5

WITH SARAH WRIGHT

What Nozick Did for Decision Theory

Robert Nozick's seminal 1969 essay ("Newcomb's Problem and Two Principles of Choice") introduced to philosophers the puzzle known as Newcomb's problem. Nozick returned to the problem in his 1974 essay "Reflections on Newcomb's Problem" and then again in *The Nature of Rationality* (1993). We describe the problem, then explain what it tells us about the nature and limits of decision theory. The problem is as follows.

1. Newcomb's Problem

You are presented with two boxes. A transparent box holds $1,000.[1] A second box is opaque, but you know it contains either nothing or $1 million. You are offered a choice. You may take only the opaque box, or you may take both: that is, you take the opaque box by itself, or the opaque box plus the box that you know to contain $1,000. It seems obvious you should take both boxes, that is, you take the extra $1,000 on top of whatever is in the opaque box.

Here is the catch. There is a predictor who predicted whether you will take one or two boxes. The $1 million is in the box if and only if the predictor predicted you will take only the opaque box. The predictor has a history of being correct 90 percent of the time.[2] What should you do?

1. In Nozick's original presentation of Newcomb's problem, both boxes are opaque. We treat the box with the $1,000 as transparent because it makes for easier exposition and does not affect the structure of the original case.

2. In the original presentation, you are said to have "enormous confidence" that the predictor can correctly predict your behavior. We have quantified this expression for use in later calculations.

The literature is full of answers. What we wonder is: what was Nozick's point in asking? Another question: why did the question spark such fierce debate? Why doesn't decision theory simply settle the matter?

2. World State Partitions

Suppose we represent Newcomb's problem as a decision matrix. Should the matrix look like this?

	Predictor predicts one	*Predictor predicts two boxes*
You take one box	M	0
You take two boxes	M+K	K

M=$1 million; K=$1,000

Or should the matrix look like this?

	Predictor is correct	*Predictor is mistaken*
You take one box	M	0
You take two boxes	K	M+K

The second matrix is a different way of representing the same problem. The actions are the same: the top row represents taking one box; the bottom row represents taking two. The columns differ in terms of how they *partition* the states of the world forming the background of your choice (although in either case the world state descriptions are meant to be mutually exclusive and jointly exhaustive; their probabilities sum to one). The first matrix partitions the world into two possible states: the predictor predicts either one box or two. The second matrix partitions the same world along different lines: the predictor is either correct or mistaken.

In Leonard Savage's decision framework, each column in a decision matrix represents a state of the world, and to each column attaches a probability. In the first matrix, the probabilities of columns appear irrelevant to the question of what you should do, because no matter which column you are in (that is, no matter what the world state) you are $1,000 better off taking two boxes. Taking two is a *dominant strategy*.

If you frame the problem the second way, though, you see neither strategy as dominant. So, in weighing your options, you want to know the probability of each column. In the second matrix, there is a 90 percent chance that the first column represents the background state of the world in which you make your decision. Therefore, to people who partition world states the second way, clearly you should take one box; taking one box implies a 90 percent chance of winning a million.

The problem: to those who partition states the first way, clearly you should take both boxes. What would make it better to see the problem in one way rather than the other? Here is a preliminary answer. The first way of partitioning states, implying that you have a dominant strategy, is a mistake when the probability of your being in a given state (that is, a given column) depends on which action you choose. Thus, if choosing one box would make it likely you were in the state represented by the first column, and choosing two would make it likely you were in the state represented by the second column, then choosing one box approximates choosing $1 million in preference to $1,000. Playing your apparent dominant strategy would be a big mistake.

The second way of partitioning world states can likewise be a mistake, but explaining why requires a bit more background.

3. From Savage to Jeffrey

Forget about the Newcomb story for a moment, and just look at Nozick's (1969) example of an uninterpreted matrix of values.

