A Resource-Bounded Agent Addresses the Newcomb Problem

John L. Pollock
Department of Philosophy
University of Arizona
Tucson, Arizona 85721
pollock@arizona.edu
http://www.u.arizona.edu/~pollock

Abstract

In the Newcomb problem, the standard arguments for taking either one box or both boxes adduce what seem to be relevant considerations, but they are not complete arguments, and attempts to complete the arguments rely upon incorrect principles of rational decision making. It is argued that by considering how the predictor is making his prediction, we can generate a more complete argument, and this in turn supports a form of causal decision theory.

keywords: Newcomb problem, rational decision making, resource-bounded, causal probability, causal decision making.

1. The Newcomb Problem

You are feeling lousy. Your spouse left you for his/her secretary; your business partner has been arrested by the FBI for doctoring your books and defrauding your investors; your mother hates you; and your dog insists on peeing on your carpet. You are sitting in a bar reflecting morosely on the sorry state of the world, when a cheerful little old man sits down next to you. He introduces himself as an experimental economist, with a large grant from the Rockefeller Foundation for studying human decision making. "Cheer up, I have just the thing to make you feel better," he says. "I will give you a chance to win a million dollars, and it won't cost you a thing. Just meet me at my office tomorrow morning." So, come morning, you find yourself at the university, where your new friend puts you in the charge of two white-coated lab technicians. They lead you into a room where you are confronted with a table, and on the table are two boxes. One of the boxes, labeled "A", is transparent, and you can see that it contains a large amount of cash. You are informed that it contains $1000. The other box, labeled "B", is opaque, so you don't know what it contains. The technicians inform you that it is either empty, or contains one million dollars. The little old man who met you in the bar decided last night, after meeting you, whether to put the million dollars in it, and no one has tampered with it in the interim. You are now given a choice. You can either to take just box B, or you can take both boxes, and you get to keep the contents. This sounds like a no-brainer. Of course, you will take both boxes. But then the technicians inform you that there is a complicating factor. The little old man has the remarkable ability to predict, with great accuracy, whether people will take one box or two, and he decides how to load box B on the basis of his prediction. If he predicts you will take both boxes, he puts nothing in box B, but if he predicts you will take just the one box (i.e., box B), he will put a million dollars in box B. *Now* what should you do? Should you take one box or two?

You are given that the predictor is extremely accurate in his predictions. So as you reflect upon your choice, you realize that the probability is very high that box B will be empty if you take both boxes, in which case you will get only $1000. On the other hand, the probability is high that box B will contain one million dollars if take only box B, in which case you will walk about with one million dollars. You have read all about making decisions on the basis of expected values. The expected value of taking both boxes is only slightly more than $1000, whereas the expected value of taking only one box is only slightly less than one million dollars. Surely, then, you should take only box B.

But then you think about it a little more. Whatever money is going to be in box B is already there. The predictor put it there last night. So what you choose now can have no effect on what is in box B. If there is a million dollars in box B, and you take only the one box, you will get one million

dollars, but if you take both boxes you will get one million plus one thousand dollars. On the other hand, if there is nothing in box B, and you take only the one box, you will end up with nothing, whereas if you take both boxes you will take home $1000. In either case, you will get $1000 more if you take both boxes. So shouldn't you take both boxes? This is the "dominance argument".

This is odd. We have one argument that recommends taking just box B, and another argument that recommends taking both boxes. Each argument seems initially compelling, and each is based upon a familiar form of reasoning in decision theory, but they recommend opposite behaviors. This is *The Newcomb Problem*. (Nozick 1969)

Presented with the Newcomb Problem, philosophers split into two groups — the *one-boxers* and the *two-boxers*. Each insists that their position is based upon obviously correct general principles of practical reasoning. Each group is adamant that their solution is the correct one, and no rational person could disagree with them. The one-boxers often attempt to support their position further by observing that one-boxers will, on the average, go home with almost one million dollars, while two-boxers will tend to go home with only $1000. They charge, "If you are so smart, why ain'cha rich?" (See Gibbard and Harper 1978, and David Lewis, 1981, both of whom reject the argument). But the two-boxers retort that the one-boxers are subject to "chooser's remorse". After choosing to take just one box, and finding out what its contents are, they can always reflect that had they taken both boxes, they would have gotten an extra $1000. By rejecting box A, they just threw away $1000.

Both groups insist that their reasoning is based on impeccable general principles of rational decision making and so they must be right. The expected value argument is based on the familiar *optimality principle* according to which, when choosing between alternative actions, rationality dictates choosing an action that maximizes expected value. This principle is taught to students in every introductory economics or decision theory class. The dominance argument is based on the *dominance principle* according to which, if there is some condition *C* such that, if in condition *C* you do *A* rather than doing *B* you will get a larger payoff, and if in condition ~*C* you do *A* rather than doing *B* you will also get a larger payoff, then you should do *A* rather than *B* regardless of whether *C* is true. If it can be correctly claimed that either of these principles is intuitively obviously correct, then in using it the one-boxer or two-boxer would seem to have an impeccable argument for his choice. Unfortunately, both principles are actually false, and rather clearly so, for the reasons I will now discuss.

1.1 Why the Optimality Principle Fails

The optimality principle says that if we are choosing between two actions *A* and *B*, we should choose *A* in preference to *B* iff the expected-value of doing *A* is higher than the expected-value of doing *B*. This principle is widely endorsed and widely taught, and many philosophers seem to think that it is intuitively obviously correct. Unfortunately, there can be no real doubt that the principle is false, and because of that I suspect that we do not have any clear intuitions about it one way or the other until it is taught to us. When it comes to constructing theories of rationality, in either practical reasoning or epistemic reasoning, our intuitive assessments of concrete cases provide our data, and then we seek general principles that capture our intuitions. Although the optimality principle seems correct in simple cases, it is easy to contrive complex cases in which it recommends intuitively incorrect choices. For a general discussion of this, see Pollock (2006). Here I will just give a brief sketch of a couple of the ways in which the optimality principle goes wrong.

One kind of case in which it fails occurs when we are uncertain whether we will be able to perform the actions we are choosing between. Suppose, for example, that we are choosing between staying home and watching TV, or going to a movie at the theatre. You would enjoy the movie more than watching TV, so the optimality principle prescribes going to the movie. But suppose that the only way to get to the theatre is to take a bus. There has been talk of a bus strike, but you do not know the outcome. You estimate that there is only a 50% chance that the buses are running and hence only a 50% chance that you will be able to go to the movie. This is surely relevant to your decision whether to stay home and watch TV or go to the movie, but it is a factor that is ignored in computing the expected values of performing the two actions. The expected value computation can be made more complicated so as to accommodate such considerations (Pollock 2003, 2006), but the optimality principle gives the wrong answer in its simple form.

A second and more serious way in which the optimality principle can give the wrong answer is that it evaluates actions in isolation, but to make sensible decisions we must often consider combinations of actions rather than individual actions (Pollock 2005, 2006). Suppose you have two decisions to make. You have to go to the bank, and you also want to go to lunch. You must decide whether to go to the bank before or after lunch, and also where to go to lunch. The optimality

principle evaluates actions one at a time and has you choose them individually on the basis of their being optimal. The problem is that decisions can interact. Carrying out one decision may alter the probabilities and utilities involved in another decision, thereby changing what action is optimal. It could be that, prior to deciding where to go to lunch, because you are very hungry the optimal decision would be to postpone going to the bank until after lunch. But if you decide to have lunch at a restaurant far from the bank and you have other things to do in that part of town that could occupy you for the rest of the afternoon, this may make it better to go to the bank before lunch. Alternatively, because you are very hungry and want to eat before going to the bank, it might be better to choose a different restaurant. The point is that actions can interfere with one another, with the result that if several actions are to be chosen, their being individually optimal does not guarantee that the group of them will be optimal. This strongly suggests that the object of decision-theoretic evaluation should be the entire group of actions rather than the individual actions.

This same conclusion can be defended in a second way. Often, the best way to achieve a goal is to perform several actions that achieve it "cooperatively". Performing the actions in isolation may achieve little of value. Again, we must choose actions in groups rather than individually. To illustrate, suppose we have a table suspended from the ceiling by cables, and it is laid with expensive glassware. We can raise the right side of the table by activating a servomotor that retracts the cable on the right, and we can raise the left side by activating a different servomotor. We want to raise the table. Activating the right servomotor by itself would raise only the right side of the table and so spill the glassware onto the floor. Thus it has a negative expected value. Similarly for activating only the left servomotor. What we must do is activate both servomotors. That has a positive expected value even though it is composed of two actions having negative expected values. This illustrates again that actions cannot always be considered in isolation. Sometimes decision-theoretic choices must be between groups of actions, and the performance of a single action becomes rational only because it is part of a group of actions whose choice is dictated by practical rationality.

