UNCERTAINTY IN NEWCOMB PROBLEMS

Aaron L Bramson
December 10, 2007

INTRODUCTION

The general decision-theoretic notion of *uncertainty* is any scenario under which an aspect of it is not known with certainty. The lack of certainty usually applies to the state of the world before the decision is made or the outcome of the decision or sometimes to the type of the agent (which is really just part of the state of the world). These are cases of decision under *risk*, known probabilities for each of the possible states or outcomes. There is often some indeterminacy in these probabilities; and while updating techniques abound, principles for initial probabilities are rare and usually would be nonsensical.[1] The technical term for scenarios including unknown states, unknown outcomes, unknown probabilities, unknown preferences, or other unknown features that render some stage of the utility maximizing process valueless are ones of *deep uncertainty*. One of the antecedent conditions for applying decision theory to a problem is that it is not a case of deep uncertainty because the choice function of classical decision theory is "choose the outcome with the greatest expected utility" and if there is no expected utility for some option then the choice function does not apply.[2] There simply isn't a thing that you

[1] For example, in research using Bayes Nets, even if the probabilities of all the individual events are known, the *a priori* conditional probabilities must be assigned. Since different distributions can produce different outcomes, and there can be no principled techniques to determine these, one must sweep large areas of the possible assignments, to get relative statics of the models' behavior.

[2] This claim is intended to appreciate the distinction among 1) applies, 2) fails to apply, and 3) does not apply. *X applies to y* implies *X(y)* is true, *X fails to apply to y* implies *X(y)* is false, and X *does not apply to y* implies that *X(y)*

should or would do *according to decision theory*. There are several standard, common sources of deep uncertainty, but in this short paper I will propose a new kind of uncertainty:

INDECISIVENESS

At first consideration, deep uncertainty is not the only source of indecisiveness. In classical decision theory any scenario where it isn't the case that one outcome is strictly preferred to all others will result in a kind of indecision. The classic parable of a donkey starving to death between two identical haystacks illustrates this point. The moral of that story is to point out that indifference shouldn't really lead to indecision; that decision theory as it normally stands isn't enough but adding any kind of tie-breaker is.[3] There is no principled reason to choose one tie-breaking scheme over another and so decision theory, if applied to the tie-breaking problem, will face a problem of deep uncertainty. Let's be clear, the outcomes can't be decided over because they are equally valued, but the tie-breaking schemes can't be decided over because they are all valueless – this is why the latter, but not the former, is a matter of deep uncertainty.

In the case of having nothing to base a prior distribution upon one can again resolve the problem by fiat, as many have suggested (refs), by assuming a uniform distribution. The claim is that without any reason to bias the distribution in any way the default position is to apply equal values to all options. It seems to me (and others (refs)) that a uniform distribution is just as much a substantive assumption as an exponential distribution or normal distribution or any other "frequently observed" distributions. Another option is to try a large variety of values, but it is a rare situation that anything close to a thorough sampling is possible…the parameter space is often astronomically or infinitely large. One might propose some insane project of determining the distribution of distributions of all sorts of data types and then basing prior probabilities on the most probable distribution for the current data type, but we won't discuss that here.[4] There just isn't any good way to pick a prior distribution. There is no data on it, no

---

is ungrammatical or is nonsense for other conceptual reasons (e.g. category error) and hence does not have a truth value.

[3] Note that any kind of tie-breaker, and not necessarily a fair one, will work because the decision maker is indifferent to the outcomes. When there are multiple maximal expected-utility outcomes then all that is needed is that one is chosen as *the* act performed.

[4] This seems scientifically valid as a project, but since this space is infinite it would be hard to justify any decision made on any finite sampling of data types and distributions of data of that type. The fact that such a project is in practice impossible makes the proposal to base a decision on some consideration like this rather insane. But perhaps I am not the first person to think to do this.

precedence, no decision criterion to apply at all. So again, it is not that every prior distribution has equal utility, but rather that no utility can be assigned to any prior probability distribution – it's deep uncertainty.

