Submitted to Physics Today

LBL-3069
Preprint

PARADOXES FOR PHYSICISTS

A. A. Ross-Bonney

July 1, 1974

Prepared for the U.S. Atomic Energy Commission
under Contract W-7405-ENG-48

TWO-WEEK LOAN COPY

This is a Library Circulating Copy which may be borrowed for two weeks. For a personal retention copy, call Tech. Info. Division, Ext. 5545

LAWRENCE BERKELEY LABORATORY
UNIVERSITY OF CALIFORNIA
BERKELEY

LBL-3069

DISCLAIMER

This document was prepared as an account of work sponsored by the United States Government. While this document is believed to contain correct information, neither the United States Government nor any agency thereof, nor the Regents of the University of California, nor any of their employees, makes any warranty, express or implied, or assumes any legal responsibility for the accuracy, completeness, or usefulness of any information, apparatus, product, or process disclosed, or represents that its use would not infringe privately owned rights. Reference herein to any specific commercial product, process, or service by its trade name, trademark, manufacturer, or otherwise, does not necessarily constitute or imply its endorsement, recommendation, or favoring by the United States Government or any agency thereof, or the Regents of the University of California. The views and opinions of authors expressed herein do not necessarily state or reflect those of the United States Government or any agency thereof or the Regents of the University of California.

I. INTRODUCTION

"Veniente occurrite morbo"[1] Persius (AD 34-62)

This paper was born as a one page appendix to a paper on the interpretations of Quantum Mechanics. But, like Topsy[2], it seems to have forgotten that and just growed, until now it is larger in pages, if not content, than its parent. The original purpose of the appendix was a brief elucidation of a comment on a particular application of the differences between physical and logical necessity. The present purpose is more general, the reasons for writing it more diffuse.

Firstly, paradoxes and puzzles are fun. More precisely, they are fun most of the time. But once they hook one it is virtually impossible to be free of them until one solves them, at least to one's own satisfaction. Yet, even if this appears done, the incubus is still not quite removed, for one is liable to feel the need to convince others of the correct nature of (and lack of cheating in) one's own solution. In my case, I have not yet read any satisfactory solution to Newcomb's[3, 4, 5] Paradox, or any write-up of the Class A Blackout. They, therefore, require exposure. For other more popular paradoxes, my solution is shifted in emphasis, or partially in content, compared to other solutions that I have seen or heard. This is because I want to relate them to methods in physics (and because I am a physicist).

Secondly, and implicit in the last sentence, I now believe that most, if not all, of these paradoxes illuminate in some way the

processes or procedures of physics. We all know that physics is not just interrelated with mathematics and the other physical science, but also with philosophy and logic; but generally we don't "waste" time contemplating these relations and influences. Whether such contemplation is, or is not, a *wasteful* distraction from active physics research it is certainly a distraction! Indeed, however, it does seem to me that any attempt to increase one's overall rather than detailed understanding of a theory requires more general considerations, and so hopefully, these paradoxes made a desirable, though small, contribution.

Thirdly, if it is too pretentious to hope to expand general understanding, at least this work has served to remind me that I should avoid certain types of loose thinking and may correspondingly perhaps, provide an amusing reminder to others. Wherever there is (was) a paradox you may very likely take it as a warning of a snare that will entrap you under similar, but less obvious, circumstances. Frequently also, paradoxes show us that we have discordant hidden premises; that in some way, initially obscure, we are accepting contradictory premises or inconsistent axioms.

Not all paradoxes are minor. For example, Russell's paradox (disproof of the existence of a set of all sets that do not contain themselves) produced a crisis in understanding the foundations of mathematics. The great Gottlob Frege[6] had the second Volume[7] on "The Foundations of Mathematics" in press when he received a letter from Russell. He added an Appendix[8] to that Volume wherein he describes how shattering it was to receive the letter; and also starts on a new

line of thought produced by Russell's paradox. Similarly Kurt Gödel's[9, 10] paradoxical theorem of 1931, showing the essential incompleteness of mathematics altered the flow of mathematical thought. Again, similarly Tarski's[11] work (published shortly thereafter), which showed that a consistent language cannot consistently describe its own meaning or truth, was of enormous importance to logicians. I contend elsewhere that, though not of this magnitude, the Einstein, Podolsky, and Rosen Paradox is crucial to *interpretations* of Quantum Mechanics.

No paradox here is of this order, but there are hints of mightier things in some of them. I fear that a philosopher, or logician, may find my treatment, especially of some paradoxes, rather cavalier, and perhaps even accuse me of cutting, instead of unravelling, the Gordian Knot. I am, however, addressing myself to scientists, as a scientist. This is not intended as an excuse for sloppy thinking but to explain a somewhat different approach and a brusquer verbal treatment.

Paradoxically, I write on paradoxes because I don't believe in them. More explicitly, I optimistically assume that the human mind can solve the puzzles thought up by the human mind. In one sense, outlined already, it is quite acceptable that a paradox may be insoluable: that the solution may be of the sort that can stop an argument. The disputants solve their argument by uncovering the previously hidden difference in assumptions each made; but they do not resolve their differences, since each may logically now reach a different conclusion. The next step of discussing the uncovered differences may, but need not,

be made. So the solution of a paradox is not quite like a mathematical solution. Our paradoxical solutions may also involve general thought, premises about the nature of the world, even psychological or emotional considerations. In this sense there is a necessary incompleteness; which is best illustrated in Section VII (in Newcomb's Paradox), and by the very incomplete discussions of confirmation in Section II and probability in Section V.

Some, or all, of my statements and analyses may, of course, not merely be considered cavalier but down-right wrong. I shall be interested in any and all written criticisms. But since even a simple paradox put across smartly in conversation is a very treacherous beast, I reserve the right to accept written rather than oral criticisms, and to respond correspondingly.

The various paradoxes in this paper are often cross-related. You will notice family resemblances. Thus it was not possible to construct a uniquely sensible order in which to discuss them. Where I feel it is necessary, or desirable, to solve one before another, you will certainly find them in that order. Otherwise, they are roughly in order of difficulty; but this is a matter largely of guess-work, since practical difficulty is a matter of both character and training.

In Section II there are some very simple paradoxes in deductive and inductive logic and in confirmation theory. In Section III, I resolve the Barber Shop Paradox[12] in my own way because I think it's fun (and because I think it is very important to hit the precise "point of falsehood" as much as to *prove* which argument is correct).

Section IV contains some slightly different time and semantic paradoxes. Section V touches on some conceptual difficulties involved in statistical arguments, and thus also in confirmation theory. Section VI describes a paradox hitherto, as far as I know, unpublished and a fine example of a "perfect-word" paradox. In Section VII, I believe that I provide a solution to Newcomb's paradox. This paradox is particularly interesting with its rather quantum mechanical and theological overtones. According to Gardner[4] several persons have tried to solve the paradox by bringing in Quantum Mechanics!

