}(Oi/A) \times U(0i).$$

[58] On causal decision theory, see Allan Gibbard and William Harper, "Counterfactuals and Two Kinds of Expected Utility," in Hooker, Leach, and McClennen (eds.), *Foundations and Applications of Decision Theory* (Dordrecht, Holland: Reidl, 1978), pp. 125–62; David Lewis, "Causal Decision Theory," *Australasian Journal of Philosophy,* 59 (1981), 5–30; J. H. Sobel, "Circumstances and Dominance in a Causal Decision Theory," *Synthese,* 63 (1985).

Nor did I notice the possibility of specific situations where the states were probabilistically independent of the actions yet causally influenced by them—Gibbard and Harper's Reoboam example—which should have marked a fourth row in the three-rowed chart on p. 132 of my original article.

[59] On the maximization of conditionally expected utility, though not the term "evidential utility," see my 1963 Princeton University doctoral dissertation, *The Normative Theory of Individual Choice* (since published: New York: Garland Press, 1990). See p. 232: "The probabilities that are to be used in determining the expected utility of an action must now be the conditional probabilities of the states given that the action is done. (This is true generally. However when the states are probability-independent of the actions, the conditional probability of each state given that one of the actions is done will be equal to the probability of the state, so the latter may be used.)'' There also the formula for conditional expected utility was stated for the cases of the two particular actions being discussed there, though not the general formula for variable action. The general formula is presented in Richard Jeffrey, *The Logic of Decision* (New York: McGraw Hill, 1965).

The issues that concern us in this book all arise when the probabilities, conditional or otherwise, subjective or objective, are sharply defined. Other issues have led some to formulate theories using probability intervals (see, for example, Isaac Levi, *Hard Choices* [Cambridge: Cambridge University Press, 1986]); how exactly the views stated here might be restated within such frameworks is a question for investigation.

The causal decision theorists too use not simply the unconditional probability of the outcome but a probability relating the outcome to the action, this time not simply the conditional probability, prob $(0i/A)$, but some causal-probabilistic relation indicating direct causal influence; the corresponding formula with these causal probabilities states the causally expected utility of act A, CEU(A).

Despite these and other technical elaborations —backtracking subjunctives, explicit incorporation of tickles and meta-tickles, the ratifiability of decisions, and so forth —and despite attempts to show the problem is irremediably ill-defined or incoherent[60] —the controversy continues unabated. No resolution has been completely convincing.

Newcomb's Problem is a complicated one, other cases involve still further complications, the reasoning seems quite compelling on *all* sides —and we are fallible creatures. It would be unreasonable to place absolute confidence in any one particular line of reasoning for such cases, in any one particular principle of decision.[61]

The amount in the first box, the $1,000, has received little attention.[62] If the dominance argument —the second argument

[60] Attempts to reject the problem as ill-formed, ill-defined, or impossible in principle include Isaac Levi, "Newcomb's Many Problems," *Theory and Decision,* 6 (1975), 161–75; J. L. Mackie, "Newcomb's Paradox and the Direction of Causation," *Canadian Journal of Philosophy,* 7 (1977), 213–25; William Talbott, "Standard and Non-standard Newcomb Problems," *Synthese,* 70 (1987), 415–58. For a defense of the problem against many such criticisms, see Jordan Howard Sobel, "Newcomblike Problems," *Midwest Studies in Philosophy,* 15 (1990), 224–55.

[61] Some years ago, in a graduate seminar several students, particularly David Cope, queried how anyone could be certain either of causal or of evidential decision theory, given the strong arguments on both sides. I am grateful for this discussion, for it set me along the following train of thought. (However, Howard Sobel writes me to say that things are not symmetrical, for it is only the causal theorists who have tried not only to produce arguments on their own side but to diagnose the [purported] errors of the opposing arguments, in line with the desideratum I proposed in my original article.)

[62] An exception is J. Howard Sobel, who in "Infallible Predictors," *Philosophical Review,* 92 (1988), 3–24, closes the paper by considering "a limit Newcomb Problem" in which the amount in the first box is increasing from $1,000 to (almost) $1 million. However, Sobel does not also consider the situation of reducing

above — is correct, then you will be better off taking what is in both boxes even when the amount of money in the first box is much smaller, $1 for example, or even one cent or a 1/10,000th chance of one cent. However, few of us would choose both boxes in such a case, granting *no* force to the other argument that if we take only what is in the second box we are almost certain to get $M. On the other hand, if the first argument above is correct and is understood as an expected utility argument (with the embedded conditional probabilities not needing to express any influence), then the amount of money X in the first box could be much larger than $1,000 yet the person would still choose to take only what is in the second box. Let us assume that the probability of the being correctly predicting your action (for each choice you might make) is .99. Where u denotes the utility function, the expected utility of taking only what is in the second box is $.99u(\$M)$, while the expected utility of taking what is in both boxes is $.99u(X) + .01u(\$M+X)$. If we suppose that the utility of money is linear with its amount in this range, then this expected utility of taking what is in both boxes is $u(X) + .01u(\$M)$. In this case, the expected utility of taking what is only in the second box will be greater than the expected utility of taking what is in both boxes if $.99u(\$M)$ is greater than $u(X) + .01u(\$M)$ — that is, if $.98u(\$M)$ is greater than $u(X)$. On the assumption that utility is linear with amount of money, then, the person will choose to take only what is in the second box whenever the amount in the first box is less than $980,000. So, for example, in a choice problem having the same structure as Newcomb's Problem but where the first box contains $979,000 (and the second box, as before, contains $M or nothing), the person would not take the contents of both boxes but only what is in the second box. No doubt, the utility of money is not

---

the $1,000 in the first box to almost nothing. In Kenneth MacCrimmon and Stig Larsson, "Utility Theory: Axioms versus 'Paradoxes,"' in Maurice Allais and Ole Hagen (eds,), *Expected Utility Hypothesis and the Allais Paradox* (Dordrecht, Holland: Reidl, 1979), p. 393, the consequences of varying the amount in the second box, though not in the first, are considered.

linear with its amount in this range, but this is no great distortion for our purposes —it is the utility of $M + X$ that will be proportionally less than its monetary amount. The general point holds nevertheless: for very large amounts of money in the first box, $900,000 for example, provided the being is highly accurate in his predictions, a proponent of the first argument would take only what is in the second box. Few of us, however, would feel comfortable following the first argument in this case, granting *no* force to the other argument that we are better off in either case taking what is in both boxes.

By varying the amount of money in the first box we can make people extremely uncomfortable with their otherwise favored argument for choice in Newcomb's initial problem. People who initially chose both boxes are unwilling to follow the dominance argument when the amount in the first box is lowered to $1; people who initially chose only the second box are unwilling to follow the expected utility argument (with conditional probabilities that do not mark influence) when the amount in the first box is raised to $900,000. This suggests that no one has *complete* confidence in the argument he or she follows for Newcomb's initial example — no one is willing unreservedly and across the board to apply the reasoning that seems to move him or her in that case.