	State 1	*State 2*
Act 1	10	4
Act 2	8	3

We have not yet assigned numbers to the likelihood of being in state 1 as opposed to state 2. It appears not to matter, because act 1 dominates act 2. But what if your choice of action *affects* the probability of being in state 1? This is the factor that, when present, makes dominance reasoning untrustworthy. To use dominance reasoning, we must start by asking whether we know of a way of partitioning world states such that (1) the probability of being in a given column does not depend on which action you choose, and (2) a dominant strategy exists when the world is partitioned that way.

Since Savage conceived of probabilities as attaching to matrix columns, the only background probability that could be modeled was a kind independent of which row (that is, which action) the agent chooses. Richard Jeffrey revolutionized the framework by developing a theory that lets us consider cases where the likelihood of being in a given state depends on the act chosen. Jeffrey modeled such cases by supplementing his desirability matrix (giving values of each consequence) with a probability matrix (giving a probability of each state given each act).

Desirability Matrix

	Predictor predicts one	*Predictor predicts two boxes*
You take one box	M	0
You take two boxes	M + K	K

Probability Matrix

	Predictor predicts 1 box	*Predictor predicts 2 boxes*
You take 1 box	P(predicted 1 / you take 1) = .9	P(predicted 2 / you take 1) = .1
You take 2 boxes	P(predicted 1 / you take 2) = .1	P(predicted 2 / you take 2) = .9

This type of matrix lets us represent the Newcomb predictor's accuracy in terms of conditional probabilities. Thus, P (predicted 1 / you take 1) is read as "the probability of predictor having predicted one box given that you took one," which in this case equals 0.9. [3] Jeffrey's probability matrix lets us represent the probabilities of states as depending on the actions chosen, which enables us to develop models of situations where agents can influence the probability of being in a given state. Putting together the two kinds of information—multiplying this conditional probability by the value of taking one box when the predictor has predicted one box, and so on—we can calculate expected utilities and represent Newcomb's problem as follows:

	Predictor predicts 1 box	*Predictor predicts 2 boxes*	*Expected utility*
You take 1 box	M x .9	0 x .1	= $900,000
You take 2 boxes	(M + K) x .1	K x .9	= $101,000

The conditional probabilities are what dominance reasoning ignores, but once we build them into the matrix, Jeffrey's framework appears to give us decisive reason to choose one box. In effect, Jeffrey's apparatus revealed a potential problem with dominance reasoning: sometimes, the appearance of a dominant strategy cannot be taken at face value.

Nozick's contribution, four years later, was to notice a somewhat analogous problem with expected utility reasoning. That is, sometimes the fact of being the option with maximum expected utility cannot be taken at face value either.

4. Probabilistic Connection without Causal Influence

What exactly is it that cannot be taken at face value? To put it somewhat paradoxically, Nozick noticed that the action having maximum expected utility is not necessarily the action that maximizes expected utility.

3. This is the standard construal of the relevant conditionals, and we need the standard construal in order to depict the problem in the standard way. Isaac Levi (1975), though, notes that a natural way to represent the idea that the predictor is right 90 percent of the time is with a converse conditional: for example, p(you take 1 / predicted 1) = 0.9, not p(predicted 1 / you take 1) = 0.9. As Levi proves, that p(you take one / predicted 1) = 0.9 does not entail that p(predicted 1 / you take 1) = 0.9. We henceforth assume this caveat goes without saying.

We explained when dominance reasoning leads to error: that is, when it leads us to overlook ways in which probabilities of world states depend on which act we choose. We now can explain when expected utility reasoning leads to a parallel sort of mistake. To get at what is wrong with being a one-boxer, consider a case Nozick proposed that has a causal structure relevantly like Newcomb's: ACADEMIC DISEASE.

Joe knows that Stu or Tom is his father. Stu died of an inheritable disease later in life. Tom did not. Stu's disease also was genetically linked to an academic propensity. Joe now has to choose a career. Other things equal, Joe would rather be an academic, but reasons that if he goes into athletics, he is less likely to be a child of Stu, and thus less likely to get the disease. Nozick says it would be wild to decide on that basis, since there is nothing Joe can do to *make* it less likely that Stu is his father, thus nothing he can do to make it less likely that he will get the disease.[4]

The crucial fact, to Nozick, is not whether a world state is probabilistically linked to your action but whether it is *influenced* by your action.[5] Whether M is in the opaque box seems to depend on whether you choose one box or two. That probabilistic dependence suggests that dominance reasoning is a million-dollar mistake in the Newcomb context. Contra dominance reasoning, you do best when you take one box rather than two.