I end this introduction, by wishing that the reader shall be amused; and by hoping that I shall not, like the metaphysician, be damed as "A blind man in a dark room -- looking for a black cat -- that isn't there."[13]

II. LOGIC NECESSITY AND CONFIRMATION

"Necessity never made a good bargain"

"Some are weatherwise, some are otherwise"

Poor Richard's Almanac (1735)

Although most of us make occasional mistakes in formulating a logical argument, particularly if excited, these errors are usually relatively easy to trace. What can introduce confusion is a conjunction between discussion of physical reality -- the world of humanity or the world of physics, for example -- and the operations of logic (or mathematics).

When I say a logical error, I mean an error in argument not necessarily involving a factual error. Here is one of the simplest illustrations of this type of error. Suppose the following is a valid sequence ("hypothetical" of Section III) which I will label H_1.

"If (proposition) A is true, then (proposition) B is true." (H_1)

Proposition B is now found to be indeed true. The elementary error is to conclude that A must be true. Similarly if we have an hypothesis H_2:

"If A occurs at time t_0, then B will occur at time $t_1(>t_0)$" (H_2)

Even though A does occur at t_0 and B at t_1, it is erroneous to say this proves a causal relation between A and B. Yet we hear this

frequently as in "If you elect me, Mr. Claimall, President, the economy will recover." Mr. Claimall is elected. The economic scene does improve. What does Mr. Claimall say at the next election?

Now I will recount three simple paradoxes, where the paradox is caused by a confusion between logical necessity and physical necessity. Although these three cannot really be expected to generate a paradoxical impression on a physicist's or mathematician's mind, the first does illustrate an argument of this sort which once caused genuine difficulty, and the second erroneous argument is sometimes explained mathematically rather than logically; the third can still apparently cause errors.

In the first instance it is the English philosopher Hobbes[14] who argues. Thus: "Let the case be put, for example, of the weather. It is necessary that tomorrow *it shall rain or not rain*. If therefore it be not necessary that it shall rain, it is necessary that it shall not rain; otherwise there is no necessity about the proposition." (My italics) Besides showing that typically English concern about the weather, Hobbes displays for us a simple confusion between logical and physical necessity (or perhaps the physical necessity logically required by the theory). It is logically necessary that we not have concurrently, that a proposition is true and that the same proposition is not true (at least if it is not to be either meaningless or absurd). Hobbes' necessity tell us nothing about the weather. We cannot separate the italicized sentences into pieces as he does, because it is the combination which has logical necessity (assuming the world exists tomorrow, and so forth).

The second, but by far the oldest, of the three is Zeno's paradox of Achilles and the tortoise. After two thousand years most of the juice has been squeezed out of it. But since it illustrates a highly recognizable form, a confusion between logical necessity and physical necessity, I think it is worth a very brief re-analysis.

Achilles and the tortoise agree to run a two mile race. Achilles can run ten times as fast as the tortoise, so he generously refrains from starting until the tortoise has covered the first mile. Now when Achilles has run this mile the tortoise has proceeded another one tenth of a mile. When Achilles has run this one tenth of a mile, the tortoise is still one hundreth of a mile ahead. This line of reasoning can be continued indefinitely: it leads apparently to the conclusion that Achilles can never catch up with the tortoise. A simpler arithmetical argument, however, leads us to the conclusion that when Achilles reaches the winning post, the poor tortoise has still eight tenths of a mile to travel, and that Achilles actually passes the tortoise after one and one-ninth miles. Clearly there must be a fallacy in one argument or the other, and, equally clearly, it lies in the first argument.

For, if I could in any way continue to divide something of finite size into smaller and smaller pieces then I would certainly have more and more of them; though if all the pieces, undamaged, are placed together the total size of them all remains the same finite size as that of the original thing. There is a logically necessary connection between "smaller and smaller" and "more and more," but it exerts no physical compulsion on the nature of material objects; cannot compel

material to be continuous and cuttable, nor, in reverse, atomistic. That I can divide, in my imagination the time between now and 12 p.m. into smaller and smaller intervals of ever increasing number does not affect the hands of my watch as (so far) they inexorably approach that hour. My mental exercise does not tell me what is the smallest time interval that can be experienced by me in various circumstances. It assumes something about the nature of time but neither compels that there be or not be a smallest measurable time interval.

Thus the first argument applied to Achilles may suggest a basis for contemplating the mathematical concept of converging series, but it does not affect the outcome of the race. *Nor is it necessary for the resolution of the paradox that one know how to sum an infinite series.*

A presently more interesting, but not dissimilar, confusion between logically connected propositions and physical behavior was displayed by Leibniz.[15] He starts with the surely sound proposition that we cannot use a definition without first making sure that it contains no contradiction; since from contradictory concepts we can draw contradictory conclusions which is absurd. Then he says that he used to explain this with the example that the idea of a fastest motion contained an absurdity. Assuming that a wheel is turning with this fastest motion, he argued, then if one of the spokes of the wheel is lengthened so that it extends beyond the rim, then the end point of the spoke will be moving faster than a nail on the rim. Therefore, the motion of the nail is not the fastest motion. This contradicts the

hypothesis. He adds, in my opinion rather gratuitously, that although we might seem to have an idea of something because we understand what the words mean, yet the fact is that we cannot have an idea of impossible things.

Certainly, I would maintain that Leibniz' initial contention is correct. In all the rest, however, he appears to confuse logical impossibility with physical impossibility. It is logically impossible (more exactly we ought to say "impossible logically") to construct a hypothesis that states, or implies, that there is a fastest speed for material objects and that it is also possible for Leibniz' wheel to exist. On the conceptual side, without necessarily believing six impossible things before breakfast, it does not seem in the least impossible for the human mind to "have an idea" of all sorts of things which cannot all be possible at the same time. It may perhaps only be the exceptional person who is simultaneously well-informed about the present "laws of nature" and who formulates concepts which appear to contradict them. Such a person need not abandon the laws of logic when evaluating the possible consequences of their concepts.

To a physicist this may all seem a bit tongue-in-cheek, because all physicists nowadays are so familiar with the special theory of relativity; but in 1971 Professor Antony Flew[16] in a generally admirable book on philosophy, can refer to "the elegant proof by Leibniz of the impossibility of a fastest conceivable motion." On disputed territory it is essential that we retain clearly in our minds the separation between logical impossibility and physical impossibility. I am not

arguing here that logic itself is totally separable from the world and language, but that in any non-universal problem the separation can be made. In case I seem to suggest a superiority of physicists over philosophers, I draw to your attention the beliefs of most (with notable exceptions) physicists before 1957. Before that year, it was largely taken for granted, as natural law, that parity was conserved, that the universe had no right or left handed preferences. Moreover many of us assumed this, without logical support (that might be supposed to come from even minimal examination) that non-violation was experimentally confirmed in all known interactions. Much more appropriate, though less topical, illustration was the belief that the existence of electromagnetic waves logically required the existence of an aether. Who did not, in another era, accept this?

