HEMPEL, GOOD, and BAYES[*]

by

Itzhak Gilboa[+]

May 1993, Revised August 2003

Abstract

This paper analyzes some decision and belief paradoxes from a Bayesian viewpoint, focusing on Hempel's "paradox of confirmation" and Good's variation thereof. It is shown that a straightforward Bayesian analysis resolves the paradoxes discussed. These examples are used to support the view that the Bayesian paradigm is a very effective tool for providing a coherent and intuitive representation of belief.

[*] I wish to thank many people for conversations and correspondence which motivated and influenced this note. In particular, I would like to mention Ilan Eshel, Peter Gardenfors, Gabi Gayer, Eva Gilboa-Schechtman, Shlomo Gilboa, Sergiu Hart, Isaac Levi, Eric Maskin, Roger Myerson, Ady Pauzner, Judea Pearl, Dov Samet, David Schmeidler, Teddy Seidenfeld and Bob Wilson.

[+] Berglas School of Economics, Tel-Aviv University, and Cowles Foundation, Yale Univeristy. This paper was written while the author was at Kellogg Graduate School of Management, Northwestern University.

1. Introduction

The "Bayesian paradigm" is a loosely-defined term. By all accounts, it is based on the idea that any uncertainty should be represented by a probability function (a "prior"), which is to be updated in the face of new information according to Bayes' rule (to obtain a "posterior"). In the context of decision theory or any application thereof, it is often taken to imply also that decisions are, or should be made so as to maximize expected utility. At times, the "Bayesian paradigm" is also interpreted as prescribing a "rational" way to define the state space, on which probability is defined.

In this paper, the "Bayesian paradigm" (BP) will refer only to the aspects of this approach which relate to belief representation and update, ignoring the decision-theoretic side. However, as far as beliefs are concerned, we will take a broad interpretation of BP, and, in particular, assume that when it is used, the set of states-of-the-world is defined to be all logically-consistent functions from propositions to truth values, where the set of propositions is assumed to be rich enough to describe all conceivably-relevant aspects of the situation modeled.

While BP enjoys the status of a dominant paradigm in economic, decision and game theory, it is by no means free of criticism. Descriptively interpreted, there is ample evidence (both behavioral and cognitive) that the commands of the Bayesian paradigm are consistently violated. From a normative viewpoint, BP is also sometimes criticized as impractical or even useless.

We share many of the concerns regarding the universal applicability of the Bayesian paradigm. However, there is one thing which BP seems to do very well, probably better than any other approach: providing a tool for coherent and intuitive qualitative representation of beliefs. That is, if one is to think clearly about a problem that relates to beliefs and to their update, and if one is guaranteed not to be called upon to actually quantify the beliefs in question, the Bayesian paradigm is a highly recommended tool. It may not be as useful when used as a descriptive theory of belief formation and update; it may well prove hardly applicable if actual priors are to be specified; but it can boast stellar performance when it comes to resolving confusion.

The failures of the Bayesian paradigm will not be discussed here. The purpose of the following sections is to convince the reader of the strength of BP as a tool for qualitative reasoning. To this end, we will discuss a few well-known paradoxes in the philosophy of science, and show that they fail to embarrass the Bayesian.

The analysis of these paradoxes does not take any originality of thought. On the contrary, the Bayesian "resolutions" of these paradoxes is arrived at by an almost-algorithmic process. All that is needed is to follow the steps in the (unwritten) cook-book of the Bayesian cuisine, and the alleged problems dissolve.[1] This lack of originality is precisely the point of this paper: the Bayesian paradigm is successful enough not to require innovative adaptations or ad-hoc modifications in order to deal with some well-known and widely-discussed puzzles.

The rest of this paper is organized as follows. Section 2 attempts to clarify the notions of "paradox" and "resolution" as used in this context. Section 3 deals with Hempel's "paradox of confirmation" ("The Ravens Paradox"). In section 4 we discuss a variant of it, introduced by Good. Section 5 briefly discusses the behavioral version of Newcomb's paradox, and the "Monty Hall Paradox". Finally, section 6 concludes.

2. Paradox and Resolution

Since the rest of this paper deals with various "paradoxes", it may be helpful to define the relevant terms at the outset.

A *paradox* can be thought of as a logical inconsistency of a set of axioms, in which one believes or to which one is psychologically attached. One of the greatest paradoxes in the history of mathematics is considered to be the fact, known to the Greek mathematicians, that $\sqrt{2}$ is an irrational number. The axiom that all real numbers are rational was inconsistent with other axioms implying the above (including the fact that there is such a thing as $\sqrt{2}$). This could have been a source of great puzzlement and even distress to anyone who believed in all of the axioms simultaneously. Naturally, we have no reason to be embarrassed by this contradiction today, since we do not have any attachment to one of the problematic axioms, i.e., we learned to accept the fact that some real numbers are not rational.[2]

This definition of a paradox is inherently subjective and quantitative. What is paradoxical to some will not be paradoxical to others. Further, some puzzles may be more paradoxical than others, depending on the degree of belief in or psychological attachment to the axioms involved. Since both these features, subjectivity and fuzziness, seem to apply also to the everyday usage of the word, we view them as merits, rather than flaws, of this definition.