Why is this not the end of the conversation? The problem is that what normally makes conditional probabilities relevant is missing in the Newcomb situation. If our picking one box would *make* it more likely that the predictor will put $1 million in the opaque box, that would be part of the action's overall utility. Conditional probabilities often indicate a tendency of the action to influence the probability of being in a given state. However, they *need* not, and when they *do* not, they are not relevant. And that kind of conditional probability is the kind we are given in Newcomb's problem. Probably.

5. The Prisoner's Dilemma

Likewise, in a prisoner's dilemma, if one's deciding to cooperate would *cause* one's partner to cooperate, that is one of the effects of one's action, and thus is part of the action's overall utility.[6] Here is the general form of a prisoner's dilemma.[7]

4. Another case: John Calvin said the devout go to heaven, but there was ambiguity about why. Do they go to heaven *because* they are devout? If so, expected utility gives the right answer: be devout. Or do they go to heaven because of predetermined grace, a side effect of which is a neurotic urge to be devout? In this second case, assuming it's more fun to be a party animal, then expected utility gives the wrong answer.

5. Nozick (1969) 123.

6. Many have wondered whether the prisoner's dilemma is a Newcomb problem (Lewis 1979, Sobel 1985). We are not making a claim of equivalence here.

7. Here is the classic case from which the dilemma gets its name. You and your coconspirator, a person about whom you care little, and from whom you do not fear retribution, have been arrested and charged with a crime. You are offered the chance to testify against your partner, in which case he gets a long sentence and you go free—unless he testifies against you as well, in which case you each get a medium sentence. If each of you refuses to testify, then each of you gets a short sentence.

Standard Prisoner's Dilemma Partition

	Partner cooperates	*Partner defects*
You cooperate	M (partner gets M)	0 (partner gets M + K)
You defect	M + K (partner gets 0)	K (partner gets K)

Partitioned in this way, the problem seems to have a dominant strategy solution. No matter what your partner does, you are better off defecting, and likewise for your partner. But there is an issue here that is analogous to Newcomb's problem. The analogous question for dominance reasoning here is, what if your partner is a lot like you? What if you reasonably predict that, if your reasoning is leading you to choose a particular action, then the same reasoning will be leading your partner to the same conclusion?

An Alternative Partition

	Partner plays like you	*Partner plays unlike you*
You cooperate	M (partner gets M)	0 (partner gets M + K)
You defect	K (partner gets K)	M + K (partner gets 0)

The second matrix is merely a different model of the same situation. In this alternative model, though, there is no dominance. If we consider it likely that "my partner plays like me," then we will calculate that the expected utility of cooperating exceeds that of defecting. Suppose "my partner plays like me" is likely. Your forming an intention to cooperate seemingly makes it more likely that your partner is likewise forming an intention to cooperate. In that case, is cooperation rational? Yes, to judge by the numbers. If we are fairly sure we are working in the first column of the second matrix, then to choose cooperation is, in effect, to choose a payoff of M over a payoff of K. But it depends on causal structure. If cooperation *makes* it more likely (that is, *causes* it to be more likely) that your partner cooperates, then cooperation is in your interest. If not—if you know your partner's choice is independent of your action—then cooperation is not in your interest.

The second matrix is somehow a mistake, although it is not transparently so. It is a mistake insofar as there is in fact a dominant strategy that the matrix fails to reveal. And it is a mistake insofar as the feature that would give the second matrix its point—that it successfully partitions world states in such a way as to reflect *relevant* conditional probabilities—is not actually in place.[8]

8. As Nozick later came to express the point, in a prisoner's dilemma, *causal* reasoning recommends the dominant strategy; *evidential* reasoning recommends cooperation when you think the other party is so similar to you that your cooperation can be taken as evidence that the other player will cooperate (1993, 48).