A person might have differing amounts of confidence in various principles of decision (and their associated arguments). For the moment we can restrict ourselves to just the two principles of maximizing (conditionally) expected utility, as these are formulated by causal decision theory and by evidential decision theory. These differing amounts of confidence might be represented by degrees of confidence between zero and one inclusive that sum to one; or by degrees that do not sum to one, leaving open the possibility that both of the principles are incorrect for a given case; or by confidence-weightings that are not degrees between zero and one. For some particular person, let Wc be the weight he or she gives to the expected utility principle of causal decision theory,

and let We be the weight he or she gives to the expected utility principle of evidential decision theory. Let CEU(A) be the causally expected utility of act A, the utility of that act as it would be computed in accordance with (some favored one of the versions of) causal decision theory; let EEU(A) be the evidentially expected utility of act A, the utility of that act as it would be computed in accordance with evidential decision theory. Associated with each act will be a decision-value DV, a weighted value of its causally expected utility and its evidentially expected utility, as weighted by that person's confidence in being guided by each of these two kinds of expected utility.

$$DV(A) = Wc \times \mathrm{CEU}(A) + We \times \mathrm{EEU}(A).$$

And the person is to choose an act with maximal decision-value.[63]

I suggest that we go further and say not merely that we are uncertain about which *one* of these two principles, CEU and EEU, is (all by itself) correct, but that both of these two principles are legitimate, and each must be given its respective due. The weights, then, are not measures of uncertainty but measures of the legitimate force of each principle. We thus have a *normative* theory which directs a person to choose an act with maximal decision-value.

A maximizer of decision-value, if he gives nonzero weights to Wc and We will be led to shift his choice in Newcomb's Problem: from one box to two when the amount in the first box is raised sufficiently; from two boxes to one, when the amount in the first box is lowered sufficiently. Such changes are predictable for maximizers of decision-value. (Thus, the theory of maximizing DV has testable, qualitative, behavioral consequences, at least for those who conform to that normative theory.)

[63] If less than complete confidence in one principle leads to following a combination of them, what happens if one does not have complete confidence in this combination? If there is a determinate other principle that one has *some* confidence in, then, insofar as the argument depends only upon actual degrees of confidence, it seems that other principle should also be included in the weighting.

There are many different mathematical structures that would give CEU and EEU a role, but the *DV* formula is especially simple and it would be premature to look now at anything more complicated. The weighted *DV* structure, all by itself, of course, does not give anyone much guidance. How great should the weights be? Must a person use the same weights in all decision situations, or might the weights vary for different types of decision situation, or more systematically according to where a decision situation falls along some dimension D—the further to the left the more plausible the use of one of the decision criteria (and hence the greater weight it receives), the further to the right the more plausible the use of the other one? I would welcome a theory to specify or restrict the weights, just as I would welcome a theory to specify or restrict prior probabilities within a Bayesian structure and one to specify or restrict the substantive content of preferences within the usual ordering axioms. Still, in each case the general structure can be illuminating.

That some weight is to be given to both factors, CEU and EEU, means that EEU will receive some weight even in decisions about cases where there is no causal influence of the act upon the relevant outcome — for example, the cases where a choice of a career indicates (but does not affect) differing probabilities of catching or already having a terrible disease. In my original article I thought it absurd to give such considerations any weight. Yet I knew that the evidential component of the DV formula has had major social consequences in human history, as the literature on Calvinism and the role its view of *signs* (though not causes) of election played in the development of capitalism attests. (It can be a causal consequence of an action that a person believes something that act indicates but does not cause and is made happy by this belief. But someone who introduces this as a reason for doing the action must take care not to countenance such happy consequences as a reason for holding the belief.)[64]

[64] For a divergent view of evidentialist considerations, holding that these are appealing only when they match cooperative reasoning in interpersonal situations,

Theorists of rationality have been intent upon formulating the one correct and complete principle to be applied unreservedly in all decision situations. But they have not yet reached this —at any rate we do not have complete confidence that they have. In this situation, won't a prudent and rational individual hedge her bets? I want to say more: namely, that no one of the principles alone is wholly adequate —it's not simply that we haven't yet found the knockdown argument for the one that is correct. I do not say that the framework of decision-value alone will bring theorists to agree. They will continue to differ in the weights they assign to the specific decision principles, even were they to agree about which principles should be included. It is this disagreement about weights that explains the differing choices in Newcomb's Problem, but it is the fact that we do give *weights* (rather than sole allegiance to one principle) that explains the switching of the decision as the amount in the first box is varied. The *DV* structure represents the fact that each of EEU and CEU captures legitimate reasons (of a sort), and we do not want to dismiss completely either sort.[65]

It is somewhat strange that writers on decision theory generally have shown such confidence in their views. For if we formulate the issue about the correct principle of decision as a decision problem, one about which principle of decision should be followed[66] — we might imagine that pills have been developed that can trans-

---

see Susan Hurley, "Newcomb's Problem, Prisoners' Dilemma, and Collective Action," *Synthese,* 86 (1991), 173–96.

[65] "But what explains the disagreement between proponents of CEU and EEU? Is it a factual or a value disagreement?" This questions assumes both proponents share an EU formula and asks whether their disagreement resides within the probability or the utility component. Yet if the *DV* formula is correct, there are *other* things to disagree about, including the weights *Wc* and *We,* the nature of the formula, and also —to anticipate the next paragraphs —the inclusion of other factors. To ask "fact or value?" —allowing no other alternative —is to assume that what *must* be in common is the simple EU framework and that only *within* it can disagreement arise.

[66] David Gauthier considers the question of what disposition of choice a person should choose to have in *Morals by Agreement* (Oxford: Oxford University Press, 1985), chapter 6, secs. 2–3.

form us into consistent followers of each principle —then it is not obvious what the contending principles of decision will answer, and in particular it is not obvious whether they each will put itself forward as the preferred alternative. That depends upon what the world will be like. If it will offer many situations like Newcomb's Problem itself, with significant payoffs, then taking the EEU-pill can be predicted to have better *causal* consequences, so that the CEU principle will recommend taking the EEU-pill rather than the CEU-pill. If on the other hand the world will offer very many significant situations with the structure of my disease example ("Newcomb's Problem and Two Principles of Choice," p. 125) or the many similar ones (Gibbard and Harper's Solomon example, etc.: see note 58 above), then the person following the EEU principle (without any "tickle" addition) will frequently forgo significant benefits (because of the misfortunes it portends) ; since this can be predicted, the EEU principle itself will recommend taking the CEU-pill as an act that has higher EEU utility than does taking the EEU-pill. (In this case, the CEU principle also recommends taking the CEU-pill. Is there an example where the EEU of taking the CEU-pill is higher — and so the EEU principle recommends that —although the CEU of taking the pill is not higher, and so the CEU principle does not recommend taking it? Difficulties then would abound.) Just as there is no one particular inductive policy, no one Carnapian c-function, that is best or most effective no matter what the character of the world actually is, so there may be no one best principle of rational decision.[67] And just as we want our inductive procedures to allow for learning, to contain parameters that get specified through some experience of the world, so too we want our principles of rational decision to contain param-