Whereas any proposition may serve as an axiom in a mathematical context, in science one may wish to distinguish between axioms that reflect a-priori intuition and propositions that describe data. For the purposes of the present discussion, we would like to exclude the latter from the set of axioms considered. Thus, if a theory one likes happens to be refuted by observations, the resulting embarrassment will not qualify as a paradox in this paper.

[1] As discussed below, however, "the Bayesian paradigm" as we know it today may have been refined thanks to some of these paradoxes.

[2] I thank Ilan Eshel for pointing out this example and the general point regarding the socio-psychological nature of paradoxes from. (In an undergraduate course in mathematical genetics, Tel-Aviv University, 1981-2.)

The distinction between observations and intuition is admittedly problematic. Intuition is an integral part of the observations one attempts to explain by a scientific theory. In particular, for much of the social sciences it seems that data obtained by introspection are legitimate as well as indispensable. Consider, for instance, Allais's Paradox (Allais (1953)) in decision theory. It does not qualify as a paradox simply because experimental work has shown that von Neumann and Morgenstern's (1944) independence axiom is violated by many decision makers. But this refutation of the theory will be paradoxical to the extent that one feels that the "right" choices in Allais's experiment do violate postulates that one would like to satisfy.

The above notwithstanding, our discussion will be more fruitful if we draw a line between paradoxes and refutations: in a paradox the contradiction is obtained from axioms that are generally intuitive; in a refutation, by contrast, axioms of this nature are contradicted by specific examples.

There are several ways in which a paradox may be resolved; indeed, one may classify paradoxes (*ex post*) by the resolutions they call for. First, it may be the case that, sad as it is, the paradox is real in the sense that it points out a true incompatibility of axioms with clear formal (or easily formalizable) content. In this case one has no choice but to give up some of the axioms, or at least weaken them. The example of $\sqrt{2}$ seems to be of this nature: the axiom that every real number is rational had eventually been renounced.

At the other extreme, one may find, upon careful inspection, that the contradiction does not exist after all, and is simply a result of a mathematical mistake. This case of a spurious paradox is hardly of great interest, but it is an important benchmark.

Finally, there are many paradoxes that are neither real nor spurious, because the axioms they rely on are not precisely formalized. This seems to be the case with many of the more interesting paradoxes, and sometimes also the more instructive ones. A spurious paradox may serve, at best, as a nice puzzle. A real one forces us to discard some axioms so as to make our beliefs consistent. But a paradox that points at an ambiguity in the language in which the axioms are formulated very often opens new horizons. Russell's paradox, and more generally, the family of paradoxes relying on self-reference, are instructive in this sense.

Many of the interesting and instructive paradoxes hinge on some beliefs (or axioms) of which one is not aware until attempting to resolve them. For instance, Goodman's paradox (Goodman (1965)) may be interpreted as showing that the notion of simplicity is intrinsically language-dependent. This is quite surprising when encountered for the first time. Yet most people would not be aware that they implicitly assume the contrary until faced with Goodman's example (or a variation thereof). Similarly, most people tend to believe that all seemingly-meaningful propositions are indeed meaningful and can be assigned truth values in a consistent manner, but they would not be aware of it until encountering some version of the liar's paradox.

It follows that a resolution to a paradox may consist of some combination of the following: (i) checking the mathematical proof of the contradiction carefully; (ii) spelling out the intuitive but implicit assumptions, and formalizing them to a certain degree of clarity; and finally (iii) discarding some of the basic axioms.

The subjective nature of paradoxes and the fact that their resolutions may involve formalization of hidden assumptions imply that there are many ways to skin a paradox. In the following we will therefore describe each paradox in a verbal, informal way, and then focus on one of its aspects that seems the most intriguing, or the most challenging to a theory of belief representation. It goes without saying that the focus is subjectively chosen. Indeed, in most cases the account of the paradox given here, as well as the "focus", will differ from the original ones.

3. Hempel's Paradox

The following account, though taken out of context, is fairly faithful to Hempel's original example (Hempel (1945,1966)):

Story

Suppose we wish to test the rule/law/hypothesis that all ravens are black. A procedure that seems acceptable to all is to randomly select ravens, and check each of them for blackness. One counter-example would, of course, suffice to refute the rule. On the other hand, the general rule will never be proven by examples. However, the more ravens we test, the stronger is our belief in the truthfulness of the general rule, should they all turn out to be black.

Notice that "all ravens are black" is logically equivalent to "all that is not black is not a raven," or simply "all non-blacks are non-ravens". Thus one may test the latter rule rather than the original one. And by the same methodology, one may randomly select non-black objects, test them for "ravenhood", and then either refute the rule or increase its plausibility.

From here to embarrassment the way is short. Pick a non-black item from your desk, or consider a red herring. As a non-black object, it qualifies for the sample; as a non-raven, it should lend support to the rule tested. Yet it seems patently absurd to use such evidence to confirm the blackness of ravens.