Another class of scenarios that generate indecisiveness is the various forms of Newcomb problems (including Death in Damascus, Smoking Gene, and other such setups). In these problems there are two choices, A and B, and doing A is evidence that B is the better (higher utility outcome generating) action and doing B is evidence that A is the better action. Another way to read these is that whichever action you feel prone to do, being prone to do it is evidence that you shouldn't do it. Some analyses (Artzenious, Joyce) demonstrate effectively that no action is ratifiable and the proper conclusion to the decision theoretic process is a distribution of probabilities of actions wherein each outcome yields equal utility. There is some interpretation work to be done in the meaning of these probabilities of action, but the result is that decision theory does not yield a unique recommendation of action. Let's see how these Newcomb problems can be seen as revealing deep uncertainty.

THE CLASSIC NEWCOMB PROBLEM

The problem: an agent (you) are presented with two boxes, A and B, and the options to 1) take A or 2) take A and B. There is certainly $1000 in B. A reliable predictor has put $1,000,000 in A if she predicts that you will choose option 1, and $0 in A if she predicts that you will choose option 2. What to choose? Decision theory seems clearly to recommend taking both boxes because, whatever the predictor predicted, you are $1000 wealthier if you take both and so two-boxing dominates. But the predictor is *ex hypothesi* a reliable predictor and so two-boxing must be considered a reliable indicator that you'll get only $1000. But the more you think you're a one-boxer (and destined to get the $1,000,000) the more reason you have to two-box (to get the extra $1000).

Causal decision theory seems to have no way out of recommending strict two-boxing because the predictor put the money in the box already; nothing one chooses can causally alter the predictor's choice of putting the money in the box or not. Two-boxing is the dominant decision because regardless of whether the money is in the first box, you are better off taking the second box. But assumptions of causal links be damned the evidence for predictor's accuracy

lurks ominously in the Newcomb problem. One is compelled to ask, "How could the predictor be so accurate without having a causal connection to the choice?".

WHENCE THE EMPLOYED NOTION OF CAUSATION

The causal decision theorist's commitment to two-boxing rests on the assumption that the choice of boxes cannot causally relate to the money in them. Indeed there is a strong tradition of assuming that causation can only flow in the same direction as time (Lewis, etc.), and this seems to be an assumption we all want to keep. One "quick fix" that causal theorists have employed is to propose an unknown *common cause* of both the predictor's prediction (and hence the money in the box) and the chooser's choice (a la Fisher's smoking/cancer gene). Doing so gets one to a point where one is indifferent between the two options (Artzenius, Joyce), bringing the discussion back to indecisiveness. Some might leave it there and be satisfied.

But what justifies the common cause assumption? That possibility is certainly not built into the problem, nor is it explicitly precluded. That is only one possible, ad hoc fix to mend the evidence with some extant folk notion of causation. Whatever flavor of scientific doctrine one subscribes to, the evidence that the predictor will have had chosen an empty opaque box if one later chooses to two-box is strong. If we maintain anything like a Humean notion of causation (Hume, Lewis) then such strong evidence is exactly what gets encoded as causal laws. And so the notion of causation employed must either allow for the possibility of an unknown causal link of some kind or allow for holes in the fabric of causation (i.e. uncaused events, causal loops, magic, causation backwards in time, etc.).

Allowing holes in causation is in some ways easier, but it is much less satisfactory to our scientific-minded models of how things work. Proposing an unknown common cause doesn't seem any better than just allowing for some unknown causal link; the latter includes the former and there is no additional evidence to support the common cause hypothesis. The only thing pointing to a common cause is people's lack of creativity; that's all we can think of. The evidence points to some causal link or other, but our current understanding of things does not include a good candidate. The case is quite strange (and artificial), after all. But once we accept that there must be *some* causal link we are in a state of deep uncertainty about what the cause might be. One might just pick the common cause hypothesis, but deep uncertainty is not the same as indifference, and so this move is not justified. Picking is only justified when we know

we don't care, not when we don't know. When causal explanation is involved we generally maintain that only one explanation is correct, we just don't know which one in this case. Furthermore, in the description of the Newcomb problem, we have absolutely no basis for making any further claim.

The conclusion of causal decision theory here is that we have no idea what action is better because we are in a state of deep uncertainty with respect to the casual link between the prediction and the choice with no way (within the problem) to clear it up. I will attempt an analogous example: you are given the choice of which Martian moon to land a probe. Which moon do you choose? If you are like most people then you may not have even known that Mars has two moons or be able to name them (Phobos and Deimos). And even if you could, you probably don't know enough about them to decide which is more probe-worthy – especially since I didn't tell you what the probe can do. You could pick one my employing some superlatives (the bigger one, the one closer to Mars, etc.), but even if the picking processes is justified the moon you pick is not a justified choice. It's not that you believe that either moon is an equally qualified target for probing; you simply have no idea what you should do.[5] There isn't something that you should do in such circumstances.