A high conditional probability of your partner cooperating, given that you cooperate, could be a sign that your cooperation tends to induce your partner's cooperation.[9] However, a second possibility is that a high conditional probability may reflect the fact that the same reasoning that leads you to cooperate tends *independently* to lead like-minded partners to cooperate as well. In this second case, expected utility reasoning gives the wrong answer—because it leads you to think of your cooperation as *producing* the preferred outcome, when in fact your action has nothing to do with the process you hope will culminate in your partner deciding to cooperate.

We normally and appropriately treat probabilistic dependence as a sign of causal influence. If we *know* there is no such influence, probabilistic dependence becomes irrelevant.

6. Unknown Causal Structure

So, what if we *suspect* influence? Then we must decide whether and how to take such suspicion into account.

The *Fisher Smoking Hypothesis* begins with the observation that cancer correlates to smoking. We then suppose there are two alternative causal structures that could underlie the correlation. Perhaps smoking causes cancer. In that case, expected utility reasoning gives the right answer: don't smoke. Or, perhaps smoking does not cause cancer, but if not, then why are smokers at a higher statistical risk of getting lung cancer? The Fisher hypothesis is that smoking and cancer are each caused by a defective gene. Thus, there is a statistical correlation not because smoking causes cancer but because smoking and cancer are effects of a common (genetic) cause. In that case, expected utility reasoning gives the wrong answer.

Fisher Hypothesis

	You won't get cancer	*You will get cancer*
You don't smoke	M	0
You smoke	M + K	K

K = the pleasure of smoking; M = utility of not getting cancer

The symbols M and K are awkward here, but the point of using them is to reveal what the Fisher hypothesis is hypothesizing: namely, that smoking is a Newcomb problem, a situation where cancer is probabilistically but not causally linked to smoking. (Compare this matrix to the first matrix in section 2.) If the Fisher hypothesis is correct, you apparently have a dominant strategy. Regardless of whether you are going to get cancer, smoking is fun (we are supposing).

9. That could be so in, for example, an iterated prisoner's dilemma, where the game is played over several rounds, and in any given round, your partner can respond to the manner in which you played the previous round. In this case, expected utility gives the right answer: cooperate, so long as your partner is cooperating in return, and so long as your partner likewise would respond to defection by defecting.

Does that mean you should smoke? No! It depends on whether the world state is independent of the action. Although it looks like smoking dominates here, what if Fisher's hypothesis is wrong? What if the reason for the probabilistic link between smoking and cancer is that smoking does indeed *cause* cancer? The next partition models that uncertainty.

Alternative Partition

	Smoking doesn't cause cancer	*Smoking causes cancer*
You don't smoke	M	M
You smoke	M+K	K

Here we have a different partition, where state probabilities clearly are unaffected by the choice of action.[10] That is, my choice of action clearly is not what determines the general truth of the matter regarding whether smoking causes cancer. But in this matrix, smoking does not dominate. Accordingly, this alternative partition suggests the *correct* conclusion that whether you should smoke (assuming smoking is fun) depends on whether smoking causes cancer.[11]

The second matrix is a better way of representing the problem if smoking turns out to cause cancer. What if you do not know? In that case, is it still better to represent the problem in the second way? Yes, because the second matrix makes plain that the question of causal structure is *the* question. The second matrix does not tell us what to do, but it does focus us on the pivotal question.

Maximizing expected utility gets wrong answers in cases like Fisher Smoking Hypothesis. However, the only reason we know it gets the wrong answer is that by stipulation we supposedly know smoking does not cause cancer. If smoking and getting cancer have a common cause but smoking does not per se have direct causal bearing on whether a person gets cancer, then the lowering of expected life span, given that you smoke, is misleading. Therefore, just knowing the numbers (utilities,

10. There is a background risk of cancer, of course. In the foregoing matrix, M is the utility of being subject only to this background risk, avoiding the added risk that goes with smoking if smoking causes cancer. Note that the result is the same in both columns of the top row, meaning that so long as you do not smoke, it does not matter whether smoking causes cancer.