The two examples illustrate two different phenomena. In the first, actions interfere with each other, changing their execution costs and hence their expected values from what they would be in isolation. In the second, actions collaborate to achieve goals cooperatively, thus changing the expected values by changing the probabilities of outcomes. These examples might be viewed as cases in which it is unclear that actions even have well-defined expected values in isolation. To compute the expected value of an action we must take account of the context in which it occurs. If the expected values are not well-defined, then the optimality principle cannot be applied to these decision problems. Alternatively, if we suppose that the expected values of the actions in isolation are well-defined, then what is important about these examples is that in each case we cannot choose the group of actions by choosing the individual actions in the group on the basis of their expected values. In the first example, the expected value of the group cannot be computed by summing the expected values of the actions in the group. In the second example, the members of the group would not be chosen individually on their own strength. In these examples, it is the group itself that should be the object of rational choice, and the individual actions are only derivatively rational, by being contained in the rationally chosen group of actions. I put this in my (2005, 2006) by saying that the proper objects of decision making are plans rather than individual actions. In simple cases the plans may consist of single actions, but in complex cases rational decision making is more complicated than the optimality principle recognizes.

It might be supposed that we can repair the optimality principle by simply applying it to plans rather than actions. This is the procedure followed by most theories of decision-theoretic planning in artificial intelligence. The proposal would be that it is rational to adopt a plan iff there is no competing plan with a higher expected-value, and it is rational to choose an action iff it is prescribed by a rationally adopted plan. Let us call this *plan-based decision theory*. Regrettably, plan-based decision theory faces some insurmountable logical problems. The first is that plans are logical entities of potentially unbounded complexity. Plan-based decision theory would have us survey and compare all possible plans in order to determine whether they compete with a given plan and, if they do, to determine whether they have a higher expected value. But this is an impossible task. No real agent can consider all possible competitors to a given plan, so he cannot make decisions in accordance with plan-based decision theory.

The first problem is devastating enough, but it is worth noting that there is a second problem (taken from my 1992). Even if we could somehow survey and compare an infinite array of plans, plan-based decision theory would not yield rationally correct decisions. Plan-based decision theory is simply wrong as a theory of rational choice. This arises from the fact that for any plan there will

almost always exist a competing plan with a higher expected value. To illustrate, suppose that I am choosing between roasting chicken and barbecuing lamb chops for dinner. Suppose the former has the higher expected value. This implies that the plan of barbecuing lamb chops for dinner is not rationally adoptable, but it does not imply that the plan of roasting chicken for dinner is adoptable, because some other plan with a higher expected value may compete with it. And we can generally construct such a competing plan by simply adding steps to the earlier competing plan. For this purpose, we select the new steps so that they constitute a subplan aimed at achieving some valuable unrelated goal. For instance, we can consider the plan of barbecuing lamb chops for dinner and then later going to a movie. This plan still competes with the plan of roasting chicken for dinner, but it has a higher expected value. Thus the plan of roasting chicken for dinner is not rationally adoptable. However, the competing plan is not rationally adoptable either, because it is trumped by the plan of roasting chicken for dinner and then later going to the same movie.

It seems clear that given two competing plans P_1 and P_2, if the expected value of P_1 is greater than that of P_2, the comparison can generally be reversed by finding another plan P_3 that pursues unrelated goals and then merging P_2 and P_3 to form P_2+P_3. If P_3 is well chosen, this will have the result that P_2+P_3 still competes with P_1 and the expected value of P_2+P_3 is higher than the expected value of P_1. If this is always possible, then there are no optimal plans and simple plan-based decision theory implies that it is not rational to adopt any plan.

In an attempt to avoid this problem, it might be objected that P_2+P_3 is not an appropriate object of decision-theoretic choice, because it merges two unrelated plans. However, we often merge plans for unrelated goals. If I plan to run two errands (aimed at achieving two unrelated goals), and both errands require me to go in the same direction, I may merge the two plans by running both errands on a single trip.

The inescapable conclusion is that the rational adoptability of a plan cannot require that it have a higher expected value than all its competitors. The problem is that plans can have rich structures and can pursue multiple goals, and as such they are indefinitely extendable. We can almost always construct competing plans with higher expected values by adding subplans pursuing new goals. Thus there is no way to define optimality so that it is reasonable to expect there to be optimal plans. Consequently, simple plan-based decision-theory fails.

The only obvious way to avoid this problem is to apply the optimality principle exclusively to *universal plans*, which prescribe courses of action for all possible situations for the duration of the agents existence. For example, Savage (1954) toys with this idea. But again, it is not a possible decision principle for realistically resource bounded agents. No real agent could either construct or survey and compare all universal plans (which are infinitely complex, and of which there are infinitely many).

There are more epicycles to be followed in the search for a correct theory of rational decision making that takes account of the fact that actions must, in general, be chosen as parts of plans, and also that expected-values are presumably somehow relevant but plans cannot be chosen simply by comparing their expected-values to those of all their alternatives. These epicycles are pursued in my (2006) where a theory of decision-theoretic planning is proposed that is claimed to accommodate all of these observations. However, what is important for present purposes is just that the optimality principle, which seemed initially like a correct general principle of practical decision making, is incorrect. I doubt that we tend to endorse it because we find it intuitively obvious. Rather, it is something we are taught, and its credibility derives from the fact that it gets most simple cases right. But there are many ways in which it can get complex cases wrong. The failures I have illustrated here do not bear directly on the Newcomb Problem, but the Newcomb Problem is complex in its own ways, and having seen that the optimality principle can fail in complex cases, it would be naïve to simply insist that it must be right in the case of the Newcomb Problem. In fact, causal decision theorists think they know what is wrong with the optimality principle as applied to the Newcomb Problem. I will say more about that shortly. Formulating a correct theory of rational decision making is a philosophical problem, and the ultimate test of a theory is that it conforms to our intuitions in clear cases. The Newcomb Problem is not a clear case. There are many smart philosophers in each camp. So we cannot resolve this issue by appealing to bare intuitions, and we cannot resolve it by appealing to the optimality principle either, because it is broken and needs fixing, and without further argument it is not clear how a properly repaired optimality principle would apply to the Newcomb Problem. We can propose repairs, but we cannot evaluate them by appealing to how they resolve the Newcomb Problem, because it is not intuitively obvious how the Newcomb Problem should be resolved.

1.2 Why the Dominance Principle Fails

Let us turn then to the dominance principle. Is it any better off? First, let us formulate it precisely. Let the *conditional expected-value* of an action EV(A/C) be the expected-value computed using probabilities conditional on C being true. The dominance principle then proposes that if EV(A/C) > EV(B/C) and EV($A/\sim C$) > EV($B/\sim C$), then you should choose to do A rather than B. It turns out that the dominance principle can be derived from the optimality principle in the case in which the condition C is probabilistically independent of actions A and B, i.e., when PROB(C/A) = PROB(C), and PROB(C/B) = PROB(C). But when C is not probabilistically independent of A and B, then the dominance principle can disagree with the optimality principle. That is what is happening in the Newcomb Problem. There, the condition C is the condition that box B contains one million dollars. The dominance principle is applied by arguing that if box B contains one million dollars then you are better off taking both boxes, and if box B does not contain one million dollars then you are still better off taking both boxes. However, whether there is a million dollars in box B is not probabilistically independent of your taking both boxes. Your taking both boxes makes it less probable that box B contains one million dollars.

Should we endorse the dominance principle even in cases in which C is not probabilistically independent of the actions we are choosing between? There are cases in which we clearly should not. Consider a case that is a bit like the Newcomb Problem, but in this case box B is not loaded until after you make your decision. If you decide to take both boxes, nothing is put in box B, but if you decide to take only box B then one million dollars is put in box B. I take it that it is uncontroversial that in this case you should take only box B. But let C be the condition "There will be a million dollars in box B." It is still true that if C is true, you will be better off taking both boxes, and if C is false you will be better off taking both boxes. So the (unrestricted) dominance principle will recommend taking both boxes, and that is clearly incorrect. This illustrates that dominance reasoning often does not work in cases in which C is not probabilistically independent of the actions. And the independence condition fails in the Newcomb Problem, so this is a reason for being suspicious of the argument for two-boxing.

To summarize the discussion thus far, we first noted that we cannot resolve the Newcomb Problem by appealing to our unaided intuitions. Although many people have strong intuitions about whether they should take one box or two, the fact that many intelligent people disagree with them should give them pause. To resolve this issue satisfactorily, we are going to have to muster arguments. Unfortunately, the standard arguments for either one-boxing or two-boxing are flawed. The considerations to which they appeal seem intuitively relevant, but the general principles used to fill out the arguments are false. So we must look further for good arguments.