Focus

The rule discussed is of the form $(\forall x \in A)(Q(x))$ for some proposition $Q(x)$. In this case, A is some set of objects, and $Q(x)$ is an implication of the type $R(x) \rightarrow B(x)$, where the predicate $R(x)$ is interpreted as "x is a raven" and $B(x)$ -- as "x is black". Equivalently, $Q(x)$ may be written as $\neg B(x) \rightarrow \neg R(x)$.

A positive example is an element $a \in A$ such that $Q(a)$ holds. The implicit assumption we would like to focus on is that positive examples confirm the rule, i.e., that having observed a positive example, our belief in the general rule increases.

Since the exercise discussed has to do with hypothesis testing, sampling, and belief, we find it natural to discuss it in statistical terms. Furthermore, we would like to focus on Bayesian statistics, and will attempt to show that the Bayesian paradigm offers a natural resolution to the paradox. Thus we assume that there is a prior probability measure P, which has a value

$$P\big((\forall x \in A)(Q(x))\big) = p_0$$

and is updated to

$$P\big((\forall x \in A)(Q(x)) \mid Q(a)\big) = p_1$$

for some $a \in A$.

The precise meaning of "the rule is confirmed by the positive example" will turn out to be crucial. Let the *weak confirmation axiom* state that $p_1 \geq p_0$ and the *strong confirmation axiom* -- that $p_1 > p_0$.

Resolution[3]

It is easy to see that the weak confirmation axiom can not be violated by the Bayesian paradigm. Indeed, assume for simplicity that A is finite, and $|A| = n$. (The finiteness assumption is immaterial. For countable sets the analysis is identical; for uncountable ones -- identical up to measurability constraints.) Assuming that the only relevant aspects of the world are the ravenhood and blackness of the objects in A, define the state space to be

$$\Omega = \left\{\omega \mid \omega : A \to \{0,1\}^2 \right\}.$$

For $\omega \in \Omega$ and $a \in A$, $\omega(a) = (v_1, v_2)$ should be read as follows: at ω, the object a is a raven iff $v_1 = 1$; it is black iff $v_2 = 1$. Thus the rule $(\forall x \in A)(Q(x))$ corresponds to the event

$$Q = \left\{\omega \in \Omega \mid \forall x \in A, \omega(x) \neq (1,0)\right\}.$$

Hence, Ω contains 4^n states of the world, out of which 3^n are in Q.

For every $a \in A$, define an event

$$Q_a = \left\{\omega \in \Omega \mid \omega(a) \neq (1,0)\right\}.$$

Thus

$$Q = \bigcap\nolimits_{a \in A} Q_a \quad \text{and} \quad Q_a \supseteq Q \text{ for all } a \in A.$$

[3] For other Bayesian resolutions of Hempel's paradox, see Korb (1994). His resolution differs from ours in several ways. In particular, the resolution proposed here retains the weak confirmation axiom, and it is used to highlight to algorthmic nature of the Bayesian analysis.

It is therefore a trivial observation that the Bayesian paradigm cannot violate the weak confirmation axiom:

$$P\left(Q \mid Q_a\right) = \frac{P(Q \cap Q_a)}{P(Q_a)} = \frac{P(Q)}{P(Q_a)} \geq P(Q) \ .$$

Yet there is no reason to expect it to satisfy the strong confirmation axiom, which would require a strict inequality above.

At this point we have clarified what could be meant by a "confirmation" of a rule. We note that the story above does not suggest that the weak confirmation axiom is violated: testing a red herring may not increase the plausibility of all ravens being black, yet it would not decrease it either. Thus, if one only believes in the weak confirmation axiom, the Bayesian analysis poses no problems.

On the other hand, those who believe in the strong version of the axiom are still troubled, and they may wish to renounce the Bayesian paradigm before they discard this axiom. In the next few paragraphs we attempt to convince the reader that the strong confirmation axiom does not make much sense in general, nor, indeed, in the red herring example.

Consider the following story: an ornithologist who wants to test the rule mentioned above goes out to the field and observes a raven, who happens to be black. So far, it seems that we finally have a reasonable ornithologist for a change. However, after this successful observation, our scientist stays put and observes the same raven again. And again. By the end of the day, she writes a paper, reporting 1000 observations of ravens, all of which happened to be black, thus strongly supporting the claim that all ravens are black.

It seems obvious that the number 1000 is misleading here. After all, what additional information was gained from the second observation on? Note that the issue here is not merely one of quantitative difference: it is not only the case that 999 new ravens would be more convincing that 999 additional observations of a known raven. The difference is qualitative: observing the same raven again does not add *anything* to our belief in the rule.[4] More generally, when we condition on an event that is *already known,* our beliefs naturally do not change. This, of course, is captured by the Bayes' update formula.

The red herring example is quite similar: if we know that an object a is a herring, we already have $P(Q_a) = 1$. That is, we assign zero probability to the event that this object will turn out to refute the general rule, simply because we know that this object is not a raven.