These humbling recollections lead us naturally to the logic of the process of confirming or disproving physical hypotheses; a prospect already suggested in the discussion of H_2. According to Salmon[17] this process is still far from being understood. Since it appears also to contain paradoxes, and in addition confirmation is to some degree involved in some of the paradoxes in this paper, I will summarize briefly here a few of the problems.

At first blush it can seem that in inductive reasoning we commit the elementary error described in the second paragraph of this section. The experimental scientist will deduce a true-consequent of a theory. If this consequent turns out to be factually true then it is supposed to confirm the theory. But its factual truth need in no way

do this (as in the deductive fallacy) yet it certainly can. How can this be? It can be in general as follows: if A in H_1 stands for the theory and B is not just a single consequence but its logically equivalent (as, for example, B = not not -A) to A then an increase in the probability that B is true will correspondingly increase the probability of A. But let us take care! We may well run into a fallacy as bad as H_2.

Suppose we accept the overall supposition that animals exist with legs, eyes, mouths, etc., and we define a dragon as a scaley animal with wings that breathes out fire. I now propose the Bonney Theory:

"Dragons have red eyes." (B1)

then I say that this is logically equivalent to

"All animals that don't have red eyes are non-dragons." (B2)

So now I can argue that B2 is exceedingly well-confirmed by the observations of zoo-ologists over many years, with absolutely no credible exceptions. In other words it is highly probable that it is true. Therefore B1 is very well-confirmed, and I have established the Bonney Theory on a large experimental basis. The fallacy is a bit like that of the Barber[18] (adult male) in the village of Noville who shaves all the men in Noville who do not shave themselves. In order to have a logical equivalence to B2 I must add to B1 "or dragons do not exist." Alas. If I have seemed unfair by choosing dragons one can make similar structures with real objects. For

example[17]

"All crows are black." (C1)

versus

"All un-black things are non-crows." (C2)

or more restrictively,

"All un-black animals are non-crows." (C3)

or still more restrictively,

"All un-black birds are non-crows." (C4)

A little work will see why we prefer in general to confirm C1. Suppose confirmation of C4 means considering un-black birds at random. It becomes a matter of population size. Suppose there are M crows in existence, m_1 of which are black and m_2 of which are not. Suppose there are N other birds n_1 black and n_2 not black. We try to confirm C4 by examining n colored birds. The chance we would not find a colored crow is

$$\left(\frac{n_2}{m_2 + n_2}\right)\left(\frac{n_2 - 1}{m_2 + n_2 - 1}\right)\cdots\left(\frac{n_2 - n + 1}{m_2 + n_2 - n - 1}\right) = \prod_{\gamma = 1}^{n}\left(1 - \frac{m_2}{n_2 + m_2 - \gamma + 1}\right)$$

If the n_2 is large -- that is there are a lot of non-black non-crow birds --

compared with the number n we examine then the probability that we shall not find a colored crow is approximately

$$\left(1-\frac{m_2}{m_2+n_2-n}\right)^n \cdot \left(1+\frac{n}{n_2-n}\right)^{n_2+\frac{1}{2}} \cdot \left(1-\frac{n}{m_2+n_2}\right)^{m_2+n_2+\frac{1}{2}}$$

Even if *no* crows are black ($m_1 = 0$, $m_2 = M$), we shall get a very small chance to disprove C1 if the total population of non-black birds ($M + n_2$) is large compared to that of the M crows. Automatically C3 is still less help, and C4 less still than C3 (almost zero as the population becomes almost infinite). An examination of crows, however, would have led to immediate results in the case where m_1 is zero; and in general the probability of finding only black crows after examining n crows where some crows ($m_1 \neq 0$) are black is

$$\left(\frac{m_1}{m_1+m_2}\right) \cdot \left(\frac{m_1-1}{m_1+m_2-1}\right) \cdots\cdots \left(\frac{m_1-n+1}{m_1+m_2-n+1}\right)$$

Of course in a realistic case we cannot state the probabilities in numbers since we do not know them or we would not be experimenting, but arranging them in this way makes it clear why it may be irrelevant, or nearly so, to try for confirmation of a true consequent of the theory.

Before proceeding any further I shall attempt to dispose of a puzzle that, according to Salmon, has been a part of confirmation theory for the past thirty years. It seems to me unworthy of so much discussion, so it may be there is a fallacy in my argument.

"A ballet dance of bloodless categories" (F. H. Bradley, 1846-1924).

The grue-bleen (Nelson Goodman's) paradox can be put as follows. We arbitratily choose a future time T. An object is grue if it is green at all times $t < T$ and changes to blue at T and so remains $t > T$. An object is bleen if it is blue at all times $t < T$ and green thereafter. The hypothesis that certain stated objects (O_1's) are grue and others (O_2's) are bleen appears to be confirmed by the facts that O_1's are (and have been in the past) green and similarly O_2's are blue. Is this not confirmation that O_1's are grue and O_2's bleen? To me the answer is uncomplicated for the simple reason that the theory cannot even be stated if green O_1's and blue O_2's do not presently exist. To emphasize, no hypothesis is acceptable unless it accords with highly confirmed past-present data, if it is supposed to be valid where any such data exists. As a more sophisticated example, consider the special theory of relativity. If this theory did *not* duplicate (to the required accuracy) Newtonian mechanics at low speeds it would not have been worth considering. It was not confirmed because it did so; it was necessary that it should do so. It is additionally hard to discuss the grue-bleen hypothesis because it makes no scientific sense. Scientific hypotheses can range from small and tentative, when few observations have been made to succesively larger and more comprehensive as more and more data has been collected. Or they can be generated by a desire for more sweeping generalizations of greater beauty and elegance. The general theory of relativity surely belonged at its inception by Einstein to the latter category. But hypotheses are not generated at random! Scientists like

like patterns not splodges. Incidentally it is not correct to say as Goodman does (according to Salman[17]) that if we start with grue-bleen then green means grue prior to T and bleen thereafter. For suppose at time greater than T we find, to my consternation, that O_1's are now blue and O_2's are green. O_3's which were and are green cannot now be said to have been grue and now are bleen. Grue and bleen are not so defined.