2. Causal decision theory

Because they were antecedently convinced (by their intuitions) that two-boxing was the rational response to the Newcomb Problem, Gibbard and Harper (1978) looked for a way of modifying the optimality principle so that it recommends two-boxing. Thus was born causal decision theory. As a number of authors (Gibbard and Harper 1978; Sobel 1978, 1994; Skyrms 1980, 1982, 1984; Lewis 1981a) have observed, conditional probabilities can reflect either evidential connections or causal connections. In the Newcomb Problem, choosing two boxes raises the probability that box B is empty, but it cannot *cause* box B to be empty, because the contents of box B were already fixed prior to the decision to take two boxes. The probabilistic connection is only an evidential one. Choosing both boxes gives you reason to expect that box B is empty, but does not in any sense "make it true". Presented with this distinction, a number of philosophers have had the intuition that it is only causal connections that should be relevant to decision making. As Joyce (1998, pg. 146) remarks, "Rational agents choose acts on the basis of their *causal efficacy*, not their auspiciousness; they act to *bring about* good results even when doing so might betoken bad news."

Those who are moved by these considerations try to find a way of distinguishing between informational and causal probabilities. Then the claim is made that the appropriate probabilities to use in the optimality principle are causal probabilities. Decision theory based on causal probabilities is *causal decision theory*. Because choosing two boxes does not (it is claimed) raise the causal probability of box B being empty, the expected value of taking two boxes is higher than the expected value of taking only box B, and hence a properly formulated optimality principle will recommend taking two boxes.

Furthermore, as remarked above, the *restricted* dominance principle (which requires the condition C to be probabilistically independent of the actions) is derivable from the optimality principle. If we formulate the optimality principle in terms of causal probabilities, then we get a version of the dominance principle that appeals to causal probabilities. Again, because choosing two boxes does not (it is claimed) lower the causal probability of box B containing one million dollars, the independence condition is satisfied if we let C be the condition that box B contains one million dollars. Hence the intuitively appealing dominance argument goes through. On the other hand, in the example in which box B is loaded after you decide whether to take two boxes, and the contents are determined by your decision, the causal independence condition fails. So in that case an appeal to dominance reasoning violates the causal independence condition.

These results are congenial to the two-boxer, but they are basically question-begging. The appeal to causal probabilities is motivated by the assumption that two boxing is the rational choice, but that is exactly what is at issue. It is unsatisfactory to defend that by appealing to one's intuitions and then looking for a principle that conforms to those intuitions, because the intuitions are not generally shared. If the appeal to causal probabilities is to have any force, it must be independently motivated, without simply assuming that rationality dictates two-boxing.

To this end, it has become common to switch to the Smoking Gene Problem and use that to motivate the need for causal decision theory. This problem is due to Stalnaker (1978). Suppose you are deciding whether to smoke. Suppose you know that smoking is pleasurable, and harmless. However, there is also a "smoking gene" present in many people, and that gene both (1) causes them to desire to smoke and (2) predisposes them to get cancer (but not by smoking). Smoking is evidence that one has the smoking gene, and so it raises the probability that one will get cancer. Getting cancer more than outweighs the pleasure one will get from smoking, so when expected values are defined in terms of "classical probabilities" that mix informational and causal connections, the optimality principle recommends not smoking. But this seems clearly wrong. Smoking does not *cause* cancer. It is just evidence that one already has the smoking gene and hence may get cancer from that. If you have the smoking gene, you will still have it even if you refrain from smoking, so the latter will not prevent your getting cancer.

Unlike the Newcomb Problem, most people seem to share the intuition that rationality dictates smoking in the Smoking Gene Problem. The strategy is then to take this to support causal decision theory, and then use causal decision theory, either via the *causal optimality principle* (the optimality principle formulated in terms of causal probabilities) or the (causally restricted) dominance principle, to argue that one should be a two-boxer. There are some details to be worked out here regarding how causal probabilities are to be understood, but the general approach seems promising.

There is, however, a fly in the ointment. Terry Horgan (1981) asks how the smoking gene works its magic to get its possessors to smoke. The natural hypothesis is that it creates a desire for smoking. But if one desires smoking, that is something one can know by introspection. And if you desire smoking, that makes it likely that you have the smoking gene. Furthermore, if you have the smoking gene, that "screens off" your smoking from affecting the probability of your getting cancer. That is, where PROB is classical (mixed informational and causal) probability, PROB(cancer/smoking & gene) = PROB(cancer/gene). Thus, *given that you have the smoking gene*, the classical probability of getting cancer is the same whether you smoke or not, and hence the classical expected value of of smoking is higher than the classical expected value of not smoking (because you get pleasure from smoking). Thus, Horgan argues, classical decision theory makes the same recommendation as causal decision theory. We cannot defend causal decision theory by appealing to the Smoking Gene Problem.

It is worth making the mathematics precise here. Let us assume that the gene gets people to smoke by giving them the desire and in no other way, so

(SG1) $\mathrm{PROB}(S/G\&D) = \mathrm{PROB}(S/D)$.

Also, for getting cancer, G screens off any combination of D and S, i.e.,

(SG2) $\mathrm{PROB}(C/G\&(\sim)D\&(\sim)S) = \mathrm{PROB}(C/G)$

and

(SG3) $\mathrm{PROB}(C/\sim G\&(\sim)D\&(\sim)S) = \mathrm{PROB}(C/\sim G)$.

((SG2) and (SG3) are each short for four different principles, where the tildes in parentheses can be either present or absent.) Letting U(Pl) be the utility of the pleasure incurred by smoking, U(*C*) be the utility of having cancer (a negative number), the expected-values conditional on having the desire are:

$$\mathrm{EV}(S/D) = \mathrm{U}(\mathrm{Pl}) + \mathrm{U}(C) \cdot \mathrm{PROB}(C/S\&D)$$

$$\mathrm{EV}(\sim S/D) = \mathrm{U}(C) \cdot \mathrm{PROB}(C/\sim S\&D).$$

The following is a theorem of the probability calculus (proof in the appendix):

Theorem 1: If (SG1), (SG2) and (SG3) then $\mathrm{PROB}(C/S\&D) = \mathrm{PROB}(C/\sim S\&D)$.

It then follows that $\mathrm{EV}(S/D) > \mathrm{EV}(\sim S/D)$, i.e., given that you have the desire, the classical optimality principle prescribes smoking. Thus, as Horgan claims, we cannot use the Smoking Gene Problem, construed in this way, to support causal decision theory.

This is dire news for the two-boxer, who is deprived of his favorite argument for causal decision theory. Is there any way around Horgan's objection? I think there is. It turns on the assumption that the smoking gene gets people to smoke by giving them the desire to smoke and then feeding that desire into the ordinary machinery of rational decision making. However, not all action is the result of rational decision making. If you inadvertently rest your hand on a hot burner on the stove, you will quickly jerk it back. You do not think about it and decide to do that. It is a reflex action. Could the smoking gene work like that? Rather than giving you the desire to smoke, it might make you smoke without any deliberation on your part. When you are confronted with cigarettes, you would have a tendency to just automatically pick them up and smoke them, without thinking about the matter. You might not even notice that you were doing it.

If this seems far fetched, notice that many people seem to eat sweets in roughly that way. If they habitually keep a bowl of chocolates on their desk, they may dip into the bowl and eat a chocolate without thinking about it. It seems likely that much of the smoking by habitual smokers works similarly. Often they do not really deliberate about whether to smoke. While deep in thought about other matters, they may just mechanically reach in their pocket for a cigarette. No doubt these are learned behaviors rather than genetically determined responses, but this does illustrate that actions need not be the result of deliberation and desire. In lower animals, there seem to be rich arrays of complex behaviors that are genetically determined. These are called "tropisms". For example, the "tarantula hawk" is a very large wasp that lives in the desert southwest and reproduces by stinging and paralyzing a tarantula, laying its eggs in the living tarantula, and burying the tarantula in the sand. When the larvae hatch, they feast on the still-living tarantula. The process by which the wasp stings the tarantula, lays its eggs, and then buries the spider, is a complex hardwired behavior pattern, apparently involving no deliberation on the part of the tarantula. In particular, the wasp need not have a desire to sting a tarantula.

It is unclear whether humans exhibit similar tropisms, but they might, and smoking might be among them for those who have the cancer gene. Suppose it is. We can observe someone smoking and ask whether he is doing what he rationally ought to do. Of course, if his behavior is the result of a tropism, then his smoking is not really a voluntary (or wholly voluntary) action, but we can still ask whether his tropism is making him do something that he rationally should not do. Let us call this the *Smoking Tropism Problem*. I find that it is like the standard Smoking Gene Problem in that most people have the intuition that, because the smoker gets pleasure from smoking and smoking does not cause cancer, the rational thing to do is smoke. Is this then a counter-example to the classical optimality principle, and can it be used to support causal decision theory?