It is sometimes argued that testing the counterpositive of a rule cannot lend support to the rule itself. We argue that what makes a red herring a preposterous example is not that it is not black (i.e., that the counterpositive is tested), but rather that it is *known* not to be a raven. To see this

[4] Of course, we assume here that ravens do not change their color at some point of time. One paradox at a time.

point, assume that the non-black object tested may be a raven with some positive probability. In this case, finding out it is not a raven after all does indeed increase our belief in the rule. For instance, assume we cannot really tell which birds are ravens and which are not. We choose 1,000,000 non-black birds at random, and take them to an ornithologist for an examination. The next day, the expert tells us that none of the birds we brought was a raven. In this case, it seems very plausible to strengthen our belief that all ravens are black.

It follows, then, that there is nothing wrong with representing the rule Q by its counterpositive, and this does not seem to be the problem. Indeed, the Bayesian paradigm does not distinguish between different propositional representations of the same event.

To sum, in a Bayesian analysis the belief in the rule cannot decrease as a result of observing a positive example. Furthermore, for the observation of any positive example *that is not already known* the ex-post belief in the rule will be strictly higher than the ex-ante one. It is only when we condition on known facts that the Bayesian paradigm violates the strong axiom of confirmation. Indeed, violating this axiom in those cases seems to be a virtue of BP.

Comments

- The Bayesian paradigm is also criticized for failing to satisfy one of Hempel's axioms of confirmation, namely, that if B implies C and A is a confirmation of B, then it should also be a confirmation of C.

The intuitive appeal of this axiom is, as always, a matter of potential dispute. Some people (like the author), would start reasoning about it in Bayesian terms from the outset, concluding that it does not make much sense. Those who wish to retain it will, indeed, have to renounce the Bayesian paradigm.

In an attempt to convince the reader that the BP should be retained at the expense of the above axiom, let us consider a simple example. Suppose that people are divided into three categories: "dumb", "normal", and "smart". Most of the people are normal, and there are dumb and smart people in similar frequencies. Let us say the proportions are 1%-98%-1%.

Let A be the event (or proposition) that a randomly selected individual is not normal. Let B stand for "the individual is smart" and C -- for "the individual is not dumb". Obviously, B implies C. A does seem to "confirm" B -- if the individual is not normal, it is much more likely that she is smart than it was a priori. Yet A does not lend support to C: an unusual individual is *not* more likely to be "not dumb".

The above example is couched in almost-Bayesian terms, and may thus seem far from a "fair" test of the axiom's plausibility. Yet, in order to accept the axiom, one would like it to be applicable also in those situations where relative frequencies in a certain population are indeed given.

The example also highlights a fundamental difficulty with the axiom: if B implies C, but is not *equivalent* to it, it is not clear that a confirmation of the former should amount to a confirmation of the latter as well. As in the example above, B may be only one way in which C may hold, and not necessarily the most plausible way. If B becomes more plausible, but C-and-not-B does not, it is not entirely obvious that C should indeed be more likely. Needless to say, if B and C *are* equivalent, their plausibility according to a Bayesian model will always be identical.

One of the advantages of the Bayesian paradigm that is illustrated here is that it forces one to "close" the model. Any question of the type "what if?" has a counterpart, "and what if not?", which pops up in Bayes' formula. This can be extremely annoying if one is to estimate actual probabilities. But for qualitative reasoning it seems to be intuitively appealing, as well as inevitable for consistency.

- There are, of course, many other resolutions to Hempel's paradox. For instance, it has been argued that the probabilities involved are only conditional ones, and that there is no room for the "probability of the rule" as such. This seems unduly restrictive: why can we not assign a probability to a well-defined proposition? How do we quantify our belief in various theories?

Alternatively, one may restrict one's attention to a model in which only ravens are considered to begin with. In this case, one deals with conditional probabilities, but the conditioning is not done within the model. Similarly, it is not clear how one can combine a theory of ravens with one of nightingales.

- A common reaction to the paradox is that there are many more non-black things than there are ravens, hence it does not make sense to test non-black objects rather than ravens. This seems to miss the main point: if it were only a matter of quantitative difference, testing a red herring would have been inefficient, but not as ludicrous as it is in this example. The absurdity of the red herring stems from the fact that there is no uncertainty about the result of this test. Namely, it is the qualitative difference between certainty and uncertainty, or between weak and strong inequalities, that drives the paradox.
- The analysis presented above does not necessitate a Bayesian approach. Indeed, one may make a similar argument without explicit reference to probabilities: since we *know* at the outset that a red herring is not a counterexample to the rule, testing it and realizing that, indeed, it does not contradict the rule should lend no additional support to the latter. However, the main point of this example is that the Bayesian Paradigm *implies* this analysis. That is, while one may or may not notice the above subtlety using other approaches, the Bayesian paradigm forces one to spell out subjective probabilities, and one cannot evade the distinction between what is known (with probability 1) and what is believed (with probability smaller than 1).
- Hempel (1945) stresses the importance of background knowledge, i.e., of all the things we know while we get additional evidence. He argues that "... if we are careful to avoid this tacit

reference to additional knowledge... the paradoxes vanish." (p. 20) Thus he prefers to hold to some version of the "strong confirmation axiom" by hypothesizing a state of knowledge in which additional information is disallowed. Indeed, his belief that a confirmation of a hypothesis should be a confirmation of any implication thereof, is in line with ignoring all "additional knowledge."