11. The literature speaks of deciding under *uncertainty* when the agent cannot be certain of the world state but does know the probabilities. We speak of deciding under *ignorance* when we do not even know the probabilities. If I am ignorant, I can simply assign probabilities of 50 percent to signify that, so far as I know, one state is as likely as the other. Even if that were to get the right answer, though, it is not clear that it would be getting the right answer for the right reason. The procedure simulates the appearance of algorithmic decision theory, but what is really going on is that you are comparing two possibilities: the minor value of K versus a package consisting of K plus an M-sized catastrophe. And you are saying you are not interested in taking such a risk, and would not become interested unless you were sure that the probability of the catastrophe were, let's say, under one in a thousand. So long as you do not know that the risk is under one in a thousand, you cannot exactly know there is a higher expected utility in refusing to take the risk, but that is how a rational gambler would play it.

probabilities) isn't enough. We could know the numbers, and still be in a situation of deciding under ignorance.[12]

Going back to the Newcomb problem, we will rephrase our view. If I *know* my play has no causal bearing on what is in the opaque box, I take both boxes. If I am not sure of the situation's causal structure, and *cannot* rule out the possibility that the prediction is influenced by my choice, I have reason to take only the opaque box.

If all I know for sure is that one-boxers average $900,000 while two-boxers average roughly $100,000, I rationally opt for one box. If the opaque box were to become transparent, though, then average payoffs would no longer be relevant.[13] I would take two boxes. Notice that the only thing that changes when we make the opaque box transparent is that my level of confidence jumps regarding the independence of the state from my action. Yet, given the stakes, that is enough to change a rational contestant from a one-boxer to a two-boxer.

That tells us when expected utility reasoning is apt. Expected utility reasoning is apt for situations where we know we can influence the probability of being in a given world state. Less certainly, we also use, and are not obviously mistaken in using, expected utility reasoning in circumstances where we do not know the situation's causal structure but cannot rule out the *possibility* that our actions influence the probabilities of world states.

A probabilistic link between an act and a state can signify that the act affects the probability of the agent being in the state. Or it could be a sign of influence going the other way, that is, the state influences the likelihood of the agent choosing the act. Or it could be a sign that the state and the act are separate effects of a common cause. If we do not know what correlation has to do with causal influence, we guess. To jump to the conclusion that there is a causal relationship is known in philosophy as *post hoc, ergo propter hoc*. That jump is a fallacy. The premise that B happens after A does not guarantee the conclusion that B happens because of A.

So *post hoc* reasoning is invalid, but is it a mistake? That is a different question.

FOOD POISONING: I eat mushrooms. I get sick. I jump to the conclusion that the mushrooms made me sick. My inference is invalid, to be sure, but it also leads me to consume less poison than I otherwise would. The *post hoc* fallacy is our heuristic for coping with the kind of ignorance we face every day. We rely on that kind of reasoning. At the same time, interestingly, we do not lightly bet the farm on it. If the alternative to eating mushrooms is to starve, then I do not conclude that mushrooms

12. Some theories of causal decision attempt to justify assertions about causal connection entirely in terms of probabilistic dependence (Pollock, 2002). To use probabilistic dependence for this purpose, though, one needs to know various conditional probabilities—the sort of probabilities one would know only in virtue of understanding a situation's causal structure.

13. The parallel point also holds in the Fisher case. Suppose the average person has a gene for cancer, but the phenotypic expression of that gene occurs only in smokers. In that case, smoking causes cancer in the average person. Nevertheless, my real concern is not whether smoking would cause cancer in average smokers but whether it would cause cancer in me. So, if my own genetic "box" is transparent, so that I see I do not have the genetic risk-factor, then the fact that the average smoker gets cancer is not relevant.

make me sick. I draw that conclusion only if the cost of eschewing mushrooms is acceptable. Or perhaps the mechanism is hard-wired. We are wired to dislike foods we ate just before becoming ill. We presumably would not be wired that way except that we evolved under conditions where we had various things to eat, so we could *afford* to be wired to be dislike suspect foods.