[67] Rudolf Carnap maintained *(Logical Foundations of Probability* [Chicago: University of Chicago Press, 1950]) that sentences asserting that "the degree of confirmation of h on e is n" are, when true, analytic. Yet even he also held that *which* particular confirmation function is to be chosen (c*, c-dagger, or whatever from among the continuum of inductive methods), and therefore which one will specify this analytic relation, is a matter of pragmatic choice and will depend upon general facts about the universe.

eters that can be specified to fit the discovered character of the world in which decisions are to be made. (In each case, evolution may have accomplished a significant part of the setting of parameters to fit the actual world, but this does not mean we should expect our specific inductive policies or decision principles to be applicable in every imaginable science-fiction situation, or that we should treat them as valid *a priori.*) The framework of decision-values with its incorporated weights that can be altered over time is one way a fitting to the actual world can be accomplished.

The decision-value we specified was based upon the two components CEU and EEU, but any alternative plausible principle of decision, or factor in decision, might be added as a term with its own associated weight. In particular, we could add to the formula the symbolic utility of an action, its SU, which incorporates the utility of the various outcomes and actions *symbolized* by the act, with its own associated weight Ws. (It is best not to try to incorporate symbolic utility alongside other utilities, because it may well not obey an expected value formula and because we might want to keep separate track of symbolic utility, since we think it appropriate to give this factor different weight in different kinds of choice situations.) The formula for the decision-value of A, $DV(A)$, then would become:

$$DV(A) = Wc \times \mathrm{CEU}(A) + We \times \mathrm{EEU}(A) + Ws \times \mathrm{SU}(A) .$$

It would be instructive to investigate the formal characteristics of this decision-value structure; it would not be surprising if this principle of weighted combination, like other criteria previously investigated in the literature of decision under uncertainty, sometimes failed to exhibit certain desirable features.[68]

[68] See John Milnor, "Games against Nature," in R. M. Thrall, C. H. Coombs, and R. L. Davis, *Decision Processes* (New York: John Wiley, 1954), pp. 49–59, and R. D. Luce and Howard Raiffa, *Games and Decisions* (New York: John Wiley and Sons, 1957), pp. 275–98. Earlier I said that symbolic meaning need not carry over proportionally into probabilistic contexts. Yet the DV formula includes symbolic utility as one of the weighted components. We might wonder whether sym-

Symbolic utility is not a different kind of utility, standing to standard utility in something like the way metaphorical meaning stands to literal. Rather, symbolic utility is a different kind of *connection* —symbolic—to the familiar kind of utility. It stands alongside the already familiar connections, the causal and the evidential. The symbolic utility of an action A is determined by A's having symbolic connections to outcomes (and perhaps to other actions) which themselves have the standard kind of utility, just as the CEU of A is determined by A's causal-probabilistic connections to outcomes with the standard utility.[69]

Should we ensure that these types of connection —causal, evidential, and symbolic —are exclusive? The earlier formula for DV specified it as a weighted sum of CEU and EEU. However, the EEU of an action includes its causal components, since the conditional probabilities of outcomes given actions, prob(O/A), which the EEU theorist utilizes incorporate causal influences when such exist. In our weighted sum formula, then, should we not interpret the EEU as the expected utility represented by those (portions of) probabilities which are *not* (simply derivative from) causal ones? And similarly, shouldn't the symbolic utility SU of an action be its symbolic utility which is not (simply) derivative from and represented within those very causal and evidential connections?[70]

---

bolic utility will carry over into the weighted DV context. However, shifting to a probabilistic situation is a shift to a *different* situation, while shifting to the DV formula does not shift the choice situation.

[69] One further condition therefore needs to be imposed on the standard situation for measuring utility discussed in footnote 52. That situation must be one where the actions have no relevant evidential or symbolic connections to utility-outcomes. Utility is to be calibrated in causal contexts, where an expected value principle is followed, and the (sanitized) utilities thus found are to be utilized in situations where actions also stand in evidential and symbolic connections to valued outcomes. However, the value of these latter outcomes is measured in situations that are wholly causal.

[70] The one psychological study I know of that treats both causal and evidential connections, and seeks to disentangle them, is G. A. Quattrone and Amos Tversky, "Causal versus Diagnostic Contingencies: On Self-Deception and the Voter's Illusion," *Journal of Personality and Social Psychology,* 46 (1984), 237–48.

Should we incorporate still further components into the *DV* structure? One suggestion would be explicitly to include a component concerning the way an action fits into a person's image of himself and is self-expressive. However, our three components already cover much of this territory. Although performing an action of the sort that would be done by a certain kind of person may not *cause* the agent to be this kind of person, it may symbolize his being that way, be some evidence that he is, and have the causal consequence of making it easier for him to maintain an *image* of himself as being of that kind. This last is a real causal consequence of an action which may have significant utility. Hence this kind of consequence —how doing a particular action affects the person's self-image —can play a significant explicit role in an explanatory theory of that person's behavior, even though it is a type of consequence which the agent himself cannot easily take into account explicitly —"I am going to do *A* in order to make it easier to maintain my self-image as a person of kind *K*" —without thereby diminishing that very effect.[71] However, we should not interpret expressiveness as exhausted by these other independent categories, narrowly conceived, for, as we said earlier, the categories of the symbolic and the expressive are intertwined.

If we array the category of the linguistic alongside the categories we already have as follows, causal/evidential/linguistic/symbolic, this suggests two questions. How does the symbolic differ in nature from the merely linguistic (the symbolic does lend itself more to utility being imputed back, but this need not happen every

[71] Not every mode of action involves a connection to a consequence to be placed within a formula alongside the causal, evidential, and symbolic connections. Consider acting without motive, a mode whose variants are spoken of in the literature of Buddhism, Taoism, and Hinduism. Here, the person does not act so as to become a certain way, or to be a certain way, or to produce results, or to have evidence, or to symbolize anything. Perhaps he acts so as to *align* himself (rightly) with the deepest reality, to be aligned with this deepest reality by letting it act through him. This mode of action needs to be analyzed further, but it does not seem to involve a mode of connection to a consequence.

time; and how might the linguistic arise out of the causal and evidential? When causal and evidential connections (which arise from a branching structure of causal and statistical regularities) are common knowledge, someone might intentionally produce an evidential sign of p in order to get another to believe that p. This would be a crucial step beyond a Gricean natural meaning, which is an evidential sign, to an intentional deployment to produce a belief in another, that is, partway to a Gricean non-natural meaning wherein this intention is intended to be recognized.[72] Such an evidential sign might be produced to induce a belief in a true p that the other person cannot independently observe right then. But also —perhaps equally likely —it might first have been produced to deceive the other person into believing p, on the basis of planted evidence, when p was false. The first statement that stood for something else might have been a lie, a faked natural sign. If language defined humanity, expressing our rational capacities and distinguishing humans from other animals, then this would give an intriguing twist to the doctrine that we are born in original sin.