The Bayesian paradigm does exactly the opposite: by assuming a prior belief over all states of the world, any relevant knowledge is already incorporated into it. Thus, in the Bayesian analysis it does matter whether the result of the experiment is already known, and a hypothesis need not be supported by any evidence that supports a stronger hypothesis.

The advantage of the Bayesian paradigm as a model of qualitative reasoning about uncertainty is two-fold: from a descriptive point of view, it reflects people's reasoning even when background information does exist, and can deal with the actual state of knowledge, not only with a hypothetical one. From a normative viewpoint, it forces one to think about all the relevant aspects of the problem at hand.

4. Good's Paradox

Good (1967; 1968; 1986) has suggested a variation of Hempel's paradox. The following is a simplified version of Good's paradox that highlights the issues discussed above.

Story

Consider a population containing two items. It is known that one of the two holds: either both items are red herrings, or both are ravens, in which case one is black and one is red. Let us suppose, for simplicity, that both possibilities are equally likely. Hence the prior probability of the rule "all ravens are black" is 50%, since it holds true in the herring population but does not hold in the raven population.

We are now told that item 1 is a black raven. It follows that the population consists of ravens, in which case the posterior probability that "all ravens are black" has *decreased* to 0.

Focus

In the context of the discussion of Hempel's paradox, it appears that the Bayesian paradigm fails to satisfy the weak confirmation axiom after all: in face of a positive example, the posterior probability of the rule is lower than its prior probability.

Resolution

Borrowing the notation from section 3, for $A = \{1,2\}$, we recall that $P(Q \mid Q_1) \geq P(Q)$, that is, the fact that item 1 does not contradict the rule cannot lower the latter's probability. How is this reconciled with the fact that this probability has decreased to zero?

The resolution lies in a careful reading of the phrase "item 1 does not contradict the rule." What this amounts to is "item 1 is not a non-black raven." This does *not* imply that item 1 is a raven at all. Thus, the information that item 1 is a black raven may be decomposed into two distinct pieces of information: (i) item 1 is not a counterexample to the rule; and (ii) item 1 is a raven. Indeed, proposition (i) does not lower the probability that all ravens are black. It is the second proposition which tells us we are dealing with the raven population, for which we know the rule does not apply.

It may be helpful to spell out the states of the world in this example. Having two items, we have 16 states of the world, each of which specifies for each of the two items one of the four possibilities: the item is a black herring (BH), black raven (BR), red herring (RH) or red raven (RR). The prior probability we stipulated can be represented by the following table:

		item 2			
		BH	BR	RH	RR
	BH				
item 1	BR				0.25
	RH			0.5	
	RR		0.25		

(where blank entries denote zero probability). For the raven population, it is assumed here that items 1 and 2 are equally likely to be the red (hence also the black) raven.

The event "all ravens are black" (denoted by Q above) consists of the shaded 3×3 northwest matrix. Its prior probability is 50%. The event "item 1 does not contradict the rule" is represented by the top three rows. Indeed, given this event *alone*, the posterior probability of Q is $\frac{2}{3}$, i.e., larger than the prior. However, given that item 1 is also a raven, it becomes zero. Furthermore, this latter piece of information suffices: the event "item 1 is a raven", which is represented by rows 2 and 4, leaves zero probability on the shaded area.

To conclude, what decreases the probability of the rule "all ravens are black" is not the positive example per se; it is additional information, that does not follow from the positive example.

Comments

- Good's original point is that, contrary to common belief, "a hypothesis of the form all A's are B's is supported by seeing an A that is a B." (See Good (1986).) His example proves this point. Furthermore, it is indeed very intuitive that "seeing an A that is a B" would confirm the general law. Thus, Good's paradox is "real" to the extent that one happened to believe in this axiom.

The point we would like to stress here is that "seeing an A that is a B" is more than merely seeing a positive example. Thus Good's point does not relate to confirmation-by-positive-examples as such. Rather, it deals with a natural fallacy that has to do with the linguistic representation of the rule more than with its essence.

Furthermore, Good's paradox emphasizes the advantage of the Bayesian paradigm's language: since it deals with events, rather than propositions, it is less susceptible to linguistic fallacies. That is, using BP, one is forced to identify "all A's are B's" with "all non-B's are non-A's" and with "there is no A-and-not-B". Thus formulated, it is clearer that "item 1 is A and B" contains more information than "item 1 is not a counterexample."

- The resolution of Good's paradox suggested here hinges on the fact that it does not take a raven to be an example of the rule "all ravens are black." That is, the fact that item 1 is a positive example of the rule does not imply it is a raven, and this allows us to argue that it is not the positive example itself that decreases the posterior probability of the rule.