We have to know what the causal structure is like before we can be confident we are applying the right decision theory, or applying it in the right way. And if we cannot get that information, and have to go *solely* on information about probabilistic dependence, then what we actually do, and not at all unreasonably, is reason that if something reliably happens after A, it happens because of A. Applying expected utility reasoning, when we only suspect causal influence, is like *post hoc, ergo propter hoc* reasoning. It is fallacious, but not necessarily a bad idea. That is to say, deciding on the basis of expected utility is a heuristic. It may be a good heuristic, but it is in any case a heuristic, not an algorithm. There are times when an expected utility calculation will be a misleading indicator of how much utility there is in an action.

7. Problems of Underdescription

There is a controversy regarding whether to be a one-boxer or a two-boxer. Why? Suppose you have the opaque box under your arm and are walking out the door, not knowing whether the box contains $1 million. Then someone runs up waiving a thousand-dollar bill. Is there any controversy concerning whether you should take the thousand? None at all. Or, if that isn't clear enough, suppose you have already opened the box and know exactly what is inside. Surely now there is no controversy. Take the thousand. At some point, you *know* that nothing can change the amount of money in the opaque box.

In the Newcomb case, what makes some people one-boxers is that it is not merely the *box* that is opaque. The whole *causal structure* is mysterious. Nozick can stipulate that one is in a situation where one's action does not affect what is in the box, but it is hard to believe, after all. And if all we have is Nozick's word, $1 million is a lot to bet on it. So, if there is an opaque box sitting on a stage at the far end of the auditorium, and we can hear someone thumping around underneath the stage, and we know the predictor moonlights as a stage magician, at some point it makes no sense, given the stakes, simply to accept at face value Nozick's assurance that there is nothing we can do to affect the probability of the box containing $1 million.

Lawrence Davis says the choice we face in Newcomb is simple: which is bigger: $1 million or $1,000?[14]

> One box = $1 million (probably)
> Two boxes = $1,000 (plus nothing, probably)

However, those who advocate choosing two boxes also say the choice is simple: do you want the extra $1,000 or not?

14. Davis (1977).

One box = Whatever is in the opaque box + $0
Two boxes = Whatever is in the opaque box + $1,000

Which of these straightforward choices is the real choice? Davis's reasoning seems compelling, until we see that it embodies an assumption contrary to Nozick's stipulation that your action cannot influence the situation's causal structure. Conditional probabilities in the Newcomb case create a curiously (but by no means uniquely) misleading appearance of influence that throws us off, making us less certain of the wisdom of taking both boxes.

Although Nozick himself was a two-boxer, he said the situation is different if it is stipulated that you know the predictor is *never* wrong. In that case, even Nozick finds compelling the case for taking one box. He may be wrong about that, though. To make the case compelling, we would need to know the causal mechanism that explains how it could happen that the predictor is never wrong. Here is one such mechanism. Someone puts the opaque box in your hand, and (as she has done with everyone else) she explains to you the predictor's secret. The secret is that the predictor does not have $1 million, and would not give you $1 million even if he had it to give, and therefore *never* puts $1 million in the box. You can be certain there is no money in the box you have under your arm, and by the way, do you want the $1,000? Like *everyone* before you, each having heard this explanation of why the Predictor is never wrong, you take the $1,000. This is the only scenario we know of, compatible with Nozick's stipulation of genuine causal independence, where we can make sense of the idea that the predictor is never wrong.[15] And in this scenario, contra Nozick, it is two-boxing that is compelling.