In the Smoking Tropism Problem, although smoking raises the probability of getting cancer, that is screened off if you have an independent reason for thinking that you have the gene. In the standard Smoking Gene Problem, Horgan argued that we do have an independent reason for that, in the form of our introspecting that we desire to smoke. Do we have a similar reason for thinking we have the gene in the Smoking Tropism Problem? There is something unusual about the person who smokes because of the tropism, viz., he smokes without having a desire to smoke. That one smokes without having the desire to smoke makes it more probable that one has the gene. If one is a habitual smoker and has observed in the past that he smokes without having the desire to smoke, that gives him a reason for thinking he has the gene, and this suggests that this example will also be subject to Horgan's objection. However, the mathematics no longer works in the same way. Let *SD*

be "I have frequently smoked without having the desire to smoke". Then the analogues of (SG1), (SG2), and (SG3) above are:

(ST1) $\textsf{PROB}(S/G\&SD) = \textsf{PROB}(S/SD)$.

(ST2) $\textsf{PROB}(C/G\&(\sim)SD\&(\sim)S) = \textsf{PROB}(C/G)$

(ST3) $\textsf{PROB}(C/\sim G\&(\sim)SD\&(\sim)S) = \textsf{PROB}(C/\sim G)$.

(ST2) and (ST3) are still true, but there is no reason to think that (ST1) will be true. *SD* raises the probability of smoking by raising the probability of *G*, but knowing for sure that *G* is true will raise it more. After all, habitual smokers often smoke without thinking about it, and presumably without having an introspected desire to smoke. So the argument no longer goes through. In fact, we have the following theorem (proven in the appendix):

Theorem 2: If (ST2) and (ST3) hold and $\textsf{PROB}(S/G\&SD) > \textsf{PROB}(S/SD)$, then $\textsf{PROB}(G/S\&SD) > \textsf{PROB}(G/\sim S\&SD)$

If the difference between these probabilities is large enough, then the classical optimality principle will dictate not smoking, but that seems to be the wrong answer.

We can strengthen the argument further by focusing on the first time a person smokes. At that point he has no record of smoking, and so does not know whether he will end up smoking without having a desire to smoke. Thus he has no reason to think he has (or does not have) the tropism. If we ask, before he actually smokes, whether he should smoke, the answer seems to be that he should, because he will get pleasure from it and there will be no adverse consequences. But it is still true that his smoking will make it more likely that he will get cancer. So this seems to be a robust counter-example to the classical optimality principle, and it seems to support some variety of causal decision theory.

Before leaving the Smoking Gene Problem, let me make a final observation. Real (resource-bounded) agents cannot perform all possible reasoning in a finite amount of time. If they cannot do that, then that is not what they *should do*, and they are not behaving irrationally by not doing so. As reasoning progresses, what they should do changes to reflect how much reasoning they have performed. Philosophers have often been tempted to say that only ideal agents can be truly rational, because true rationality requires an agent to complete all relevant reasoning. One could, of course, *define* rationality that way if one so desired, but for most purposes that seems not to be the concept that interests us. For example, what I want to know in the case of the Newcomb Problem is whether *I* should take one box or two, and that is a question about me as a resource-bounded agent. For resource-bounded agents, we must distinguish between conclusions the agent *is justified* in drawing, given the current state of his or her reasoning and deliberation, and the conclusions that the agent *would be justified* in drawing if he or she could do all possible relevant reasoning. Let us say that the latter conclusions are *warranted* (Pollock 1986, 1995). This distinction has an important bearing on the Smoking Gene Problem. Horgan's argument shows that the *warranted* choice is to smoke. This conclusion depends on observing that (SG1) – (SG3) are true and noting that theorem 1 holds. However, most people who think about the Smoking Gene Problem are completely unaware of theorem 1, and would probably not endorse (SG1) – (SG3) without thinking about them for a bit. Nevertheless, they have the intuition that one ought to smoke. So that intuition does not depend on Horgan's argument. It is really an intuition about what the *justified* choice is, given what they have so far seen about the problem. The classical optimality principle cannot capture this intuition. Instead, it implies that until one has carried out the reasoning in Horgan's argument, the only justified conclusion is that one should not smoke. That, however, seems intuitively wrong to most people. So once we make the justified/warranted distinction and realize that Horgan's argument is about the warranted choice but most people's intuitions are about a ratiocinatively earlier justified choice, the original Smoking Gene Problem still provides a counter-example to the classical optimality principle, and suggests that one should instead endorse some version of causal decision theory. The point is that the debate is not just about what the warranted choice is, but also about how to reason about problems like this. A correct theory of decision-theoretic reasoning must track the justified beliefs of real resource-bounded agents, not just the warranted conclusions of an ideal agent.

3. Back to the Newcomb Problem

3.1 Rational Disagreement

The appeal to the Smoking Gene Problem was supposed to support two-boxing by supporting causal decision theory. However, many one-boxers remain recalcitrant, even if they agree that smoking is the rational behavior in the Smoking Gene Problem. They insist that the Newcomb Problem is different from the Smoking Gene Problem, even if they are unable to articulate what the difference is. We can attempt to argue otherwise by appealing to the optimality principle reformulated in terms of causal probabilities, but that is not entirely satisfactory. First, for the reasons given in section one, we know that the optimality principle is not really true, regardless of whether it is formulated in terms of causal probabilities or ordinary probabilities. It fails in various kinds of complex cases, so who is to say that it ought to work in the case of the Newcomb Problem?

Even if, in the end, we decide that we ought to take two boxes, there is an important difference between the Newcomb Problem and the Smoking Gene Problem. This is that almost everyone has the same intuitions regarding the Smoking Gene Problem, but the intuitions of smart people are split in the case of the Newcomb Problem. It is not enough to just argue that we ought to take two boxes unless we can also explain this split in intuitions. So let us look more directly at the Newcomb Problem and see if we can diagnose the source of this split.

There is an important difference between the two problems. In the Smoking Gene Problem, we understand the causal mechanism whereby smokers are prone to getting cancer. But in the Newcomb problem, the causal mechanisms are mysterious. We are told only that there is this almost magical predictor who can predict with great accuracy whether we are going to take one box or two. We are not told how the predictor makes his prediction, so we do not really understand how the money comes to be allocated in the boxes.

It might seem tempting to say that because we do not really understand the causal mechanisms in the problem, we are not sure what we should do, and so we should not make a choice. However, that is the worst thing we could do. If we opt out, we are guaranteed to get nothing, but if we make a choice, even an arbitrary and completely unmotivated choice, we are apt to come away with a substantial amount of money. So we should not opt out. Recognizing this, we think about it a bit, arrive at what seem like relevant considerations, and then we choose either one-boxing or two-boxing. Even if we have no idea what the best thing to do is, this is a win-win situation. We will come out ahead (at least in terms of expected-values), no matter what we do. And regardless of what happens, we will not be worse off than if we had not participated in the first place.

Compare the Newcomb Problem with a case in which you are presented with a panel of two buttons, and told that if you push one you will get $1000 and if you push the other you will get one million dollars, but are given no information about which button has which payoff. What should you do? Obviously, just pick a button and push it. One choice is better than the other, but you have no way of knowing which choice is better, so you should just make a choice randomly.

Consider a different case in which you have a complex decision to make, but you must make it quickly. You have all the relevant information, and you know all the relevant probabilities and utilities required for computing the expected-values of the various alternatives, but you do not have time to work out the mathematics. You have to decide *now* or forever lose your opportunity. What should you do? Again, it seems clear that you should make a choice randomly. If you do that, you may end up choosing the alternative with the lower expected-value, but you are not being irrational in doing this.

This reflects the "justified/warranted" distinction (Pollock 1986, 1995) discussed in section two. For resource-bounded agents, we can distinguish between conclusions the agent *is justified* in drawing, given the current state of his or her reasoning and deliberation, and the *warranted* conclusions that the agent *would be justified* in drawing if he or she could do all possible relevant reasoning. In the Newcomb Problem, most people are in the situation of having not seen their way through the problem clearly. They have thought of some relevant considerations and relevant arguments that incline them towards one-boxing or two-boxing, but as we have seen, those arguments are not decisive because they are based on principles that are not true in general. I suggest that if these decision-makers have done their best in trying to determine what they should do, then given that they definitely should not opt out of the problem, they are justified in whatever decision they make. But that does not imply that their decision is warranted. I take it that what the Newcomb Problem is actually about is which choice is the warranted one.