Somewhat ironically, this is precisely the issue in Hempel's paradox, namely, that a red herring *is* a positive example of the rule. The same logical equivalence which seems to confuse us in Hempel's example comes to our rescue in Good's story.

It follows that alternative resolutions to Hempel's paradox that renounce the equivalence principle may have difficulties in dealing simultaneously with Good's example.

- In the example above, the rule "all ravens are black" was vacuously true of the herring population. However, this is not essential.[5]

5. Other Puzzles

In this section we analyze two paradoxes that deal with the way one defines the space of states of the world.

5.1 Newcomb's Paradox

The original version of Newcomb's paradox involves an omniscient being, capable of predicting one's decisions. (See Nozick (1969).) The notion of omniscience is a little too metaphysical for our purposes here. It is not clear what does it mean for some other entity to know one's decisions, nor how does one get to know that this entity is indeed omniscient. It seems safer to think of other entities as behaving *as if* they knew certain things, and to entertain beliefs regarding such behavior. Correspondingly, we present here the behavioral version of Newcomb's paradox, and we attempt to resolve only this version within the Bayesian paradigm.[6]

[5] As a matter of fact, in Good's original example there are a few black ravens in this population.

[6] I was introduced to this behavioral version of the paradox by David Schmeidler.

Story

You are standing in front of two boxes, one of which is transparent, the other -- opaque. The first contains $1,000. You know (or believe with probability 1) that the opaque one may either contain $1,000,000 or nothing. The choice you are faced with is to take only the opaque box, or to take *both*. Obviously, it seems that the "right" choice is to take both boxes.

One more piece of information may be relevant: you are not the first person to be in this situation. Actually, you are the last in a line of 10,000 people. Waiting patiently for your turn, you happened to notice that all your greedy predecessors, who took both boxes, ended up with an empty opaque box, i.e., with $1,000. On the other hand, all the modest ones, who took only the opaque box (and let us assume there were quite a few), found the money in it, and walked away happily with $1,000,000. Will you still take both boxes?

Focus

Those of us who still choose the two boxes probably do not find any reason to be embarrassed. Let them skip to sub-section 5.2 with their $1,000. I assume that I am left with those readers who, like me, would choose to behave modestly. We are probably very rich now, but we are still not happy, since our behavior troubles us: how come we are that irrational? It appears that taking both boxes is a dominant act: whatever is the state of the world, you will have $1,000 more by taking both boxes as compared to taking only the opaque one.

Choosing a strictly dominated act is a decision-theoretic problem. It does not pose any problem to a theory of belief representation. However, as we will see shortly, the analysis of choice in a Newcomb situation is closely related to belief formation and update.

Resolution

The resolution to the behavioral version of the paradox is well-known, and is reported here for the sake of completeness alone. (See Jeffrey (1965) and "David's problem" in Gibbard and Harper (1978).) The main point is that one should not jump to conclusions. That is, one can never know for sure what aspects of the world depend on one's actions. Thus, in order to avoid surprises, one should take into account all possible "causal" relationships in the formulation of the states-of-the-world.

Analyzing the story above with two states of the world ("the opaque box contains $1,000,000" and "the opaque box is empty"), the very formulation of the decision model implicitly presupposes that the box's content is independent of one's choice. No matter how many other people one has observed, this implicit assumption, because it is implicit, cannot be updated, retracted, or modified. When it is time to make a choice, taking both boxes seems to be a dominant act in this two-state model.

The "unprejudiced" approach, by contrast, will not presuppose that the choice has no effect on the box. Thus, even though the story above made one believe that the one million dollars are either there or not, when formulating the states-of-the-world model, one should allow for the possibility of double-bottomed boxes, last-minute trickery, and so forth. In short, one needs to have *four* states of the world, which are functions from acts to outcomes.

Once there are four states in this problem, the dominance argument no longer holds: there is a state at which the greedy act yields a payoff of $1,000, and the modest one -- of $1,000,000. Furthermore, should every state have some positive prior probability, and should the first 9,999 other people's choices serve as valid data (namely, be considered as facing i.i.d. draws from the same distribution over the states), the posterior probability of this state may be very high, rendering the intuitive choice perfectly rational.

Comments

- In a sense, the lesson we learn from this example is that one should make all assumptions explicit. Making an implicit assumption by omitting some of the conceivable states of the world may be an irreversible error: from that point on, nothing in the formal model may indicate that the assumption could be wrong, and there is no way to revive the excluded states by a Bayesian update. By contrast, if all conceivable states are present in one's model, the model is rich enough to describe all assumptions that may be implicit in the prior probability.

In the example we discuss, one may require that every state of the world have positive probability. However, this is in general incompatible with the requirement that all conceivable states be present in the model. Hence the classical Bayesian paradigm cannot be applied in a completely "unprejudiced" way. To be precise, the states of the world can be defined in a "canonical", "unprejudiced" way. But if there are uncountably many of them (as should be in general), the prior one starts out with puts some limitations on what one may learn by Bayes' update. (See Blackwell and Dubins (1962).)