To return to more general considerations: Even if we cannot readily assign a numerical probability that a hypothesis is true we can agree in principle that an experiment confirms the hypothesis if it increases the probability that it is true. However, a hypothesis can usually only be tested in conjunction with other hypotheses, and thus problems may, and genuinely do, arise as to what we are confirming or disconfirming. Even if we have two competing hypotheses α and β and try to do a crucial experiment which ideally has result a if α is true and β is not, b if β is true and α is not, and c if neither are true, it is still likely that some third hypothesis (or more) is necessary to evaluate the experiment.

Suppose there are two excellent athletes Jones and Brown. They have just run in the same 1 mile event. However I and my friends do not expect to hear the results for another hour. Having found some records of previous events in which either Jones or Brown (but not both) have competed I estimate either has individually a 50% chance of breaking the one minute mile. Jones broke the one minute mile today is proposition A. (50%) That Brown did is proposition B.(50%) So I assign a 75% chance that either one or the other or both will have broken the one minute mile

today (Proposition C); 25% that both will (D). Then I obtain much more comprehensive records and find that they nearly always competed in the same race, that either of them considered separately has only a 40% chance of running the mile in less than a minute, but that there is a 90% chance that either they both have, or neither has. Then I have "disconfirmed" A, "disconfirmed" B, "disconfirmed" C -- (that at least one of A or B or both is true) and yet "confirmed" D, the probability that both A and B are true since this probability has increased from 25% to 35%. This has only the momentary appearance of a paradox yet is enough to make one wonder if we are not often a little too glib, as scientists, and should perhaps pay more attention to confirmation theory. On the other hand this type of confirmation has only a slight link with the actual way scientists consider a theory confirmed.

III. DODGSON'S BARBER SHOP PARADOX

"Everythings got a moral if only you can find it"
Lewis Carroll

This is more a strictly logical paradox: that is a paradox in logic. The conclusion is manifestly false, as is stated in Dodgson's presentation. The problem is to pin down the false logical step. In the next paragraph I shall summarize the paradox briefly, but it should be read in the original. A delightful version taken from hitherto unpublished galley proofs of the second (unpublished) volume on Symbolic Logic by Rev. Charles Ludwige Dodgson (better known as Lewis Carroll) is to be found (page 43) in an article by W. W. Bartley III.[12] I make no apology for including this paradox, since according to Bartley, it is still the subject of "bewildered controversy" among logicians today. But I apologize to the reader, and to Dodgson, for not resisting the attempt to formulate a solution as a continuation of Dodgson's presentation, and in his style. I hope he would appreciate the attempt as a tribute to himself, and forgive any clumsiness.

The paradox is presented in the form of an argument between Uncle Jim, Uncle Joe and "Cub," their nephew, while all three are on their way to Allen's, a barber's shop. The barber's shop is staffed by three barbers Carr, Brown and Allen. Uncle Jim wishes to be shaved by Carr, and Uncle Joe undertakes to prove that Carr must be in. During the argument the Uncles agree that the shop will not be left empty and that Allen, since having a fever, is now always so nervous that he will

only go out accompanied by Brown. At this point I will interject the definitions of a hypothetical: and of its protasis and apodosis. A hypothetical is a statement of the following form:

> If A (is true) then B (is true). A and B are propositions.
>
> The protasis of the hypothetical is A, the apodosis is B.

The outline of Uncle Joe's agrument is as follows: If Carr is out then [if Allen is out then Brown is in] is true. This is because the shop cannot be unattended. The apodosis of this theoretical is false since we know that the hypothetical "If Allen is out Brown is out" is always true. It follows since the apodosis is false the protasis must be false. Thus Carr must be in.

Included in Bartley's article[12] are two resolutions by Dodgson. My own commences where Dodgson's presentation ends in mid-sentence at the crucial (?) point when they enter the barber's shop as thus: "and on going inside we found ---" now I continue "---that Carr and Brown and Allen were all in their shop, with Brown and Allen each busily engaged in attending the needs of two elderly gentlemen. Uncle Jim's face was a fine sight! It expressed wonderfully the conflict between his pleasure that he would be shaved by Carr, and his regret that Carr's absence could not take the wind from his brother's sails. I was glad my bet had been refused. Uncle Joe was gracious enough to remain silent.

"After Uncle Jim had been shaved, Uncle Joe decided to get himself shaved by Carr too. Uncle Jim had obviously been making the most of his quiet period to contemplate the absurd reductium (that's mine) of Uncle Joe's. As soon as Uncle Joe was trapped fast in shaving soap, he began.

> 'You said that the truth of a Hypothetical didn't mean that its protasis was true or even possible? It just has to be a valid sequence?'
>
> Uncle Joe appeared to grant it.
>
> 'You said that "If I were to run to London in five minutes I would surprise people" was a true Hypothetical.' Uncle Jim continued. 'Then I say that "If I could run to London in five minutes I wouldn't surprise people" is a true hypothetical. No one would be surprised because they would be perfectly used to seeing me. If I could do it, I certainly should do it often. So there are two true Hypotheticals that are contradictory. You must conclude that protasis is false and that I can't run to London; which I told you anyway.'

"Uncle Joe looked as if he would dearly like to interrupt and there was a corresponding flicker of uncertainty on Uncle Jim's face. Recovering he turned to me.

> 'Do you know how to disprove a hypothetical my cubeb?'
>
> 'What's that?' I said before I could stop myself.

'A sort of pepper berry," replied Uncle Jim with satisfaction.

'I *should* know how to disprove a hypothetical my cubquacious Uncle' I said, 'after listening to you two so often.' I declared, 'If you can find one example where the protasis is true but the apodosis is false then the hypothetical is false. It is invalid.'

A little to my surprise both Uncles looked approving.

'All that "chopping logic" has taught our minor at least one major point. He is more than a two-cubit cub.'

In fact, I am more than two and a half cubits, but I let it pass.

'There remains one proposition and one hypothetical which we both agreed are true.' Uncle Jim said, 'I will call the first X_1 and the second Y_1. X_1 is "Carr, Allen and Brown are not all out at the same time." Y_1 is "If Allen is out, then Brown is out." We both accept that they are both true, don't we? Is one truer than the other? Cub, which is truest?'

'You know that you can't have a truest -- a truer! Either something is true, or it is not.'

'Good. So I say that Carr is out then X_1 is still true, since it is always true. Indeed, if Carr is out the hypothetical X_2 is true. X_2 is "If Allen is out, then Brown is in." As a logical sequence it is valid when

Carr is out. But Y_1 is always a valid hypothetical whether Carr is in or out. So if Carr is out the hypothetical Y_2 is certainly valid. Y_2 is "If Allen is out, Brown is out." Thus, *if* Carr is out we have two true, valid, hypotheticals. Each of these true hypotheticals has the same protasis but has a contradictory apodosis. Thus if Carr is out, the protasis of X_2 (and Y_2) must be false. Thus it follows that Allen is in. And he added emphatically if ungrammatically, 'if "Carr is out" is a false statement (that is, if Carr is in) then it doesn't matter to your over-reaching hypothetical that *then* X_2 is a *false, invalid,* hypothetical. Indeed, it *is* false then for if Carr is in the "If Allen is out, Brown is in" is false, meaning as a *hypothetical* it is not then *valid.*'

"At this point it seemed that Uncle might suffer apoplexy from his enforced silence. But Uncle Jim ran out of breath, and Carr finished his job. My poor Uncle Joe was freed. Allen and Brown were again occupied with two freshly arrived gentlemen.