3.2 How Does the Predictor Do It?

With this understanding of how smart people can rationally arrive at different decisions in the Newcomb Problem, let us turn to how the predictor can be making his prediction. The reasoning I will now go through is reasoning any decision-maker could go through, and as we will see, it has consequences for whether the decision-maker should choose one box or two, so it bears on what the warranted choice is. We are talking about a "quasi-real-world problem" in the sense that the decision maker can assume that what is going on must be consistent with at least the broad contours of the way we think the world works. Accordingly, we can assume that the predictor is not a magician, but is relying upon objective cues to make his prediction. He must know about some objective property of decision-makers that strongly correlates with how they choose in the Newcomb Problem. This might be a complex of personality traits, or a gene, or early childhood experience, or whatever. It will not make any difference to our argument what it turns out to be, as long as there is some such objective property. Taking our cue from the Smoking Gene Problem, suppose it is a gene. There are different ways in which the gene could affect the decision-maker's choice. First, it could be that some decision-makers see clearly what the rational choice is and make it, while the gene interferes with the rational deliberation of other decision-makers and causes them to make an irrational choice. But in light of the previous considerations, I take it that this is extremely unlikely — virtually no one sees clearly what the rational choice is. Decision-makers are moved by various considerations they deem relevant, but they have not seen their way through the problem with complete clarity. They are "intuitive decision-makers", moved by intuitive considerations but not able to turn those considerations into complete arguments. In the case of intuitive decision-makers, it is much easier to imagine that a gene or personality trait could incline them to put more weight on one group of seemingly relevant considerations rather than the other and so become one-boxers or two-boxers. Let us explore this possibility first, and return later to other possibilities concerning how the gene could affect decision making. Suppose there is a dimorphism in a gene such that having one version of the gene — Newcomb-A — strongly inclines intuitive decision-makers to be moved more by the appeal to the optimality principle, and the other version of the gene — Newcomb-B — strongly inclines such people to be moved more strongly by the dominance argument. The gene is not making either group of decision-makers irrational — just inclining them towards different behavior when they have not completely solved the problem.

Now suppose you become interested in the Newcomb Problem for purely academic reasons. You are convinced that there must be some objective property to which the predictor is appealing in making his predictions, and so you break into his lab in the middle of the night and rummage through his files until you discover his secret. You learn about the gene, and then go home, your curiosity satisfied. So far, it is just an academic exercise in intellectual burglary. Then, to your surprise, you get a call offering you the opportunity to participate in the Newcomb Problem experiment. So, knowing the predictor's secret, you rush out to the nearest biological lab and get your DNA tested to learn which version of the gene you have. Suppose you discover that you have Newcomb-B. Knowing how the predictor makes his prediction, you can predict with confidence that he will not put any money in box B. Given that, it would be crazy to take just box B. Clearly, your only rational choice is to take both boxes. Suppose instead that you learn you have Newcomb-A. In that case you can confidently predict that the predictor will put one million dollars in box B. But you also know that there is $1000 in box A, so again it seems clear that you should take both boxes. To take only one box would be to throw away $1000 for no reason.

Notice that, so far, I have not said anything that should be controversial. The choices I have advocated as rational are exactly the choices classical decision theory would recommend via the optimality principle. But next, suppose it is expensive to get your DNA tested. Suppose it costs $500. It occurs to you that you are going to do the same thing regardless of the result of the testing, namely, take both boxes. So why waste the money on the test? Just take both boxes and be done with it. Surely, that is the rational thing to do. It would be foolish to spend $500 to be tested when you know beforehand that it will not affect your decision. No matter what you discover, you will (and should) take both boxes.

Notice that the structure of this version of Newcomb Problem, where you know how the predictor makes his decision but you do not know which version of the gene you have, is exactly parallel to the structure of the Smoking Gene Problem where you know how the smoking gene causes cancer but you do not have any information about whether you have the gene. In the latter, having the gene causes you to get cancer, and in the former, having the gene causes box B to be empty. Your deciding to smoke makes it more likely that you have the smoking gene, and so more likely you will get cancer, but that seems irrelevant because if you are going to have the gene, you

already have it regardless of whether you smoke. And your deciding to take two boxes makes it more likely that you have Newcomb-B and hence more likely that box B will be empty, but that should be equally irrelevant because if you are going to have the gene you already have it regardless of whether you take two boxes.

There is one difference between the Smoking Gene Problem and the Newcomb Problem that might affect people's intuitions. In the Newcomb Problem, unlike in the Smoking Gene Problem, the causal connection between having Newcomb-B and box B being empty passes through the predictor — a rational agent who, you can suppose, exhibits free will and so cannot really be part of a mechanical causal chain. Should this make a difference? I think not, because we can take the predictor out of the chain. Suppose that the predictor gets tired of having to go into the lab each evening to load the money into the boxes, so he rigs up a mechanical device that automatically (and surreptitiously) tests the subject's DNA for the gene, and then loads the boxes in response to the outcome of the test and without any human intervention. Then it is hard to see how there is any relevant difference between the causal mechanisms in Newcomb Problem and the Smoking Gene Problem. Surely, if it is rational to smoke, it is rational to take two boxes.

3.3 "Why Ain'cha Rich?"

What about the "If you are so smart, why ain'cha rich?" argument? The claim is that two-boxers go home with about $1000 on average, and one-boxers go home with one million dollars, so shouldn't one be a one-boxer? But notice that this is only true for decision-makers whose decisions are influenced, as predicted, by the Newcomb gene. If they choose two boxes because they have Newcomb-B, they will on the average get about $1000, and if they choose one box because they have Newcomb-A, they will on the average get about one million dollars. However, this does not really have anything to do with their choice. Box B contains whatever it does because they have the gene, not because they choose as they do. Furthermore, the correlation between what they choose and how much money is in box B only pertains to "intuitive decision-makers" who have not reasoned the problem out as above, and whose decision is being determined by the gene. What we might call "the rationally informed" decision-maker who reasons as above will take both boxes regardless of which version of the gene he has. So it is not true that decision-makers who take both boxes as a result of reasoning the problem out in this way will tend to get only $1000. On the contrary, if we suppose the two versions of the gene are equally distributed among decision-makers, the rationally informed decision-maker will on average receive $501,000.

Of course, those who do not reason it out and take one box because they have Newcomb-A will tend to do even better, averaging about one million dollars. But that is just because they are being rewarded for having Newcomb-A, not because they take one box. In fact, they would do better yet if they took two boxes. So this is not a reason for taking one box. Rationally informed decision-makers who possess Newcomb-A but reason the problem out fully as above and take two boxes do even better, averaging about $1,001,000.

Notice further that if a decision-maker is moved to take one box by the "Why ain'cha rich?" argument, then he is presumably not being influenced by the gene. Taking one box for that reason does not make it more probable that one has Newcomb-A, and hence does not raise the probability that box B will contain one million dollars. If the gene is equally distributed across decision-makers, and being moved by the "Why ain'cha rich?" argument is not correlated with the gene, then a person taking one box for this reason will, on the average, get $500,000 — $1000 less than the average rationally informed two-boxer.

3.4 Relaxing the Assumptions

The above reasoning is predicated on two assumptions. The first is that the gene only has an effect on the "intuitive" decision-makers who have not reasoned the problem out fully as above, and works by strongly biasing their intuitive responses to the optimality argument and the dominance argument. These decision-makers are influenced by considerations that seem to them to be relevant, but they do not really have good arguments for preferring one set of considerations to the other. The second assumption is that the decision-maker has come to know how the predictor is making his prediction, although he does not know which version of the gene he has. Consider what happens when we relax these assumptions.

Let us try relaxing the second assumption first. Suppose that, although we are rationally convinced that there is some objective property P of the decision-makers that the predictor is using to make his prediction and hence to load the boxes, we do not know what it is. Should this make any difference? It is hard to see why it would. Whatever the property P is, if we were to discover

that it is the property the predictor is using, we would reason as above to the conclusion that we should take both boxes. Suppose we do not know what property *P* is, but we could find out by bribing the predictor's lab assistant. Suppose we could do that for $500. Should we? Noticing that whatever we find out will not make any difference to our decision — we will rationally take two boxes in any case — it would be foolish to spend the money. So the identity of *P* seems to be irrelevant. Accordingly, deciding rationally to take two boxes does not seem to be dependent on knowing what property *P* is. Merely knowing that there is some such property *P* should be enough.

Thus far I am using the cost of finding out what *P* is or getting yourself tested for *P* as an intuition pump. I have not proposed a general principle of reasoning that prescribes two-boxing in this case. We are still at the stage of trying to collect cases regarding which we have clear intuitions. Once we have them, we can look for general principles that capture them. The virtue of this intuition pump is that people seem to be much more in agreement that they should take two boxes than they are in the unelaborated Newcomb Problem.