We undermine the one-box intuition when we stipulate that the opaque box's payoff is safely in your pocket and you are walking out the door as someone hands you an extra $1,000. Accordingly, a credible case for one-boxing would need to focus on situations where the causal structure is opaque, where there is a real chance (unlike in the scenario just mentioned) that if I choose one box, the predictor will have put $1 million in that one box, and where we reasonably can suspect we somehow influence the predictor. To *nurture* such suspicion in myself, what I would need to do, *prior to the prediction*, is to insist that I be put in a situation where I rationally lack confidence in Nozick's stipulation that I cannot influence the prediction. I will insist that the box be placed on a stage at the far end of the auditorium, on top of the magician's trapdoor! Although I do not know my choice is influencing the predictor, neither can I be sure of the contrary. Meanwhile, if all goes well, the predictor knows I respond to such ambiguity by playing expected utilities. The predictor puts $1 million in the box, and I take away $1 million, although not the extra $1,000.

So, in the Newcomb situation, do you know that Nozick's stipulation (that you cannot now influence the prediction) is true? That is the heart of the controversy.

15. What makes it true in our story that the predictor is never wrong would not support a subjunctive counterfactual of the form "But if someone were to choose one box, then the predictor would have predicted one box." In our story, p(predictor chooses 1 / you take 1) would equal zero rather than one, but since you never take one box, it does not affect the claim that the predictor is never wrong.

What does it take, given the stakes, for you to feel absolutely confident that one-boxers are reasoning from a false premise about the problem's causal structure?

8. Absolute Confidence

Nozick's original article launched a vast literature, much of which was a debate between one-boxers and two-boxers. The real payoff of this puzzle, though, is that it stands as something like an incompleteness theorem for decision theory.

In his later book, Nozick concludes: "It would be unreasonable to place absolute confidence in any one particular line of reasoning for such cases or in any one particular principle of decision."[16] If we had absolute confidence in dominance reasoning, we would take two boxes even if there were only a dollar in the transparent box. If we had absolute confidence in expected utility reasoning, we would take only one box even if there were close to a $1 million in the transparent box. Evidently, Nozick is right. We are *not* absolutely confident in any given decision procedure. Indeed, if we were, we would be irrational.

In Newcomb's problem, there is enough vagueness in the problem's description to create doubt regarding the situation's causal structure. If we tighten the description, we resolve the ambiguity about which procedure is called for. That tells us there is work to be done prior to the stage of decision-making addressed by decision theory. *Before* we bring a decision theory to bear on a given decision problem, there are decisions to make.

A decision theory is not the kind of thing we simply follow. A decision theory is something we decide to apply: something whose applicability we must evaluate before deciding to apply it. Our chosen principle is, after all, *chosen*. So, before we embrace dominance reasoning, or expected utility reasoning, what reasoning leads to our choosing that principle? There is a pre-decision-theoretic art to the partitioning with which decision theory begins.

At one level, Nozick's point (to answer the question we began with) is that Jeffrey's conditional probabilities are relevant to decision-making only insofar as conditional probabilities do, or at least may, indicate causal influence. More profoundly, what we seek in a decision theory is an algorithm, and that may not be the right thing to seek.

A decision theory can recommend that we try to partition world states so that the probabilities of the states do not causally depend on the choice of action. Then,

(1) if we have a dominant strategy, given a partition of world states such that the probabilities of the states do not causally depend on the choice of action, we should go with it;
(2) if we have no dominant strategy, given any such partition, we should use conditional probabilities to calculate expected utilities, and go with that.

Recommendation (2) assumes that conditional probabilities indicate causal influence. As Heraclitus once said, though, nature loves to hide. The world does not

16. Nozick (1993) 43.

wear its causal structure on its sleeve. Where causal structure is opaque, *post hoc* reasoning is a good heuristic. So, for the same reasons, is using expected utility calculations. But Nozick showed that such calculations are only a heuristic. We do not have an all-purpose decision-theoretic algorithm.

Originally published in The American Philosophers, *ed. P. French and H. Wettstein*, Midwest Studies in Philosophy *28 (2004): 282–94. Reprinted with permission of Blackwell Publishing. Sarah Wright is Professor of Philosophy at the University of Georgia*.