Yet, the moral of the story may still be applied in more restricted set-ups. For instance, in the analysis of Newcomb's problem above, we discussed four states of the world. These are not all the conceivable functions from acts to outcomes, nor do they exhaust all the functions from the set of acts to the set of four outcomes described in the original story[7]. But they do suffice to resolve the paradox and justify what appears to be the intuitive choice. Thus, while it is admittedly inevitable

[7] This set includes the payoffs of $0 , $1,000 , $1,000,000 and $1,001,000.

that one would make *some* assumptions about the world, the weaker they are, the less is one prone to paradoxical reasoning.

5.2 Monty Hall's Paradox

The TV show "Let's Make a Deal", run by Monty Hall, bears some relevance to our discussion here. In actuality it is more complicated than the version we describe here, and should be analyzed as a game rather than as a one-person decision problem. However, this additional complication is irrelevant to our purpose.

Story

You are a contestant in the show. You have to choose among three shut doors, say A, B , and C. It is known that one of them conceals a prize -- a car -- and the other two do not (they conceal goats). For simplicity, we assume that all doors are equally likely to conceal the car, and without loss of generality assume you choose door A.

However, before the door is opened, the host (Monty Hall) has to open a door, showing you whatever is behind it, and then to allow you to choose a new door. Let us assume that the host has to open a door that is not the one you have chosen, and not the one concealing the car. Say Monty Hall opened door B, and you see a goat there. You can now decide to "stick" to your original choice and bet on A, or to "switch" to the other unopened door, namely C. What is your choice?

Focus

It is quite simple to see that the strategy "switch" wins the prize with probability $2/3$, and "stick" -- with probability $1/3$. What puzzles many people is the following argument: the prior probability was $1/3$ for each door. Now that door B was flung open, one should update it and get a posterior of $1/2$ for each of the remaining doors. Why is "switching" any better than "sticking", then?

Resolution

The resolution to this paradox is similar to the previous one. One simply has to start out with an appropriate states-of-the-world model. The naïve approach suggests modeling the situation with three states of the world, depending on where the car is. With this model, it is perfectly true that "the car is not behind door B" leaves the other two states equally likely.

However, this is an inappropriate model for this situation since it makes an implicit unwarranted assumption, i.e., that the *way* in which information is acquired is irrelevant[8]. An

[8] This formulation of the moral of the story is due to Roger Myerson.

appropriate model would take into account the mechanism by which information is revealed. In this case, "Monty Hall opened door B" implies that "the car is not behind B", but it says more than that. Analyzing the problem in a 9-state model, where each state of the world specifies where the car is, as well as which door Monty Hall opens, shows, indeed, that the overall success probability of "switch" is $\frac{2}{3}$.

Specifically, a 9-state model may be represented by the following matrix, specifying the probability measure:

		Monty	Hall	opens	
		A	B	C	total
car	A	0	α	$\frac{1}{3}-\alpha$	$\frac{1}{3}$
is	B	0	0	$\frac{1}{3}$	$\frac{1}{3}$
behind	C	0	$\frac{1}{3}$	0	$\frac{1}{3}$
	total	0	$\frac{1}{3}+\alpha$	$\frac{2}{3}-\alpha$	1

In the above we assume without loss of generality that the contestant's original choice was A. (The complete model would have 27 states to account for the other possibilities as well.) It is also assumed that, if Monty Hall has a choice, i.e., in case the car is actually behind door A, he opens doors B and C with conditional probabilities of 3α and $1-3\alpha$, respectively, for some $\alpha \in \left[0\,, \frac{1}{3}\right]$. For the symmetric case, i.e., $\alpha = \frac{1}{6}$, the conditional probability of the car being behind door C, given that Monty Hall opened door B, is $\frac{2}{3}$. Thus it is indeed optimal to switch[9]. In the asymmetric case the conditional probabilities (given that Monty Hall opened door B and given that he opened C) will differ from $\frac{2}{3}$, but "switch" will still have an overall $\frac{2}{3}$ probability of winning.

The main point is that, once Monty Hall has flung door B open, one should not condition on the first and third rows in the above matrix (i.e., "the car is not behind door B"). Rather, one should condition on *all that one knows*, that is, on the middle column ("Monty Hall opened door B"), which, in particular, leaves zero conditional probability for the middle row. Conditioning on the event "the car is not behind door B" simply does not make use of all the information available, and, specifically, does not take into account the very fact that this event is known.

[9] "Optimality" here assumes that the contestant wishes to maximize the probability of winning the prize. Of course, there may be other objectives as well. In particular, regret considerations are ignored.

Comment

A relevant question is, when does one know that the model specified is rich enough? Is it not the case that whatever one knows, one may still ask how has one come to know it?

The answer to this seems to be simple: after incorporating the information channels into the model, one is left with the immediate experience of knowledge (or with "sense data"). However, if one's knowledge satisfies the standard assumptions (known in modal logic as S5), and, in particular, the axiom of positive introspection, one knows that one knows whatever one knows. Thus, for a fact such as "the car is not behind B", one may notice that one also knows "I know that the car is not behind B". However, for the latter, adding the prefix "I know that..." generates an equivalent proposition. In other words, the axiom of positive introspection avoids an infinite regress.[10]

6. Conclusion

6.1 This paper argues that the Bayesian paradigm offers a coherent and intuitive way to reason qualitatively about beliefs and their update. The main point is that using BP, one may follow an almost-algorithmic modeling technique that avoids puzzles and paradoxes.