However, the above argument still depends on there being such a property *P*. Is there an alternative way the predictor could be making his predictions? If the predictor is really able to predict decision-makers' choices, there has to be some property *P* of the decision-makers that the predictor can determine before loading the boxes and which correlates with their choices. But might the correlation arise differently than I have suggested above? There are just three possibilities for how the property *P* comes to be correlated with your decision making. (1) It could bias your choice, but you can do more reasoning and override it. This is the case I have been discussing. (2) It could fully determine the choice of decision makers who have the property *P*, either forcing them to choose one box or forcing them to choose two boxes (whichever is the incorrect choice), but leave the chooser without the property free to see the rational thing to do it and then choose accordingly. (3) It could fully determine the choice of decision makers who have it, and lacking property *P* could fully determine the opposite choice of decision makers who lack it, so no one is acting voluntarily. In case (3) and in the involuntary part of case (2), the decision makers cannot reverse their choices in response to further arguments, but we can still ask whether they are being made to do things they should not do.

Consider case (2). Here, some of the decision-makers do, after all, see clearly what the warranted choice is and can give complete arguments. In light of the literature, this seems preposterous, but pretend it is true. One boxers and two boxers cannot both be seeing clearly what the warranted choice is, because they are in the same decision problem but producing different answers. So some must be getting it right and others getting it wrong. If those decision-makers who see clearly what the warranted choice is choose accordingly, then in order for the property *P* to be correlated with what decision-makers choose, it must be negatively correlated with whether they see clearly what the warranted choice is. Having the property *P* must prevent decision-makers from seeing what the warranted choice is, and lacking the property enables them to see what the warranted choice is. If the correlation is really strong, as it must be for the predictor to be so accurate, it must be that virtually everyone who lacks the property *P* sees what the warranted choice is.

This seems pretty dubious, because virtually no one is able to give a correct argument for either choice, so it seems that hardly anyone can be said to see clearly what the warranted choice is. They may make the choice that is in fact warranted, but not for wholly adequate reasons. But let us wave this objection and suppose things work as I have just described. Now consider two cases. Suppose first that one-boxing is the warranted choice. Then if the predictor determines that a decision-maker *lacks* property *P*, and so sees clearly that he should choose just box B, the predictor will put one million dollars in box B. Otherwise the predictor will leave box B empty. But now we can argue just as we did above for case (1). Suppose you discover how the predictor is making his prediction, and you get yourself tested for property *P*. If you learn that you have property *P*, you can be confident that the predictor has put nothing in box B, so clearly you should take both boxes. If you learn instead that you lack property *P*, you can be confident that the predictor has put one million dollars in box B, but you also know there is $1000 in box A, so again you should take both boxes. Again, it would be foolish to pay to get yourself tested for property *P*, or to find out what property *P* is, because what you find out will have no effect on what you should do. So, if we suppose that one-boxing is the warranted choice, it follows that two-boxing is the warranted choice. This is a reductio of the supposition.

If we suppose instead that two-boxing is the warranted choice (which it must be, in light of the previous argument), then it has to be the case that the predictor puts one million dollars in box B iff he determines that the decision-maker *has* property *P*. But again, if you get yourself tested and determine that you have property *P*, you should take both boxes, and if you find that you lack property *P*, you should also take both boxes. Again, it makes no sense to pay for information that will not affect your decision regarding what you should do, so you should take both boxes *simpliciter*. Of course, you cannot do that because your choice is fully determined by your having the property *P*, but nevertheless, that is what you should do. Hence this supposition is consistent — you should take both boxes.

Case (3) works just like the involuntary part of case (2). Again, you should take both boxes. If your having or lacking *P* determines that you will instead take one box, there is nothing you can do about that, but it is still true that you are being forced to make the wrong choice.

Thus far I have supposed that the predictor is highly reliable in his predictions, but not that he is infallible. That is the version of the Newcomb Problem that Nozick first formulated, and he proposed that one should take both boxes. But he also remarked in passing that if the predictor were somehow infallible in his predictions, then one should take just box B. That, however, seems to be wrong. If the predictor is infallible, then property *P* must be perfectly correlated with the decision-makers' decisions. The only way that could happen is if having or lacking property *P* causally determines what choice a decision-maker will make. Thus this is a version of case (3). Because reasoning, even incomplete reasoning, can normally change a decision-maker's mind, the causal link between *P* and the choice cannot pass through the decision-maker's normal deliberative processes. It must determine the choice in the same way the gene determines smoking in the Smoking Tropism case. If the decision-makers are thus being *made*, non-rationally, to choose as they do, then their choices are not voluntary actions. But we can still ask whether they are being made to do what they should do. And the preceding arguments still work. If they have property *P*, they should take both boxes, and if they lack property *P* they should take both boxes, so they should take both boxes. Hence, supposing that the predictor is infallible does not seem the change the problem in relevant ways.

Some philosophers have been tempted to respond that what I have described is not the standard Newcomb problem. In the standard formulation, the predictor can predict everything about your reasoning concerning whether to take one box or two, so if you reason as in sections 3.2 and 3.3, he will predict that and leave box B empty. But this does not really change anything. Perhaps there is an array of mutually exclusive and exhaustive properties (e.g., a gene multi-morphism rather than a dimorphism) that determine exactly how you will reason. There has to be some such set of properties if the predictor is not magical. In this case the properties cannot just bias your decision making, as in case (1), leaving you free to do some more reasoning and override the bias, because they must also correlate with your doing that further reasoning. So this must be a version of case (3). Although it may seem to you that you are making voluntary choices, in fact you are not. You choices are entirely determined by whichever property you have, and that may force you to take one box rather than two. But this does not change the argument. No matter which property you have, and hence no matter how you reason, you will do better taking two boxes. Consequently, it would not make sense to pay a lot of money to get yourself tested to see which property you have, because whatever you find out, you should still take two boxes.

4. A Horganesque Complication

I have argued that if we consider how the predictor could, possibly, be making his predictions, we are led inexorably to the conclusion that we should take two boxes. It is generally agreed, however, that the classical expected-value of taking two boxes is less than the classical expected-value of taking one box, on the grounds that the probability of there being a million dollars in box B is much higher given that one chooses one box than it is given that one chooses two boxes. So this seems to be a counter-example to classical decision theory and the optimality principle formulated in terms of classical probabilities. To many philosophers, it has seemed that the problem is that your choice has no causal influence on the contents of box B, and this motivates a search for some kind of causal probability that can be used in place of of classical probabilities. The suggestion is that if we reformulate the optimality principle in terms of causal probabilities, it will correctly prescribe taking two boxes. To work this out we must consider how causal probability is to be

understood, and we have to take seriously the observation that the optimality principle is not correct anyway when applied to complex cases. I will address these issues in section six.

But first, let us consider whether the preceding analysis of the Newcomb Problem really provides a counter-example to the classical optimality principle. Further reflection suggests a rejoinder. We can try to run a version of Horgan's argument (regarding the Smoking Gene Problem), applying it to the Newcomb Problem. Before we begin, let me emphasize that although this argument is motivated by Horgan's reasoning concerning the Smoking Gene Problem, Horgan himself does not give this argument regarding the Newcomb Problem, and Horgan is officially a one-boxer so the conclusion is at odds with his official view.

For concreteness, let us go back to the gene dimorphism version of the analysis. We have seen that if one *knows* either that one has Newcomb-A or that one has Newcomb-B, the classical optimality principle prescribes taking both boxes. If a person notes that he is an "intuitive two-boxer", that is, that he finds the dominance argument intuitively more compelling than the argument from expected-values, that gives him a reason for thinking that he has the Newcomb-B version of the gene, just as desiring to smoke gives the potential smoker a reason for thinking he has the smoking gene. Similarly, if one notes that he is an "intuitive one-boxer", that gives him a reason for thinking that he has the Newcomb-A version. Does this make it classically rational for either decision-maker to take two-boxes, and hence render this not a counter-example to the classical optimality principle? To run an argument analogous to Horgan's argument for the Smoking Gene Problem, we would need the following three premises:

(NB1) PROB(take-2-boxes/Newcomb-B & intuitive-2-boxer)
= PROB(take-2-boxes/intuitive-2-boxer).

(NB2) PROB(million/Newcomb-B &(~)intuitive-2-boxer &(~)take-2-boxes)
= PROB(million/Newcomb-B)

(NB3) PROB(million/~Newcomb-B &(~)intuitive-2-boxer &(~)take-2-boxes)
= PROB(million/~Newcomb-B).

These premises all seem true. (NB1) is true because Newcomb-B only inclines one to be a two-boxer by making one an intuitive two-boxer. (NB2) and (NB3) are true because the predictor loads the boxes simply on the basis of whether the decision-maker has Newcomb-B. Then, as above:

Theorem 3: If (NB1), (NB2) and (NB3) hold then PROB(Newcomb-B/take-2-boxes & intuitive-2-boxer) = PROB(Newcomb-B/~take-2-boxes & intuitive-2-boxer).