"Uncle was just about to inundate his brother with verbal floods when Carr bade us all a courteous farewell, and quitely departed from the shop. That silenced even Uncle Joe momentarily. But, I wasn't surprised when he said '-------'."

IV. TWO TIME PARADOXES

"If you believe that you will believe anything"

Attr. Arthur Wellesley

Some paradoxes use both time and verbal sleight of hand for their effect. One of the simplest illustrations of this type of paradox is the old one[18] about the Crocodile and the Mother. The Crocodile snatches the Mother's baby from the banks of the Nile. She pleads with him to return it. He says that if she predicts correctly what he will do, then he will return the baby unharmed. She exclaims, "You will eat it." The Crocodile promptly responds "Now I cannot return your baby because if I do you will not have predicted my actions correctly." The quick-witted Mother hastily says, "You cannot now eat my baby for if you do then I shall have predicted your actions correctly." It is to be assumed that the Crocodile is honorable. Dodgson puts it "We assume, of course, that he was a Crocodile of his word; and that his sense of honour outweighed his love of Babies."

The factual possibilities are only two. Either the Crocodile eats the baby or he restores it.[+] In either case he breaks his promise. In the first case he fulfills the Mother's prediction. He would have to restore the baby in order to keep his promise and, alas, he can no longer do this. In the second case, by restoring the baby he makes the Mother's prediction false and breaks his word. Since there is no way he can keep his word there seems little doubt that he will eat the baby.

The misleading *word* here is honorable. Since the Crocodile is quite obviously no fool he has no doubt worked out the consequences. Incidentally, if the Mother had said "You will return my baby," then clearly he could have eaten the baby and kept his promise. Thus he has probably already decided to retain his tasty morsel. More important, he knows that he may have, by his original suggestion, put himself in a position where any course of action that he later takes will lead him to break his word. No "honorable" crocodile would do such a thing even if honor be allowed to co-exist with cruelty.

I think there is a little more to it than this. Both the words prediction and promise imply a direction in time. A prediction P (made at time t_1, for example) precedes the events (E) that it predicts (which occur, or do not occur, at time t_2, say); and it cannot be said to be true or false (at time t_3, say) until the event has occurred or vice versa. Taking the direction of time as t increasing, then $t_1<t_2<t_3$. The failure to put a possible equality between t_2 and t_3 is not arguable since I used the word "said" in order to avoid this side-track. So we have with time arrow

Predict → events → Predict is true or false (PTF)

Similarly a conditional promise can be summarized. If A occurs (time t) then (I will do or cause) B (time $t + \Delta t$) where Δt must be positive however small. Thus if B is E, it follows $t + \Delta t = t_2$. or $t < t_2$. Consequently it is now logically absurd to substitute for A occurring at

time t, *PTF* which occurs at time t_3, which is greater than t_2! Thus the Crocodile is making a "false" promise of a different order. It is absurd in rather the same way that *some* self-referencing statements are, such as the famous paradox of the Athenian who states that all Athenians are liars; or the statement "This statement is false."[27] Tarski[11] argues that a consistent language cannot contain methods for describing either the meaning or truth of its own statements. This might also serve as a warning to those who wish a mathematical theory to generate its own physical interpretation. To summarize the argument then: prediction and promise both predicate a direction in time. If we accept the words as meaningful then we must exclude certain promises (and predictions) as unallowable in conjunction rather than false or true.

Rather similar to the Crocodile is the story of a law suit. P (Protagoras) makes a contract to train E (Eathius) to be a barrister. The financial agreement in the contract is that E pays half the fee immediately and the other half is to be paid only on a contingent basis. If E wins his first case in court then he pays it, if he does not then this half is waived. After the training is complete E does not quickly obtain a brief. P becomes impatient and brings suit against E in order to get the other half of his fee. Before the judge he argues "If I win, E must pay me by the Court's decision, if E wins then he must pay me because he will have won his first case." But E argues "If I win then that means I shall not have to pay, whereas if P wins I shall have lost my first case and shall not have to pay." If I were the judge I would say something like the following:

"This action is before me because I am the ultimate authority on contracts. The job of this Court is to settle disputes arising between persons under contractual obligations to each other. In this, the Court must be guided by knowledge of the law and understanding of the particular contract. It may find that a contract has been broken by one or other of the parties to the contract. It may find that the Contract was illegal and therefore not binding. It may find that though the contract was legal, and made in good faith, subsequent events have made it unenforceable. In all cases, my decision is final and binding on both parties. In this case the Court finds that the contract is unenforceable. For if the Court's decision were to be that E pays then the contract would be broken, because he must pay although he has lost his first case. If the Court decided that E need not pay, then the contract is broken because he must not pay although he has won his first case. Thus the contract is unenforceable. Thus we need only concern ourselves with the good faith and legal knowledge of the disputants. The reason that it is unenforceable is doubly due to the actions of P. By bringing this suit he has made it so. Furthermore, the contract was made when P was an attorney and E was not, and therefore the making of the contract can

be assigned primarily to P. I therefore find against P."

Just as the first solution for the crocodile seemed alright but somehow inadequate and therefore needed further analysis, so does this. A similar sequential analysis can be made but it is on the same lines and therefore will not be included here.

V. PROBABILITY AND THE FATAL GENE

"What we anticipate seldom occurs; what we least expect generally happens."

Disraeli

Probability Theory

Of the various branches of mathematics I think that probability theory is one of the most conceptually difficult. Anything much beyond the throwing of a well-balanced die is complex and even this, itself, can be difficult to grasp. However, most of us can accept that there is a 1/6 probability that one particular face of a die will appear. One might make the, at least apparently, simple argument which follows. If we computed mechanically the result of a particular throw onto a "flat" surface we would obtain a particular result. If we computed all reasonable throws then we would find that the particular result occurred precisely 1/6 of our computed throws. Thus we conclude that, if the throw is random, then that particular result has a probability of 1/6. This argument allows us also to ascribe probabilities for random throwing of a loaded die. Suppose we measure the mass distribution of a die which, unlike the first one, is not uniform. Then, again in principle, we compute the results of all reasonable throws. Perhaps one particular result appears 9/60 of the time instead of 10/60. Then we ascribe for actual future throwing of the die a probability of 9/60 for this result. This argument is circular in one major aspect. For the perfectly cubical

Die we started with equal probabilities for seeing any *face*. We "explain" it by ascribing equal probabilities to different *throws* -- for example to the initial direction and rotation of a particular height and strength of throw. We always return to some set which we simply accept as equally likely, perhaps a set of events, or a set of numbers; and label these events, or numbers, random.