Prisoner's Dilemma

The Prisoner's Dilemma is a much-discussed situation where each party's selecting a (strongly) dominant action, which appears to be the rational thing to do, leaves each of them worse off than if each had selected the more cooperative dominated action. The combination of (what appears to be) their individual rationalities leads them to forgo an attainable better situation and thus is Pareto-suboptimal.

The general situation is named after one instance of it: a sheriff offers each of two imprisoned persons awaiting trial the following options. (The situation is symmetrical between the prisoners; they cannot communicate to coordinite their actions in response to the sheriff's offer or, if they can, they have no means to enforce any

[72] H. P. Grice, "Meaning," *Philosophical Review,* 67 (1957), 377–88.

agreement they might reach.) If one prisoner confesses and the other does not, the first does not go to jail and the second will receive a twelve-year sentence; if both confess, both receive a ten-year prison sentence; if both do not confess, both receive a two-year sentence. Figure 2 represents the situation they face, where the entries in the matrix represent the number of years to be served in prison: the first number the years for the first prisoner, the second number for the second.

FIGURE 2

		PRISONER II	
		Don't Confess	Confess
PRISONER I	Don't Confess	2, 2	12, 0
	Confess	0, 12	10,10

Each reasons as follows: "If the other person confesses and I don't I will receive twelve years in prison, whereas if I do I will receive ten years; if the other person doesn't confess and I don't I will receive two years in prison, whereas if I do I will receive no years at all. In either case, whichever thing the other person does, I am better off confessing rather than not. Therefore, I will confess." Each prisoner reasons in the same way: both confess and both receive ten years in prison, whereas if both had not confessed, each would have receive only two years in prison. Individual rationalities combine to produce a joint mess. And the situation is stable in the following sense: neither one has any incentive to perform the other (more cooperative) action, given that the other party is going to confess. Their actions of confessing are in equilibrium.

The Prisoner's Dilemma situation is an instance of a more general structure (see fig. 3) where each party has a choice between two actions —call them *D* for the dominant one, *C* for the cooper-

ative —and has the following preferences among the possible outcomes of the combined actions, *a, b, c,* and *d*. Person I prefers c to *a* to *d* to *b,* while person II prefers *b* to *a* to *d* to *c*.

FIGURE 3

		II	
		C′	*D′*
I	C	*a*	*b*
	D	c	*d*

Since person I prefers *c* to *a,* and *d* to *b* , action *D* dominates action *C* and he chooses to do *D*. Since person II prefers *b* to *a,* and *d* to *c,* action *D′* dominates action *C′*, and she chooses to do *D′*. Together *D* and *D′* yield the outcome *d,* while both of them prefer the outcome *a* (which would result from *C* and *C′*) to outcome *d*. Therefore, these simple facts about the structure of the 2×2 matrix and the structure of each person's preference ordering seem sufficient to mark a Prisoner's Dilemma situation.

Some people have argued that a rational person in this situation, knowing the other also is a rational person who knows as much about the situation as he himself does, will realize that any reasoning convincing for himself will be convincing for the other as well, so if he himself concludes the dominant action is best, the other person will as well; if he concludes the cooperative action is best, the other person will as well. In this situation, then, it would be better to conclude the cooperative action is best and realizing all this, he therefore (somehow) does so. This type of argument has had a mixed reception.

ThePrisoner's Dilemma parallel's Newcomb's Problem, whether or not the two are (as some have argued) identical in all essential features. Both involve two arguments that lead to differing actions, one argument based upon the dominance principle inter-

preted in such a way as to be congenial to causal decision theory, the other argument based upon considering what each act would indicate (and what outcome therefore should be bet upon), in a way congenial to evidential decision theory. The argument that in the Prisoner's Dilemma you should expect that the other person will do as you do, even though your action does not causally affect what the other does, fits the principle of maximizing the evidentially expected utility, where the conditional probabilities need not represent any causal influence. Causal decision theory recommends performing the dominant action; evidential decision theory recommends performing the cooperative action when you think the other party is relevantly similar to yourself. It need not be that you are certain you both will act alike; it will be enough if the conditional probabilities of the other party's actions, given your own, vary sufficiently. (Notice too that evidential decision theory might lead to performing the dominant action, if you believe the other party is likely to perform a *different* act than yours, or simply if her act is independent of your own but you ascribe sufficiently pessimistic probabilities to her chances of cooperating.) As was the case with Newcomb's Problem, our confidence in each of these positions may be less than complete, and we may want to give each some legitimate weight.

In the case of Newcomb's Problem, this multiple granting of legitimate weight (or, alternatively, the lack of complete confidence) showed itself in the switching of decisions when the amount of money in the first box was varied. (Yet the structure of the problem was kept constant as judged by the two competing principles of decision which would have maintained their same decision through these changes.) In the case of the Prisoner's Dilemma, the question is what rational agents with common knowledge that they face rational agents should do. Proponents of the two differing arguments find that the abstract structure of figure 3 is sufficient to give their favored argument its compelling grip. All the dominance argument needs is that person I prefers

c to *a,* and *d* to *b,* while person II prefers *b* to *a,* and *d* to *c.* All the evidentially expected utility argument about rational agents seems to need is that each has common knowledge that each is a rational agent and that each prefers *a* to *d*. If people do lack complete confidence in these arguments, however, we should find that variations in the amount[73] of the payoffs within the abstract structure of figure 3 (while still maintaining the *order* of the party's preferences) will produce changes in the decision people would make.

Suppose that utility is measured on an interval scale, unique up to a positive linear transformation, with an arbitrary unit and an arbitrary zero point, in conformity to some variant of the standard Von Neumann–Morgenstern axioms.[74] In the situation represented by figure 4, where the matrix entries are such utility numbers, we would think that cooperation is the rational choice.

FIGURE 4

		II	
		C′	*D′*
I	C	1000,1000	0, 1001
	D	1001,0	1, 1

In general, when the cooperative solution payoffs are very much higher than the dominance ones, and when payoffs for the nonmatching actions offer only slight gains or losses over these two, then we strongly will think that cooperation is rational and

[73] More exactly — since utility is measured on an interval scale — in the ratios of differences in amounts. When the discussion to follow ignores this complication in the interests of lucidity, it can be suitably rephrased.