CAUSAL VERSUS EVIDENTIAL DECISION THEORY

Now a brief aside to consider what purchase evidential decision theory is supposed to provide. Supposedly the evidence is strongly in favor of one-boxing because (in some versions) all the one-boxers get $1,000,000 and all the two-boxers get $1,000.[6] The evidentialist is supposed to conclude from these observations that whatever the mechanism (i.e. ignoring the structure of the problem) we have reason to strictly prefer one-boxing. The reason is that we can and *should* expect to be wealthier for our decision to one-box. And the evidence lends perfect support for

[5] A different, and perhaps better, example was recommended by Steve Campbell who suggests building a preference into the example but with ignorance of the satisfier. One has to pick between two identical twins to participate in a trivia match. It is only known to the chooser that one has a PhD in world sports history and literature while the other is deaf and illiterate. Which ought we to choose (left or right). This example has less uncertainty than the Martian moons example, but it highlights better the difference between mere indifference and uncertainty.

[6] Though some of the versions offer less than perfect predictions in order to give probabilistic risk assessments some leverage to change the recommendation, those considerations won't be relevant here so we can ignore them without loss of generality to other specific formulations.

this evidentialist conclusion. How long can evidence for future events and purported causes for future events remain in conflict? I will now argue that the answer is "not long".

BRINGING SCIENCE TO BEAR ON NEWCOMB

Going back to the question of where our knowledge of causal connections comes from, two answers from immediately to mind: theoretic notions and from evidence. Those theoretic notions are also grounded in evidence, but through a theory-building process. We can't get into what the theory-building practice involves here, but the point I want to bring in is that theories must be able to stand, at least temporarily, against specific anomalous results. The Newcomb problem's result is intuitively something anomalous to causal "business as usual"; the evidence points to causes that are not in our working theory. More strongly, some might want to claim that the evidence points to causes that cannot consistently be included in the relevant theory. If the conflict is real and the evidence persists then it is the theory, no matter how otherwise elegant and satisfying, that must change to accommodate the evidence. That is what science, and its systems of causal relations we sometimes call 'theories', is expected to do in such circumstances.

In the short term[7] causal decision theory recommends two-boxing because given the provided description of the situation and our folk (or otherwise working) notions of causation it is dominant. In the medium-run we may be open to exploring other causal possibilities but we should then discover deep uncertainty regarding the possible causes and so causal decision theory does not recommend any action. In the long term, our understanding of causal relationships must reflect the available evidence, and so we should expect to either 1) find strong evidence that a specific causal link is responsible for the correlation or 2) conclude that all such cases are outlandish coincidences. Newcomb problems are weird, artificial examples that depend upon accepting quite unrealistic propositions (e.g. about the accuracy of predictors or that death is going to meet us, etc.). If we could recreate Newcomb problem situations in a lab and then investigate different causal link hypotheses then we could make progress on this problem (in the scientific sense). If not, then the causal decision theorist can justifiably claim that these scenarios are only possible in the land of make-believe and that our intuitions against

[7] I don't have specific time-scales in mind, but rather stages of inquiry each requiring more understanding. Greater understanding, we can safely assume, takes some amount of time to develop.

working notions of causation are generated by the outlandish coincidences purported to be facts in the problem description. Then we are justified in two-boxing, artificially generated evidence be damned.

CONCLUSIONS

I claim that we ought to be, generally speaking, long-term causal decision theorists. Sometimes decisions must be made before that kind of analysis is performable, or the scenarios are too difficult (e.g. impossible, costly, time-consuming) to construct in practice. If we are thus limited, then the next best thing is to be a medium-term causal decision theorist. That will put us in a state of deep uncertainty regarding many of these Newcomb problem-styled scenarios, forced to pick without reasons or justification (even if we are confident that *some* reasons must exist as in the Martian moons example). Such picking is outside the domain of decision theory. The short-term decision theorist may do better to trust the evidence (whether on the hopes that underlying causes will be found) or to stand on received dogma regarding causal principles, but given the problem descriptions we are given we simply don't have access to what one *ought* to do in Newcomb problem cases.