Despite its circularity, this argument can assist in understanding the manner in which we ascribe probabilities to predictions when we have a theory and also when we do not; and how such predictions must be limited.

Let us return to a particular die and use our most accurate techniques to measure the mass distribution. Suppose now that our measurements, and their accuracy assure us the cube is so well balanced that a man throwing it should not uncover any discrepancy in ten thousand years.[20] We then ask Mr. Smith, our lab technician, to stand in a particular position and throw the die while we record the results up to a prescribed number, call it N, of throws. Naturally we don't expect that five, for example, will have turned up precisely 1/6 of the number of throws, however, using the same arguments based on random throwing we can ascribe a probability to a particular result we may obtain. Let us further suppose that the calculated probability for our particular (total) result is one millionth; that the five has been face up the most, say n_5 times, the two the least, and that the other four faces have shown up a roughly equal number of times. Had we not already investigated the mass distribution of the die it seems likely that we would now assume

that our result was not due to chance but that the die was loaded.[21] Consequently we would either attempt to determine this by direct investigation or, if that were impractical, we would re-assign what I might call an experimental probability to the chance of each face turning up. Thus for the five-face we would now assign n_5/N instead of 1/6. If we were unethical we might use our die to go out and make bets.[22]

But we did elinimate a loaded die. We have not, of course, eliminated "chance." But I think most of us would instead regard with suspicion the throwing habits of our technician, Mr. Smith. We could probably eliminate this possible source by blind-folding and gloving Mr. Smith in future runs. But if we do this we may also be eliminating our opportunity to determine whether we have trickery[21] or not, for we shall have informed Mr. Smith indirectly of our suspicions. Thus, even if later on we repeat the original experiment under identical physical conditions, Mr. Smith, unless stupid, will know we suspect him and he may choose to suspend his skills and throw at random. To be able to use our experimental probabilities sensibly the experiment must duplicate the original one. Attempting to do this would compel us, at very least, not to admit to analyzing the results, and suggest also that we install some camouflaged movie cameras when Mr. Smith was not around. If in the next sequence the new numbers still give fractions similar to the preceding, we should certainly examine our visual records very carefully. If in the second sequence the effect seems to have departed, we will assign our original results to chance, or assume we gave ourselves away, or that Mr.

Smith tired of his devices. If either of the latter two assumptions are made our only remaining opportunity is to interrogate the technician. It might be emphasized, before deserting Mr. Smith, that this case illustrates not only in what way we assign "experimental probabilities," but most particularly how our course of action is complicated when human behavior is involved in an experiment.

Painful Inherited Disease

Another example of use and misuse of statistics and probability is the story told by Martin Gardner,[4] and used by him, to display perfectly wild reasoning (in the sense of quite obviously wrong reasoning). A man knows that one of two men S and T is his Father (presumably with equal à priori probabilities). S had a genetically dominant gene that produces a painful, and soon fatal, effect in the mid-thirties: he is dead. T does not have the gene. There is a well-confirmed theory that persons who inherit this unpleasant gene also inherit a tendency to display intellectual and scholarly behavior. I will assume that well-confirmed in this context means that medical statisticians have discovered a correlation between possession of the gene and the above behavior. The man would have prefered to go to graduate school, but also had a chance to become a professional baseball player. However, he prefers a long life playing baseball to a short one in academia. He reasons that he should choose baseball and thus reduce the probability that he has the fatal gene. Like Gardner I think most people will agree that his reasoning is fallacious, and obviously so. Since his gene content is already fixed, and has been since his conception, his actual behavior can neither help nor hurt him,

What I want to emphasize is the relation of this to our understanding of experimentally derived probabilities. Suppose (with some difficulty) that there were a large number of persons in the same position as our man. Suppose further that, despite a preference for academia, many of them do argue in this manner. Then the next generation of medical statisticians are going to find a much smaller (perhaps statistically not significant) correlation between possession of the inherited gene and academic pursuits. Knowledge of experimentally derived probabilities have changed the "probabilities" derived from the next experiment. This type of effect is, of course, well recognized by psychologists and testers of pharmaceutical products. Prophecies can not only be self-fulfilling but also self-defeating.

Gardner points out most correctly that although the above story appears intuitively to cast some light on Newcomb's Paradox, yet Newcomb's still seems obscure and therefore it is necessary to explain why Newcomb's is more difficult, if it is similar. I propose to attempt this: to show where they can be considered alike and where different. For this reason I have tried, above, to make explicit some possible implications of the gene story. For a longer discussion of the fallacy of the man's argument Gardner's article[4] should be consulted.

But preceding this I will explore another minor paradox which also casts light on misleading verbal assumptions.

IV. THE CLASS A BLACKOUT

"I've a great fancy to see my own funeral before I die."

Maria Edgeworth

For a couple of years I worked at the Clarendon Laboratory and became a member of Oxford University, that seat (or hot-bed) of philosophy. Perhaps because of the influence of Ryle[24] and Ayer[25] and others, much of the philosophical discussion was on semantics and paradoxes. As a child I remember hearing Professor Joad, a philosopher naturally, speak on a program on the B.B.C. called "The Brains Trust." He started almost any remark with "It all depends on what you mean by -----." This sort of expression resounded when I was at Oxford. Although I never personally spent much time with philosophers or logicians, nine out of ten of all the people whom I did meet engaged in physics and other fields of research work were prepared to discuss "philosophy" at the drop of a hat. I encountered the paradox of the Class A Blackout (C.A.B.) a number of times. I never heard it discussed in any formal encounters between Physics and Philosophy; nor do I know who invented or discovered it. I have not seen anything written about it. For this reason alone it is probably worth including; and also because it is different from the others. It is, I think, a largely semantic paradox, and only escapes triviality because of the importance of the word with which it plays. That word is "know."

The problem is as follows: In the Royal Navy the ships are required to practice sailing in Black Out conditions with a certain frequency. There are two types of blackout: Class A and Class B.

The handbook which lays down all the rules and regulations governing Her Majesty's ships, describes the two types. "The blackout is Class A if none of the crew, except the Captain and First Officer, know before the six hours preceding its occurence that there is to be a blackout the coming night. All other blackouts are Class B." One Sunday at 5 a.m. the Captain orders that notices be posted and announcements made over the ships communications system at 6 a.m. All these say "There will be a Class A blackout within the following week. It will occur before 6 a.m. next Sunday."