Hence for the intuitive two-boxer, having Newcomb-B is statistically independent of taking two boxes, and hence there being a million dollars in box B is independent of taking two boxes. But if you take two boxes you will also get the $1000, so the classical expected-value of taking two boxes is higher than the expected-value of taking one box. We can run the analogous argument for the intuitive one-boxer, so in either case classical decision theory tells us to take both boxes.

Let me make this argument more precise. The decision-maker who has reasoned as in sections 3.2 and 3.3 is justified in concluding that he should take two boxes, and would not be justified in concluding that he should take just one box. It was noted that PROB(million/take two boxes) is low but PROB(million/take one box) is high, and so if we use these probabilities in computing expected-values, EV(take two boxes) < EV(take one box). However, it was insisted that these probabilities are not causal probabilities. What box you take cannot causally influence whether there is a million dollars in box B, because box B was loaded before you make your decision and was not tampered with subsequently. Because there being a million dollars in box B is causally independent of your taking both boxes, the causal probability $\text{C-PROB}_{\text{take two boxes}}(\text{million}) = \text{PROB}(\text{million}) = \text{C-PROB}_{\text{take one box}}(\text{million})$. Hence if we compute the expected-values using only causal probabilities, EV(take two boxes) > EV(take one box) (because taking two boxes also gets you the $1000 in box A).

However, if the decision maker reasons further as in the present section, he finds further relevant probabilities that still lead him to conclude that he should take both boxes:

PROB(million/take two boxes & intuitive two-boxer)
= PROB(million)
= PROB(million/take one box & intuitive two-boxer)
= PROB(million/take two boxes & intuitive one-boxer)

$= \text{PROB}(\text{million}/\text{take one box} \ \& \ \text{intuitive one-boxer})$

and hence, regardless of whether you are an intuitive two-boxer or an intuitive one-boxer, EV(take two boxes) > EV(take one box). Furthermore, these classical probabilities agree with the causal probabilities, because we still have causal independence:

$\text{C-PROB}_{\text{take two boxes}}(\text{million}/\text{intuitive-two-boxer})$
$= \text{PROB}(\text{million})$
$= \text{C-PROB}_{\text{take one box}}(\text{million}/\text{intuitive-two-boxer})$
$= \text{C-PROB}_{\text{take two boxes}}(\text{million}/\text{intuitive-one-boxer})$
$= \text{C-PROB}_{\text{take one box}}(\text{million}/\text{intuitive-one-boxer}).$

So the causal decision theorist gets the same result for the expected-values. The two varieties of decision theory disagree when applied to the probabilities known in sections 3.2 and 3.3, but agree again when applied to the richer array of probabilities noted in the present section.

In computing the probabilities to use in decision making, one should take account of the most information you have about the situation, so given that the decision-maker knows whether he is an intuitive one-boxer or an intuitive two-boxer, he should use the probabilities that are conditional upon that further knowledge.

Thus far the Horganesque argument depends on the assumption that the gene (or whatever property *P* is) biases decision makers' responses to the considerations that seem intuitively relevant to the problem, making them intuitive one-boxers or intuitive two-boxers, but not in a way that cannot be overridden by further reasoning. Given that in the real world, most of the philosophers who argue about the Newcomb problem seem to have open minds to the extent that they are prepared to change their views given a sufficiently good counter-argument, it seems that this is the only way the gene could work in the real world. However, as above we can imagine situations in which the connection between the gene and the decision making is tighter, and the decision maker cannot override the causal effect of the gene by doing further reasoning. If the decision maker knows that the gene works this way, and he can tell by introspection (before making his decision) whether he is a one-boxer or a two-boxer, then he can run the same Horganesque argument and conclude that he ought to take both boxes.

I have argued that there are only three ways the gene (or property *P*) could come to be strongly correlated with decision makers' choices, and as above we can reason in each case that if the decision maker knows that is the case he is in then the rational choice is to take both boxes. But unless he can reason as above that he is in case (1), it seems likely that the decision maker will not know which case he is in. Then what should he do? Presumably, which case a decision maker is in is statistically independent of whether he chooses one box or two, so it follows from the version of the dominance principle that in turn follows from the classical optimality principle that the decision maker should choose two boxes.

What I have called "the Horganesque argument" is reminiscent of an argument due to Eells (1984), known in the literature as the "tickle defense". I will not go into details here, but Eells argues that a rational agent cannot believe that $\text{PROB}(\text{million}/\text{take two boxes}) < \text{PROB}(\text{million}/\text{take one box})$. Taken literally, this is clearly wrong. The intuitive two-boxers and one-boxers all believe this, and I argued above that they are rational. However, what Eells means by "rational" is "ideally rational in a Bayesian sense". That is, he assumes that rational agents have degrees of belief, and they are subjective probabilities in the sense that they are coherent, i.e., conform to the probability calculus. He also assumes that the probabilities at issue in the Newcomb problem as subjective probabilities. However, I do not accept these assumptions. I have argued in my (2006) that subjective probabilities do not make sense for real resource-bounded agents, and I do not think that any real cognizer can have coherent degrees of belief. The Bayesian notion of rationality is, at best, a notion of ideal rationality only applicable to ideal agents, not to real agents. However, we need not get involved in a dispute about that here. For present purposes, it is best to remain noncommittal about the kind of probability employed in decision-theoretic reasoning. The main point here is that although the Horganesque argument has a similar conclusion, it is quite different from Eells argument and not subject to the kinds of objections that have been raised to Eells argument (e.g., Sobel 1994).

5. Response to the Horganesque Argument

The Horganesque argument is extremely interesting. It shows that the decision-maker who reasons as in sections 3.2 and 3.3 has not, after all, reasoned the problem out fully. What I there called "the rationally informed" decision-maker turns out to be only partly rationally informed. He is more rationally informed than the intuitive one-boxer or the intuitive two-boxer, but he has not seen that more reasoning is possible. This additional reasoning still leads to the conclusion that the rational decision is two-boxing, but now it does so in a way that accords with the classical optimality principle. The intuitive one-boxer who defends his decision by appealing to the classical optimality principle should, in light of this further reasoning, change his stripes and become a two-boxer. This seems to me to be completely compelling. No matter what one's theoretical commitments, once one has reasoned the problem out this far, one should abandon one's one-box inclinations and become a two-boxer.

On the other hand, this diagnosis of the Newcomb Problem seems to undercut the support that the earlier reasoning appeared to provide for causal decision theory. But does it? In a surprising way, I think this further diagnosis actually supports causal decision theory. First, notice that we have independent support for causal decision theory from the Smoking Gene Problem (although we have not yet spelled out just how causal decision theory is going to go). Second, I argued that the decision-maker who reasons as in sections 3.2 and 3.3 should, rationally, take both boxes. What this new argument shows is that the decision to take both boxes on the basis of the arguments of sections 3.2 and 3.3 must be regarded as a justified conclusion rather than a warranted conclusion. It is not yet warranted because there is more relevant reasoning to be done. But, interestingly enough, the additional reasoning does not reverse the conclusion. When the additional reasoning is completed, we still end up with the conclusion that one should take both boxes. I take it that this is, if anything, further support for the contention that the decision-maker of sections 3.2 and 3.3 should choose both boxes. The argument there is still an appeal to intuitions. I find that the intuitions there are more widely shared than the intuitions mustered by the initial formulation of the Newcomb Problem, but a died-in-the-wool one-boxer might still deny those intuitions. I take it that in light of this further argument, that is no longer possible, and this lends further support for the claim that the intuitions of sections 3.2 and 3.3 were correct assessments of what conclusion the decision-maker who has gotten that far in his reasoning is justified in drawing. But if the decision-maker who got as far in his reasoning as sections 3.2 and 3.3 was justified in choosing two boxes, he was justified without having seen the truth of theorem 3, and perhaps without having realized that (NB1) – (NB3) are true. This shows that the classical optimality principle is not a correct principle of decision making. The correct reasoning that was being done there can be accommodated by some version of causal decision theory, but not by classical decision theory.

The debate is not just about what conclusion is warranted, but also about what the correct principles of reasoning are. What is crucial for present purposes is not that you ought to take both boxes, but that you should take both boxes even before you see the additional Horganesque probabilities. This is precisely analogous to the observation that in the Smoking Gene Problem you are justified in smoking even before you see the probabilities involved in Horgan's argument. These are both counter-examples to classical decision theory. That classical decision theory agrees with this assessment once you take account of the Horganesque probabilities does not undercut the counter-example. If anything it lends it further credence by making it indisputable that one should take two boxes.