The guiding principles of the Bayesian paradigm are: (i) define the state space in an "unprejudiced" way, without making any implicit assumptions by exclusion of some conceivable states; (ii) form a prior over this state space; and (iii) update the prior according to Bayes' rule, conditioning on all the information available.

While some of these principles are sharpened and better understood thanks to the paradoxes discussed, it is important to note that they are general in nature. These are not ad-hoc rules concocted in order to cope with this paradox or the other. Furthermore, with the possible exception of (i), they predated the paradoxes. Thus the success of the Bayesian paradigm in resolving the puzzles should be taken as evidence of its merit.

6.2 It should probably be emphasized again that BP is suggested here as a tool for *qualitative* reasoning about uncertainty. It is not argued, nor is it the author's belief, that it is also useful for quantitative applications. For instance, while Hempel's paradox is qualitatively explained by the Bayesian approach, it seems hopeless to actually estimate one's prior probability that "all ravens are black," as an event in the state space, in which the blackness and ravenhood of every object is determined. Fortunately, one does not need to have an actual numerical estimate of the probability function in order to understand its mathematical behavior.

6.3 The "canonical Bayesian paradigm", as described here, prescribes that the states of the world be formulated in an "unprejudiced" way, making no implicit assumptions, and, in particular, allowing all potential "causal" relationships between one's acts and the resulting outcomes.

[10] This was a conclusion of a conversation with Dov Samet.

However, as argued in Gilboa and Schmeidler (1995) , the notion of a "prior" on such a space is somewhat metaphysical. The behavioral derivation of a prior as in de Finetti (1937) and Savage (1954) is incompatible with the "canonical" state space. Specifically, Savage's set of "conceivable acts" on this space is by two orders of magnitude larger than the set of actually available ones. Thus the preference order over the conceivable acts is not observable even in principle.

Deriving the notion of a "prior" from cognitive data ("qualitative probability" relations) seems shaky as well. It is hardly convincing to argue that one has directly available probabilities, or consistent and quantifiable plausibility judgments on the canonical state space.

This point is closely related to the previous one: suggesting BP as a tool for qualitative analysis, a metaphor, or an argumentation technique does not entail a literal interpretation of the prior as quantification of belief. In this paper we only support the canonical BP as a framework for qualitative reasoning, and not as a scientific theory in the usual sense.

References

Allais, M. (1953), "Le Comportement de L'Homme Rationel devant le Risque: critique des Postulates et Axioms de l'Ecole Americaine." Econometrica **21**: 503-546.

Blackwell, D. and L. Dubins (1962), "Merging of Opinions with Increasing Information." Annals of Mathematical Statistics **38**: 882-886.

de Finetti, B. (1937), "La Prevision: Ses Lois Logiques, Ses Sources Subjectives." Annales de l'Institute Henri Poincare **7**: 1-68.

Gibbard, A. and W. L. Harper (1978). "Counterfactuals and Two Kinds of Expected Utility." Foundations and Applications of Decision Theory **1**: 125-162.

Gilboa, I. and D. Schmeidler (1995), "Case-Based Decision Theory", The Quarterly Journal of Economics, 110: 605-639.

Good, I. J. (1967), "The White Shoe Is a Red Herring." British Journal for the Philosophy of Science **17**: 322.

Good, I. J. (1968), "The White Shoe qua Herring Is Pink." British Journal for the Philosophy of Science **19**: 156-157.

Good, I. J. (1986), "A Minor Comment Concerning Hempel's Paradox of Confirmation." Journal of Statistics, Computation and Simulation **24**: 320-321.

Goodman, N. (1965). Fact, Fiction and Forecast. Indianapolis, Bobbs-Merrill.

Hempel, C. G. (1945), "Studies in the Logic of Confirmation I." Mind **54**: 1-26.

Hempel, C. G. (1966), "Studies in the Logic of Confirmation", in Probability, Confirmation and Simplicity. New York, Odyssey Press: 145-183.

Jeffrey, R. C. (1965). The Logic of Decision. New-York, McGraw-Hill.

Korb, K. (1994), "Infinitely Many Resolutions of Hempel's Paradox," in Theoretical Aspects of Reasoning About Knowledge, R. Fagin (ed.), Pacific Grove, CA, Morgan Kaufmann:138-149.

Nozick, R. (1969), "Newcomb's Problem and Two Principles of Choice", in Essays in Honor of Carl G. Hempel. Dordrecht, Holland, Reidel. 107-133.

Savage, L. J. (1954). The Foundations of Statistics. New York, John Wiley and Sons.

von Neumann, J. and O. Morgenstern (1944). Theory of Games and Economic Behavior. Princeton, N.J., Princeton University Press.