A bright young sailor called Mike who hasn't yet forgotten that he used to be a graduate student makes the following argument to a group of sailors. "If there has not been a Blackout by then, then on next Saturday morning I shall know that there will be a blackout that night. But if I *know* there will be a blackout, then it *cannot* be a Class A blackout." To his ignorant friends he read the excerpt defining a Class A blackout. "Therefore it follows that there cannot be a Class A blackout next Saturday night. So when I wake up on Friday morning, if there has not yet been a blackout, I shall know that there will be one on Friday night. But if I know this on Friday morning, then there *cannot* be a Class A blackout on Friday night. From which it follows that on Thursday morning, if there has not been a blackout yet, then I shall know that it must occur on Thursday night. So it cannot occur on Thursday night, or it would not be a Class A blackout. So on Wednesday morning . . ." "Yes," said Jim "there can't be one on Wednesday night. So none on Tuesday night. So none on Monday night. So . . ."

"So I know it must occur tonight." interrupted Mike disliking to lose the spotlight. "But if I know it must occur tonight, it cannot be a Class A blackout. There cannot be a Class A blackout at all."

"You'll look a proper fool when he has one," Jim grinned. "If there can't be a Class A Blackout you don't know when a blackout may occur. If you don't know (and we don't) then the Captain can have his Class A blackout whenever he likes."

"But if he can, then on Saturday morning I shall . . ."

"Not again." said old sailor Jack, who hadn't gone beyond grade school. "Everyone knows that if that son of a Captain says Class A blurry blackout there will be a blurry Class blurry A blurry blackout. And you'd better *know* it boy!"

The solution to this particular paradox never seemed to me very difficult. In Oxford I rarely appeared to convince any of its afficionados verbally that I had the solution, so now I will try in writing. I should like to see comments on its adequacy or the reverse.

Obviously the argument leads to an absurdity, since either a Class A blackout is possible or it is not. I have heard it argued, and tried to argue this myself, that the Captain's original statement is of a nature similar to a self-referencing statement of the type mentioned in Section IV. That is like "This statement (in quotation marks) is false."[28] I do not think this is quite adequate, although I would prefer to be a crew-member on a ship whose Captain did not make such statements. Rather I think it is an example of what I mentally list as a "perfect word" paradox. The perfect word most involved is KNOW.

If the word know is used about past or present (almost present?) events one would certainly require a considerable amount of argument in any attempt to make the meaning precise, if this is possible. Nonetheless it has a far more "perfect" connotation than any usage we know (!!) how to make about *future* events.

In a world view that has a preferred direction in time there must always be some question of the "truth" of a future event. Even if we should accept a so-called Newtonian, or completely causal view (not that Newton himself would necessarily have espoused this as a world view, or other-world view), it would seem difficult to decide on the "truth" of a possible future event. Even if, in a limited region, we suppose ourselves to understand the causal laws at work, surely not even the most self-confident person can suppose themselves to be certain that some extraneous event may not throw off a true-false calculation. To exclude such a possibility totally we should need to know not only all the physical laws, but also all the initial conditions *and* not be included ourselves. To summarize: however arguable may be the use of the word "know" with reference to the past, its use about the future must involve in some much more strenuous way a degree of confidence, or probability. Thus, "I know Ms. Rose will come at eight o'clock" may express anything from wishful thinking to the solid conviction based on knowledge (can't avoid the word ! but always in the past tense) of the speaker, Mr. Brown, that he has never known Ms. Rose to break a promise, that he has plenty of experience of Ms. Rose's behavior, and, perhaps, that he has never heard any mutual acquaintance describe a possible instance when she did break

a promise. Nonetheless it cannot be concluded for certain that this time she will not break her promise (no man is an honorable man until death completes his life); but more strongly there is always the possibility, as far as the speaker knows, that some accident may make her unable to keep her promise. If she, and perhaps her car, are healthy, and times are peaceful he may feel pretty confident. "Know" in other words has degrees of conviction (belief). An example of a yet more confident "know" are expressions like "I know the sun will rise tomorrow." For modern man there is not only a man's own experience and his ancestor's past experience expressed in the statement but also an acceptance of a vast and well-confirmed body of knowledge. For example, there exist now astronomical observations of many sorts such as that on life-time of stars at particular developmental stages, and much well-confirmed theory about the nature of the sun, a well-confirmed theory of the gravitational behavior of the planetary system (including the earth, naturally), and a belief in our ability to detect any incoming disturbance which could be adequate to perturb the planetary system before tomorrow. Thus the "know" about the sun is orders of magnitude more assured then the "know" about Ms. Rose. This is substantiated by the use of the former type of "know" to emphasize the certainty of the latter. For example, Mr. Brown might well say "I know Ms. Rose will come at eight o'clock tonight, as surely as I know the sun will rise tomorrow."

In this light we can discuss the various "knows" in the C.A.B. paradox, (particularly those in Mike's argument).

First, there is the original "know" in the regulations. This particular one brings back the law, but, one might assume that the intention of the formulators was that a Class A Blackout should mean that the crew shall not be warned in anyway that a Blackout is ordered earlier than six hours before it is planned to occur. But this is just a guess, since what they appear to be saying is that none of the crew save the Captain and First Mate, shall be highly convinced of the coming blackout until it is less than 6 hours off. The extent to which rumors derived, perhaps, from conversations between the Captain and First Officer overheard by a sailor might produce such a conviction (and prevent a Class A Blackout) is unclear. This amounts in the end to this and only this: a naval court is the body that decides whether the Captain in fact, at some past time, did have a Class A blackout on his ship. Legally and actually it only makes sense to *decide* the Class A-ness of an actual blackout *in retrospect*.

Nonetheless it cannot be nonsense to ask if certain circumstances are compatible with a C.A.B.? So we must still be more explicit about the confusion in the C.A.B. paradox.