6. Causal Decision Theory

Throughout this paper, I have alluded to causal probabilities. It is time to give some thought to how causal probabilities are to be understood. A number of different theories of causal probability have been proposed, but I think they are all more complex than necessary. I have discussed this at length in my (2002, 2006), so at this point I will just sketch my alternative theory. The probabilities that seem to lead to incorrect applications of the optimality principle are always "backtracking" probabilities. That is, they result from an action A making it probable that something P was true *in the past*, and P making it probable that something Q is true in the future. If the probabilities are high, then $\textsf{PROB}(Q/A)$ is high, but this does not seem to be relevant to deciding whether to perform

A. What is crucial here is that A is screened off by P, i.e., $\text{PROB}(Q/A\&P) = \text{PROB}(Q/P)$. The intuition is then that A does not contribute to making Q true.

Let us just consider the case in which Q is something that is true or false at an instant, e.g., the proposition that I have cancer at a specific time t. Suppose similarly that A is a "point-dated" action that occurs at a specific instant. In my (2002, 2006) I discuss how to relax these assumptions. My proposal is that in computing the causal probability of Q given A, we hold the past fixed. Let B be the conjunction of all past facts (i.e., facts about the world prior to the time A is to be performed) that are relevant to whether Q will be true. I will call *B the background*. If we knew B, then we could identify the causal probability $\text{C-PROB}_A(Q)$ with the classical probability $\text{PROB}(Q/B\&A)$. For example, in the Smoking Gene Problem the only past fact that is relevant to whether I get cancer is whether I have the smoking gene. Suppose I know that I do. Then $\text{C-PROB}_{\text{smoke}}(\text{cancer}) = \text{PROB}(\text{cancer}/\text{smoke} \,\&\, \text{gene}) = \text{PROB}(\text{cancer}/\text{gene})$.

Generally, we will not know the truth values of all conjuncts of B. For example, in the standard Smoking Gene Problem we do not know whether we have the gene. In cases like this, there are various possible backgrounds $B_1,\ldots,B_n$, and we do not know which is true. If we have a probability distribution over the possible backgrounds, we can define

$$\text{C-PROB}_A(Q) = \sum_{i\leq n} \text{PROB}(B_i) \cdot \text{PROB}(Q/A\&B_i).$$

The possible backgrounds in the Smoking Gene Problem are *gene* and *~gene*, so

$$\begin{aligned}\text{C-PROB}_{\text{smoke}}(\text{cancer}) &= \text{PROB}(\text{gene}) \cdot \text{PROB}(\text{cancer}/\text{smoke} \,\&\, \text{gene}) \\ &\quad + \text{PROB}(\sim\text{gene}) \cdot \text{PROB}(\text{cancer}/\text{smoke} \,\&\, \sim\text{gene}) \\ &= \text{PROB}(\text{gene}) \cdot \text{PROB}(\text{cancer}/\text{gene}) \\ &\quad + \text{PROB}(\sim\text{gene}) \cdot \text{PROB}(\text{cancer}/\sim\text{gene}) \\ &= \text{PROB}(\text{cancer}).\end{aligned}$$

Hence for causal probabilities, getting cancer is independent of smoking. Similarly, in the Newcomb Problem, there being a million dollars in box B is independent of the decision maker's choice. Thus if we formulate the optimality principle in terms of causal probabilities, it prescribes taking both boxes.

Of course, as we have seen, the optimality principle is not really true, regardless of whether it is formulated in terms of causal probabilities. It can make intuitively incorrect prescriptions in cases in which we may not be able to perform some of the alternative actions, or in cases in which actions cannot be chosen in isolation but must instead be chosen as parts of more comprehensive plans. The details of this are worked out in my (2006). However, in simple cases in which there is no question about being able to perform the actions, and they are independent of anything else we might do, it is plausible that the optimality principle yields the correct prescriptions if it is formulated in terms of causal probabilities. I think that this is the case for both the Smoking Gene Problem and the Newcomb Problem.

Appendix: Proofs of Theorems

Theorem 1: If (SG1), (SG2) and (SG3) then $\text{PROB}(C/S\&D) = \text{PROB}(C/\sim S\&D)$.

Proof: By the probability calculus and (SG1):

$$\text{PROB}(S\&G/D) = \text{PROB}(S/G\&D) \cdot \text{PROB}(G/D) = \text{PROB}(S/D) \cdot \text{PROB}(G/D).$$

By the probability calculus:

$$\text{PROB}(S\&G/D) = \text{PROB}(G/S\&D) \cdot \text{PROB}(S/D).$$

Thus

$$\text{PROB}(G/S\&D) = \text{PROB}(G/D).$$

Then by the probability calculus, $\text{PROB}(G/S\&D) = \text{PROB}(G/{\sim}S\&D) = \text{PROB}(G/D)$. Consequently, by (SG2) and (SG3):

$$\text{PROB}(C/S\&D) = \text{PROB}(C/S\&D\&G) \cdot \text{PROB}(G/S\&D) + \text{PROB}(C/S\&D\&{\sim}G) \cdot \text{PROB}({\sim}G/S\&D)$$

$$= \text{PROB}(C/G) \cdot \text{PROB}(G/D) + \text{PROB}(C/{\sim}G) \cdot \text{PROB}({\sim}G/D)$$

and

$$\text{PROB}(C/{\sim}S\&D) = \text{PROB}(C/{\sim}S\&D\&G)\cdot\text{PROB}(G/{\sim}S\&D) + \text{PROB}(C/{\sim}S\&D\&{\sim}G)\cdot\text{PROB}({\sim}G/{\sim}S\&D)$$

$$= \text{PROB}(C/G) \cdot \text{PROB}(G/D) + \text{PROB}(C/{\sim}G) \cdot \text{PROB}({\sim}G/D).$$

Theorem 2: If (ST2) and (ST3) hold and $\text{PROB}(S/G\&SD) > \text{PROB}(S/SD)$, then $\text{PROB}(G/S\&SD) > \text{PROB}(G/{\sim}S\&SD)$

Proof: Suppose $\text{PROB}(S/G\&SD) > \text{PROB}(S/SD)$. Then by the probability calculus:

$$\text{PROB}(S\&G/SD) = \text{PROB}(S/G\&SD) \cdot \text{PROB}(G/SD) > \text{PROB}(S/SD) \cdot \text{PROB}(G/SD).$$

By the probability calculus:

$$\text{PROB}(S\&G/SD) = \text{PROB}(G/S\&SD) \cdot \text{PROB}(S/SD).$$

Thus

$$\text{PROB}(G/S\&SD) > \text{PROB}(G/SD).$$

Then by the probability calculus, $\text{PROB}(G/S\&SD) > \text{PROB}(G/{\sim}S\&SD)$.

Bibliography

Eells, Ellory
1984 *Rational Decision and Causality*, Cambridge: Cambridge University Press.
Gibbard, Alan and William Harper
1978 "Counterfactuals and two kinds of expected value", in *Foundations and Applications of Decision Theory*, ed. C. A. Hooker, J. J. Leach and E. F. McClennen, Reidel, Dordrecht, 125-162.
Joyce, James
1998 *The Foundations of Causal Decision Theory*. Cambridge, Cambridge
Lewis, David
1981 "Why Ain'cha Rich?", *Noûs* **15**, 377-380.
1981a "Causal decision theory", *Australasian Journal of Philosophy* **59**, 5-30.
Horgan, Terry
1981 "Counterfactuals and the Newcomb Problem", *The Journal of Philosophy*, **78** 331-356.
Nozick, Robert
1969 "Newcomb's nProblem and Two Principles of Choice", in *Essays in Honor of Carl G. Hempel*, ed. Nicholas Rescher et al., Dordrecht: *D*. Reidel Publishing Co., 114-46.
Pollock, John
1986 *Contemporary Theories of Knowledge*, Rowman and Littlefield.
1992 "New foundations for practical reasoning", *Minds and Machines*, **2**, 113-144.
1995 *Cognitive Carpentry*, MIT Press.

2002 "Causal probability", *Synthese* **132** (2002), 143-185.
2003 "Rational choice and action omnipotence", *Philosophical Review* **111**, 1-23.
2005 "Plans and decisions", *Theory and Decision* **57**, 79-107.
2006 *Thinking about Acting: Logical Foundations for Rational Decision Making*, New York: Oxford University Press.

Savage, Leonard
1954 *The Foundations of Statistics*, Dover, New York.

Skyrms, Brian
1980 *Causal Necessity*, Yale University Press, New Haven.
1982 "Causal decision theory", *Journal of Philosophy* **79**, 695-711.
1984 *Pragmatics and Empiricism*, Yale University Press, New Haven.

Sobel, Howard
1978 *Probability, Chance, and Choice: A Theory of Rational Agency*, unpublished paper presented at a workshop on Pragmatism and Conditionals at the University of Western Ontario, May, 1978.
1994 *Taking Chances: Essays on Rational Choice*. New York, Cambridge University Press.

Stalnaker, Robert
1978 "Letter to David Lewis", reprinted in *Ifs*, ed. W. Harper, R. Stalnaker, *G.* Pearce, Reidel, Dordrecth, Holland.