[74] See John Von Neumann and Oscar Morgenstern, *The Theory of Games and Economic Behavior,* 2nd ed. (Princeton: Princeton University Press, 1947), appendix. An examination of philosophical issues about the Von Neumann–Morgenstern and similar sets of conditions is contained in my *The Normative Theory of Individual Choice.*

will find that the dominance argument has little force. Alternatively, in figure 5, the cooperation solution is only slightly better than the dominant one, and the extreme values in the payoffs for the nonmatching actions diverge greatly. When we have no special ties to the other party or particular knowledge of the other party's probabilities of action, then we will think it is rational to perform the dominant action in the figure 5 situation, not running any risk of the other party's performing his dominant action, which he has a large incentive to do. (And if I go through this reasoning, and think he also is very likely to be like me, then I may well settle upon the dominant action in this case, comfortable with the realization that he will also.)

FIGURE 5

		II	
		C′	*D′*
I	C	3, 3	−200, 500
	D	500, −200	2, 2

These shifts in the decision one would make, which depend upon the (ratios of the differences in the) particular numerical utility entries in the matrix, are in accordance with the earlier principle of maximizing decision-value, for people who give some weight to each of the particular principles CEU and EEU. At what precise point their decision will shift as the utilities are varied will depend upon how confident they are in each of these principles (i.e., what weights they implicitly assign to them) and also upon the probabilities they assign to the other person's action being the same as their own. Notice, however, that even if this last is given a probability of 1, and even if the agent gives greater weight to the EEU principle than to the CEU principle, she will not necessarily perform the cooperative action. If the utility stakes are big

enough and fit the situation in figure 5, that fact can combine with the weight that is given to the CEU principle, or with the dominance principle itself (in its causal variant), or with some other principle that gives weight to the security level, to yield a recommendation of the dominant action. Even absolute confidence that the other person will act as you do is not enough to guarantee your performing the cooperative action —in the absence of absolute confidence in, or weight to, the EEU principle.[75] (I have been assuming until now that it is one particular version of the *DV* principle, with its particular weights fixed, that a person applies in all decision situations. However, it might be that for a given set of constituent principles of decision, a person assigns them different weights depending upon the type of decision situation she faces. Still, each type of situation where more than one particular principle received positive weight would be fitted by some *DV* structure or other.)

In the previous section we incorporated the symbolic utility of doing an action, its SU, within the *DV* structure alongside CEU

[75] "But wouldn't a correct theory *insist* that when the probability is 1 that the other person will behave as you do, you *should* choose the cooperative action in the Prisoner's Dilemma situation, whatever the magnitude of the utility differences in the matrix? And so isn't this divergence an *objection* to the *DV* structure?" We might wonder, though, whether the person has (one level up) complete confidence in his probability estimate of 1 and whether the lack of complete confidence might affect his action in this high-risk situation (see Daniel Ellsberg, "Risk, Ambiguity, and the Savage Axioms," *Quarterly Journal of Economics,* 75 [1961], 643–69).

Notice too that the argument proceeds too quickly from (1) common rationality, to (2) they will do the same thing, to (3) crossing out the upper-right and lower-left boxes in the matrix, representing divergent actions, to (4) arguing that, given that the choice is between the two remaining boxes, both should choose the one they prefer —each prefers —that is, both should do the cooperative action. Assuming common knowledge of rationality allows us to assume that we will reason in the same way and will end up doing the same thing. But perhaps *that* will result from our each reasoning about all four boxes in the matrix, and our each concluding that in the light of the joint strategic situation persented by the full matrix, including all four boxes, I (and he or she) should do the noncooperative action, and so both end up in the lower-right noncooperative box—*thus* satisfying the condition that we act identically. Our knowing in advance that we will do the same thing means we know we will not end up in the upper-right or lower-left box, but this doesn't mean we can therefore first delete them and then reason about the remaining situation. For perhaps the reasoning whereby we *will* end up performing the same action depends upon our *not* first deleting those divergent corners.

and EEU. It might be thought that if an action *does* have symbolic utility, then this will show itself *completely* in the utility entries in the matrix for that action (e.g., perhaps each of the entries gets raised by a certain fixed amount that stands for the act's symbolic utility), so that there need not be any separate SU factor. However, the symbolic value of an act is not determined solely by *that* act; what the act means or symbolizes can depend upon what other acts are available with what payoffs, and what acts also are available to the other party or parties. What the act symbolizes is something it symbolizes when done in *that* particular situation, in preference to *those* particular alternatives. If an act symbolizes "being a cooperative person," that will not simply be because it has the two possible payoffs it does, but because it occupies a particular position within the two-person matrix —viz., being a dominated action that (when joined with the other person's dominated action) yields a higher payoff to each than does the combination of the dominant actions. Hence, its SU is not a function of those features captured by treating an act in isolation, simply as a mapping of states onto consequences.[76] An act's symbolic value may depend upon the whole decision or game matrix. It is not appropriately represented by some addition to or subtraction from the utilities of consequences *within* the matrix. Many writers assume that *anything* can formally be built into the consequences[77] —how it *feels* to perform the action, the fact that you have done it, or the fact that it falls under particular deontological principles. But if the *reasons* for doing an act *A* affect its utility, then to build this utility of an action into *A*'s *consequences* would thereby alter the act and change the reasons for doing it; but the utility of *that* altered action will depend upon the reasons for

[76] This is how L. J. Savage treats acts within the formalism of his decision theory; cf. his *The Foundations of Statistics* (New York: Wiley, 1954). However, an act cannot be reduced in this way, even apart from issues about its possible symbolic value. See my *The Normative Theory of lndividual Choice,* pp. 184-93.

[77] See, for example, Peter Hammond, "Consequentialist Foundations for Expected Utility," *Theory and Decision,* 25 (1988), 25–78.

doing *it,* and to build this into its consequences would alter the reasons for doing the now doubly altered act, and so forth. Moreover, the utilities of an *outcome* can change if the action is done for certain reasons.[78] What we want the utilities of the outcomes to represent, therefore, is the *conditional* utilities of the outcomes given that the action is done for certain reasons.[79] This creates a problem for consequentialism in dealing with dynamic consistency issues; for it might be that the fact of having reached a particular subtree of the decision-tree gives you information which alters the utility of a future outcome. If we attempt to cope with this by insisting that the utilities within the tree always be fully specified conditional utilities, then we cannot have the *same* outcomes at any two different places in the decision-tree —to the detriment of stating general normative principles to govern such trees. (For *each* fact about an act, there might be a description that enables you to list that fact as a consequence of the act, but it does not follow that there is a description such that, for *all* facts about the act, that description incorporates them within the act's consequences. The order of the quantifiers matters.)

These considerations show that in Prisoner's Dilemma situations an action should be conceived as having a utility of its own,

[78] As a result of Newcomb's Problem, cases have been investigated where the *probability* of an outcome alters with the reasons for doing the action, thus giving rise to the literature on "ratifiability."

[79] Or even the conditional utility of the outcome given that the action is done for certain reasons *and leads to the outcome.* In the economic literature on auctions, it is pointed out that a person's estimate of the value of an outcome might change when he discovers that his particular bid was the winning one, when this indicates that other knowledgeable bidders had information, or reached conclusions, that led them to value the outcome less than he did. The ratifiability literature notes that the fact that "I decide to do *A"* can affect the estimate of the probability of a consequence *C* of *A,* in that prob(*C/I* decide to do *A*) is not equal to prob(*C*), while the auction literature notes that "my doing *A* is successful in bringing about *C"* can affect the utility of *C*, perhaps by altering the probabilities of other information which affects the utility of *C.* Thus, a fully formulated decision theory not only must utilize conditional utility (see my *The Normative Theory of Individual Choice,* pp. 144–58), but the conditional utility it utilizes must be not simply *u*(outcome *O*/the action *A* is done) but rather *u*(outcome *O*/the action *A* is done, for reasons *R,* and this *A* done for *R* leads to *O*).

not simply as involving a constant utility addition *within* a row of its matrix.[80] But I wish to claim something stronger —namely, that this utility is a *symbolic* utility. This is not simply the usual kind of utility applied to an action rather than an outcome. This utility involves a different kind of connection. In some Prisoner's Dilemma situations, doing the dominated action —what is usually, called the "cooperative action" —may have symbolic value for the person. It may stand for his being a cooperative person in interactions with others, a willing and noncarping participant in joint ventures of mutual benefit. Cooperating in this situation then may get grouped with other activities of cooperation that are not embedded in Prisoner's Dilemma situations; not cooperating in this particular Prisoner's Dilemma situation may then come to threaten his cooperating in those other situations —the line between them may not be so salient, and his motivation for cooperation in the others may also be partly symbolic. Giving great utility to being a cooperative person, in a particular Prisoner's Dilemma situation he performs the dominated act that symbolizes this.[81]

This does not mean this person will look only at that act's SU. He also will consider its particular utility entries and how these are evaluated by the CEU and the EEU principles. The decision-value of the act for him will depend upon all three of these things —its SU, CEU, and EEU —and upon the weights he gives to these. Thus, the mere fact that he gives some (positive) symbolic utility to being a cooperative person does not guarantee he will perform the cooperative action in all Prisoner's Dilemma situations.

[80] It also is worth mentioning that when the sequencing of the actions is strategically relevant, game theorists do not simply concentrate upon the matrix-representation of a game and its payoffs, but need to consider the game-tree.

[81] Can one build this into the standard decision theory by saying that one constant consequence of his performing the dominant act in the Prisoner's Dilemma situation is that he will think of himself as a noncooperative person, and then representing this in the game matrix by a negative addition, an addition of negative utility, all across the row for that action? Notice that this component of utility would be a function of his attitude toward that act as it stands within the structure of the whole matrix.

I do not claim that the only possible symbolic meaning relevant to the Prisoner's Dilemma situation is "being a cooperative person." Someone might think that performing the *dominant* action in such situations symbolizes "being rational, not being swayed by sentimentality"; thinking this quite important, he gives great symbolic utility (within his *DV* principle) to performing the dominant action, this in addition to the weight he gives to the CEU or dominance principle itself. Some writers on Newcomb's Problem who are proponents of the view that taking what is in both boxes is most rational overcome discomfort at the fact that they and people like themselves do worse on this problem than do maximizers of EEU by saying its "moral" is "if someone is very good at predicting behavior and rewards predicted irrationality richly, then irrationality will be richly rewarded." [82] I take it that such people give very great utility —is that a symbolic utility? — to being rational according to their best current estimate of what precise principles that involves. (It will be a subtle matter to distinguish between someone who gives weight to only *one* particular principle, CEU for example, and someone who gives some weight to CEU and also some lesser weight to EEU yet also attaches great symbolic utility —greatly weighted —to following her best *particular* estimate of what rationality involves.) One would guess that new complications would arise if following a particular decision principle itself has symbolic utility or if engaging in a particular kind of decision process or procedure does.

To say all this about symbolic utility is to say that our responses to the Prisoner's Dilemma are governed, in part, by our view of the kind of person we wish to be and the kinds of ways we wish to relate to others. What we do in a particular Prisoner's Dilemma situation will involve all this and invoke it to different degrees depending upon the precise (ratios of differences among) utility entries in the matrix and also upon the particular factual circum-

[82] Gibbard and Harper, "Counterfactuals and Two Kinds of Expected Utility," 151.

stances that give rise to that matrix, circumstances in which an action may come to have its own symbolic meanings, not simply because of the structure of the matrix.

We knew all this already, of course, at least as a psychological point about why people differ in their responses to Prisoner's Dilemma situations. However, the *DV* principle leaves room for general views about what sort of person to be, as this relates to and groups particular choices, not simply as a possible *psychological* explanation of why (some) people deviate from rationality, but as a legitimate component, symbolic utility, within their *rational* procedure of decision.

In a seminal paper on the repeated Prisoner's Dilemma,[83] Kreps, Milgrom, Roberts, and Wilson showed that your giving a small probability to my performing the cooperative action or your giving this to my believing that you will perform the cooperative action (or your giving a small probability to my believing that you will believe that I will perform the cooperative action) can be sufficient to make it rational for you to begin by performing the cooperative action, in order to encourage me in my cooperative action or consonant beliefs. If you believe I might do the cooperative action (or follow tit-for-tat), and you believe that I will continue to do so only if *you* behave a certain way, then you will have reasons to behave as I think you might, in order to encourage me to do the cooperative action.[84] If the situation is mutual, both will (under certain circumstances) perform the cooperative action. Now the *DV* structure, when it is common knowledge that both follow it, does (promise to) give some probability of each player believing that the other will believe that the first will perform the cooperative action, and hence some probability of each, of both,

[83] David P. Kreps, P. Milgrom, J, Roberts, and R. Wilson, "Rational Cooperation in the Finitely Repeated Prisoner's Dilemma," *Journal of Economic Theory*, 27 (1982), 245–52.

[84] As one writer puts it in summary, a player might "take an out of equilibrium action to set in motion the other player's out of equilibrium beliefs and strategies" (Eric Rasmussen, *Games and Information* [Oxford: Basil Blackwell, 1989], p. 111).

performing the cooperative action. (Notice that this point, and the remainder of this paragraph, does *not* depend upon the full *DV* structure which includes a weighting for symbolic utility. The narrower structure first presented, with a weighting only of CEU and EEU, is enough.) And this, not as a perturbation away from full rationality, and not as one's rational adjustment to the other's deviation from rationality (or to the other's belief that you might deviate from rationality), but rather as a part of common knowledge that all participants are *totally* rational. For if the principle of maximizing decision-value is a rational principle, normatively desirable, then if (as it appears to) common knowledge of *DV*-maximization gives some probability of each participant's performing the cooperative action, the argument of Kreps, Milgrom, Roberts, and Wilson applies even under common knowledge of full rationality.[85]

It would be nice to reach a sharper result than that the cooperative action will be performed if the causal, evidential, and symbolic utilities interact so as to lead to this. Under what conditions, for what specifications of weights within a *DV* structure for one (or both) of the participants, will a person choose to perform the cooperative action in the Prisoner's Dilemma situation or follow a tit-for-tat strategy in the repeated Prisoner's Dilemma.[86]

Here we can take only some tentative first steps in listing appropriate assumptions for deriving results. In addition to requir-

[85] A side note in passing: in my 1963 doctoral dissertation, I saw the necessity for game-theoretic situations of levels of knowledge infinitely extended, each knowing the structure of the game theoretic situation, each knowing the other knows, each knowing the other knows that he knows, and so on *(The Normative Theory of Individual Choice,* p. 274). But I thought this just was a nit-picking point. Little did I see the far-reaching interest and implications of the condition of common knowledge of rationality. See Robert Aumann, "Correlated Equilibrium as an Expression of Bayesian Rationality," *Econometrica,* 55 (1987), 1–18, and Drew Fudenberg and Jean Tirole, *Game Theory* (Cambridge: MIT Press, 1991), pp. 541–72.

[86] On the tit-for-tat strategy, see Robert Axelrod, "The Emergence of Cooperation among Egoists," reprinted in Campbell and Sowden, *Paradoxes of Rationality and Cooperation,* and *The Evolution of Cooperation* (New York: Basic Books, 1984).

ing that both players follow the *DV* principle, we can add an extremely weak form of the assumption that each should expect the other player to behave as he does, to be fed into the EEU component. The weak predictive principle says that the evidential conditional probability that the other player will do act *C',* conditional on your doing *C,* is greater than the unconditional evidential probability that she will do *C';* and similarly for her act *D'* conditional upon your own. A somewhat stronger principle, but still short of the symmetry assumption that the other rational player will act exactly as you do, would hold that these evidential conditional probabilities, for the first play, are greater than 1/2. Another principle specifies that the person gives *some* symbolic utility (and some symbolic weight to that) to performing the cooperative act in the Prisoner's Dilemma situation. Moreover, we might assume that performing the dominant act *D* has a negative symbolic utility of its own, in addition to the absence of the positive symbolic utility of cooperating." Let $S(A/B)$ be the symbolic utility of act *A* given that the other person does act *B*. If person I assigns positive symbolic utility to performing the cooperative action, then $S(C/C')$ is greater than or equal to $S(C/D')$, and each of these is greater than (the negative quantity) $S(D/D')$, which itself is greater than (the more negative) $S(D/C')$. When the Prisoner's Dilemma structure is repeated many times between the same two persons, the further possibilities of mutually beneficial cooperation affect the utilities within a current play, including the very first one. Moreover, the symbolic utility of an action will change from play to play, depending upon the past actions of the other party. We might see the symbolic utility of performing the cooperative action as declining, the more the other party performs her dominant action, perhaps as declining proportionally to the ratio of the number of times the other party has performed her dominant ac-

[87] I speak intuitively here, since on an interval scale of measurement, with an arbitrary zero point, there is no special significance to a measured quantity's being negative.

tion to the number of times she has performed her dominated one. Cooperating with *her* becomes less a symbol of being a cooperative person, the more she has refused to cooperate. On the other hand, the more the other person cooperates, the more symbolic utility your performing the cooperative action will have. And a comparable condition now applies to the negative symbolic utility of performing the dominant act. This disutility also declines in absolute amount the more the other person performs her dominant action and increases in absolute amount the more she performs her cooperative action. The hope is that these conditions, along with other plausible assumptions, will give us sharper results.

Finer Distinctions: Consequences and Goals

We have discussed three different modes of connection of action to outcomes —namely, causal, evidential, and symbolic — and have suggested that decision theory needs to utilize and explicitly recognize all three modes. Does decision theory also need finer discriminations *within* these categories? For example, some writers on ethics have claimed that different kinds of causal connections carry different weights in choice situations, even though the resulting probabilities may be identical. There is a significant difference, they claim, between bringing something about and allowing it to happen or abstaining from preventing it. (And we might consider further kinds of causal relation, such as facilitating or aiding its happening.) And some writers have formulated a doctrine of "double-effect," holding that there can be a moral difference (sometimes sufficient to make the difference as to whether an action is permissible) between bringing something about when this results from intending to bring it about as an end or a means to an end and knowingly bringing it about but as a side-effect of one's pursuit of some other goal. Admittedly, these are matters of some controversy,[88] yet it is striking that causal decision theory

[88] See Philippa Foot, "The Problem of Abortion and the Doctrine of the Double Effect," in her *Virtues and Vices* (Berkeley: University of California Press,

thus far has taken no notice of these arguably important distinctions; it proceeds instead with an undifferentiated notion of "causal influence." Should normative decision theory make room for such distinctions and give them a role, either in its first-person theory of choice or in its instructions for an adviser? One natural place these distinctions might enter is in the notion of conditional utility. Earlier, in speaking of auction theory, we noted that decision theory should speak of u (outcome $0/A$ is done and A causes or succeeds in bringing about 0). The precise kind of causal linkage between action and outcome within the last part of this condition might affect the utility of the resulting outcome 0, that is, yield differing conditional utilities for 0 and hence sometimes produce different decisions within a principle that utilizes such conditional utilities. Or is the import of these distinctions wholly symbolic, so that by incorporating symbolic utility within our theory we already have made an adequate place for them?[89]

I suggest we see these distinctions, not as dichotomies, but as arrayed along a (not necessarily continuous) dimension. Indeed, we have here not one dimension but two. The first involves the importance of the causal role of the action in relation to the effect or outcome or resulting state of affairs. Here we have (at least) seven relations an action may stand in to a state of affairs. In

---

1978), pp. 19–32; Judith Thompson, "Killing, Letting Die, and the Trolley Problem," and "The Trolley Problem," in her *Rights, Restitution and Risk* (Cambridge, Mass.: Harvard University Press, 1986), pp. 78–116; Warren Quinn, "Actions, Intentions, and Consequences: The Doctrine of Double-Effect,'' *Philosophy and Public Affairs,* 18 (1989), 334–51; Warren Quinn, "Actions, Intentions and Consequences: The Doctrine of Doing and Allowing," *Philosophical Review* (1989), 287–312; Frances Kamm, "Harming Some to Save Others," *Philosophical Studies,* 57 (1989), 227–60.

[89] Or, instead, are these distinctions framing effects, in the sense of Tversky and Kahneman, which show variance across (descriptions of) situations where there should be invariance? See Amos Tversky and Daniel Kahneman, "Judgment under Uncertainty: Heuristics and Biases," *Science,* 185 (1974), 1124–31; reprinted in Daniel Kahneman, Paul Slovic, and Amos Tversky (eds.), *Judgment under Uncertainty* (Cambridge: Cambridge University Press, 1982). Doesn't the relation of the bringing about/allowing to happen distinction to a baseline seem suspiciously like that of the gain/loss distinction to *its* baseline? This last, of course, is a favored example for framing effects.

decreasing importance, the action may: (1) cause the state of affairs to occur; (2) aid or facilitate its occurrence; (3) remove a barrier to its occurrence; (4) permit or allow its occurrence; (5) not prevent and not avoid its occurrence (when some act available to you would have); (6) not aid or facilitate its nonoccurrence (when some act available to you would have); (7) not aid or facilitate its nonoccurrence (and *no* act available to you would have).

The second dimension also involves the causal role of the action, in relation to the effect or outcome or resulting state of affairs, but this dimension marks not the act's importance but its *robustness.* The idea is that when something is pursued as a goal, certain subjunctives hold true of the person. He would reorganize his behavior in order to reach the goal (or to have a better chance of reaching it); in slightly different circumstances, where *this* action would not reach that goal, he would do something different instead that *would* reach the goal; he would tend to exclude alternative actions that have no possibility of reaching the goal. When something is merely a known side-effect of the action, on the other hand, the person would not alter his behavior if it turned out that his (current or planned) behavior would not produce this side-effect. Of course, if it instead produced another significant effect he wished to avoid, he might do so. It is a question of the *range* of the situations where the behavior would alter. Pursuit of something as a goal involves subjunctives across a wider range of circumstances than acting in the knowledge that something will result as a side-effect of other goal pursuits. Between these two falls aiming at something solely as a means toward the realization of some other goal. In this case, behavior would be reorganized in some situations to realize the means — unlike the side-effect case — but in a narrower range of possible situations than where the effect is an end or goal itself. (Consider, for example, that possible situation where this particular effect no longer serves as a means to the goal.)

Along this dimension of robustness of the causal role, we can distinguish (at least) six connections of a person and an action to an effect or outcome. The action can: (1) aim at the effect as an end; (2) aim at the effect solely as a means. Or it can not aim at the effect at all. And among actions that do not aim at the effect, the person might: (3) know of the effect (which is not aimed at); (4) not know of the effect (which is not aimed at) that she should know of; (5) not know of the effect (which is not aimed at), and it not be the case that she should know of it. Or (6) the state of affairs (which is not aimed at) occurs by accident.

Utilizing these two dimensions, and their categories, we can form a 7 by 6 matrix. (If the two dimensions are not completely independent, some of the boxes may be impossible.) An action and a person's relation to its effect (or to the resulting state) will be specified by its location within the matrix, that is, by its position along the two dimensions.[90] Should decision theory take account of these finer distinctions concerning the mode of causal connection of an action to an outcome, and, if so, how? Are there also finer distinctions *within* the evidential and the symbolic connections that decision theory should mark and take into account? I raise these questions not to answer them here but to place them on the agenda.

The themes discussed in these two lectures about principles and symbolic meaning apply to ethical principles also. By grouping actions together in a class, one action comes to stand for all, and the weight of all is brought to bear upon the one, any one, giving it a coordinate (symbolic) disutility. Deontological constraints might exhibit this same phenomenon. By grouping actions together into a principle forbidding them —"do not murder" —an action

[90] For other purposes, we might want to extend such a matrix, adding a third dimension to represent the magnitude of the consequence or effect. (Extending the matrix in this way for legal contexts was suggested to me by Justin Hughes. In legal contexts we might want to know how bad the effect was: how bad was the one aimed at, and how bad was the one which occurred.) But within decision theory, of course, this magnitude is already represented by the utility of the outcome.

is removed from separate utilitarian (or egoist) calculation of *its* costs and benefits. The action comes to stand for the whole group, bearing its weight upon its shoulders. This need not happen in a way that makes the constraint absolute, barring the action no matter what, but it constitutes a far greater barrier to performing it, by throwing its greatly increased (symbolic) disutility into any calculation.[91]

Recall now our discussion in the first lecture of the symbolic meaning of following ethical principles; ethical action can symbolize (and express) being a rational creature that gives itself laws, being a law-making member of a kingdom of ends, being an equal source and recognizer of worth and personality, and so forth. The utility of these grand things, symbolically expressed and instantiated by the action, becomes incorporated into that action's symbolic utility and hence into that action's decision-value. Thus, these symbolic meanings become part of one's reason for acting ethically. A person who maximizes an act's utility broadly conceived, that is, who maximizes its decision-value *(DV)*, may be led to perform ethical actions. This person would be pursuing his *own* goals (which need not be *selfish* goals). In terms of the categorization of Amartya Sen,[92] he therefore would be engaged in self-goal pursuit rather than the activity of *not* marginally pursuing *his* own overall individual goal. But note that if falling into this further category of not marginally pursuing his own overall individual goal itself comes to have symbolic utility to him, then it will enter into his *DV*. At that point, when he acts taking account of this symbolic utility, is he once again pursuing his own goal, that is, *his* revised *DV,* so that his attempt (within the *DV* framework) to enter Sen's other category is doomed to failure? How-

[91] Recall also the discussion above of how a meta-principle not to violate any principles might make any violation stand for all, thereby giving every principle heavy deontological weight.

[92] See Amartya Sen, *Ethics and Economics* (Oxford: Basil Blackwell, 1987), pp. 80–88.

ever we decide this, the more general point holds. Being ethical is among our most effective ways of symbolizing (a connection to) what we value most highly, and that is something a *rational* person would not wish to forgo.

We discussed various functions of principles in the first lecture. Accepting and adhering to a particular principle, we saw, could be considered to be a (general) act A and treated within a decision-theoretic framework that was largely instrumental. Now we have presented an alternative framework for decision theory, one that includes evidential and symbolic aspects, not simply causal instrumentality. Within *that* framework, an act of accepting a principle will have a decision-value DV, and it will be chosen (from among alternatives) when it has a maximal decision-value. This broader framework opens the way to a revised discussion of why we have principles at all —why if we have that one DV principle we also will have some others —and of why we have some particular ones.