To repeat: It is not feasible to make an absolute "know." Even if the Captain had merely said "There will be a blackout before next Sunday morning," it may be found next Sunday morning that no blackout occurred. The ship might sink. The Captain might die whereupon the First Officer might suspend the blackout. Less drastically, the Captain might countermand the blackout because he was involved in a rescue operation, or for some other reason. The degree of conviction,

of knowing, of the crew will depend both on their estimate of the Captain's character, for the more sophisticated at least, and of their general circumstances. Mike's argumentative "knows" are both misleading, and heirarchical. Take his first "know" for Saturday morning; then he says he will know there must be a blackout that night and therefore he will know there will not be a blackout that night (because it could not be a Class A blackout). Strictly we don't need to go any further. For whatever it might be appropriate to use the word know; we clearly can not "know" two contradictory things at the same time (however common a failing this may appear to be!). Mike can only say that he will be confused on Saturday morning. If his predominant conviction is that there will be a blackout that night, he might say he knows it, and perhaps laugh at the Captain. If his predominant conviction is that there will not be one then he may claim to know that, and relax accordingly. One suspects he will be unable to decide! Whichever position he takes we can see a heirarchical effect on the rest of the argument. Whatever degree of conviction he might have on Saturday morning will decrease as we go backwards through the earlier mornings. If members of the crew, in this unelevated sense, "know" different things; it is dubious that a naval court would accept that any know anything (maybe?). Personally, were I a member of the crew, I would try to be alert all week and my best guess (to *unknow* as many guesses as possible) would be a blackout Wednesday, or Thursday. This is not a logical guess in the ordinary or even extended sense, but given no "knowledge" of the Captain, I would expect the maximum

unpreparedness midweek -- and the reverse just after the notice, or before the expiration of the time period. But "know" -- NO.

"Gawd knows, a'E won't split on a pal."[26]

VII. NEWCOMB'S PARADOX

"And of all axioms this shall be the prize
'Tis better to be fortunate than wise'"[29]

Newcomb's Paradox

This very fine recent paradox was discovered (invented?) by William A. Newcomb, a theoretical physicist at the Lawrence Livermore Laboratory of the University of California. In 1970 it was first written about in a paper[5] by Robert Nozick. Recently[3, 4] Martin Gardner has twice written at length about it in the Scientific American. For a longer, and probably fairer, exposition of it the reader should consult their enjoyable articles. A brief summary follows.

There exists a very superior Being whom I will call the G. P. as an abbreviation of the Great Predictor. He has studied human psychology extensively and has already predicted your choices in very many situations. I'll call this A_1 (assumption one). Gardner adds "and the choices of many other people in the situation to be described." I will call this assumption A_2. Thus you have enormous confidence in his predictive abilities. There are two boxes. In Box 1 is $1,000. Box 2 contains either $1,000,000 or nothing.

The choice offerred you is to take Box 2 *only* or to take both boxes. If the G. P. predicts that you will take only Box 2 he puts $1 million in it. If he predicts you will either take both boxes, or that you will decide in some random way what to do, he leaves Box 2

empty. Gardner adds "you know these facts, he knows you know them and so on." I will label this quoted sentence A_3. The G. P. then makes his prediction, and either puts the $1 million in Box 2 or does not. Which will you choose?

I will call E the expected utility argument. E is as follows: The G. P. is a great predictor. If I pick only Box 2 I shall have a very high chance of getting $1 million and a very small chance of getting nothing. If I pick both boxes I shall have a very large chance of getting $1,000 and a very small chance of getting $1,001,000. (Assuming that I do want as large a sum as possible) my best chance is to take only the second box.

I will label D the dominance argument: The G. P. has already put the $1 million in Box 2, or he has not. If he has I will get $1,001,000 if I take both boxes rather than $1,000,000 if I only take Box 2. If he has not then I shall get $1,000 if I take both boxes and nothing if I take Box 2. Therefore in either situation I gain by taking both boxes.

I don't feel it should be held against an E (or any particular E-letter writer) that he assumed linear values for the desire for money, because this is simply not the real point at issue. In reality there may be some people who would choose no box; some who would choose only A because whereas $1,000 would add pleasure, $1,000,000 might add painful complications; and some who didn't care who would choose B (while of D argumentation) because they have a naturally cussed reaction. My guess is that of most people these would factually be a

small minority. But whether they are or not, they only add quite unnecessary complications to an already complicated story. Let us now exclude them.

The question at issue is not merely which argument is correct, (if either) though clearly they cannot both be. It is most of all to point the fallacy in the incorrect version, and not cease the argument after once again displaying either the correctness of E or D.

In Gardner's March 1974 article he tells us that out of the first 148 letters to Scientific American 89 were E's and 37 were D's. I shall reveal myself at once: I am outnumbered; though I shall later discuss circumstances that might make me act like an E -- though never argue like one! Having put my head on the block I will now try and disintegrate the axe.

First: let us not examine the explicit and implicit assumptions, but instead imagine the choice actually taking place. Let us suppose initially that I, and I alone, am offerred it. If I make my D-choice I find either one of 2 things. If the G. P. has slipped up, I am the rich possessor of $1,001,000 which I hope is non-taxable. Does it prove anything? It cannot prove my logic is good, because it might be that one in a million (or whatever the number is supposed to be) error. It does not show that the G. P. isn't a great P, for the same reason. Similarly if I acquire only the expected $1,000 it doesn't prove my logic is wrong. If, on the contrary, I had made an E-choice neither $1 million can prove that the

E argument was correct nor can a loss show it wrong -- unless the G. P. is supposed to be perfect.

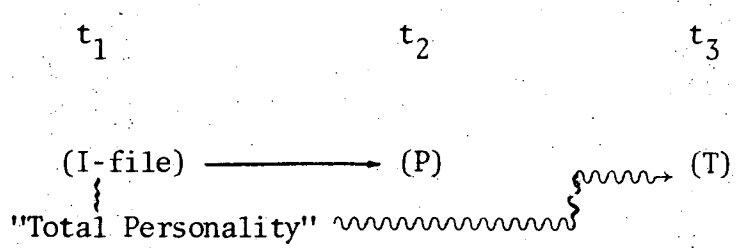
Since this supposition produces no enlightenment, let us imagine that ten thousand people are offerred the choice and are to make it simultaneously (or nearly so) -- a difficult, but surely possible, arrangement. Let us again make reasonable assumptions about the G. P.'s ability to exclude from his population persons who don't want money or who would rather have a thousand than a million, etc. Now let us suppose the day of choice is here and that three quarters (7,500) make an E choice and take Box 2 only, whereas one-quarter (2,500) make a D choice and take both boxes. Of the E choosers let us suppose all but eight have been correctly predicted by G. P. so 7,492 E persons rejoice over their millions and eight E persons have nothing. Of the D choosers let us suppose all but three have been correctly predicted. So 2,497 D persons receive their expected $1,000 and three D persons are the richest of all the ten thousand with $1,001,000. Possibly These three are almost as unhappy as the eight who have nothing. They have been insulted by the G. P.

What would these results show? I have been careful to attribute "success" to about one tenth of a percent of the D's and "failure" to about the same percentage of E's. Does this prove that the E's are right and the D's are wrong? Let us emphasize in addition that we can say also that the E-choosers have obtained on the average $998,933; and the D-choosers have only acquired $2,200 on the average. Surely this proves the superiority of E-arguing? They certainly appear

to have got plenty of utility out of their argument! But this argument, which people of other nationalities might typify pejoratively as American, is only that riches are the reward of industry, or virtue of some sort -- in this case of logical virtue.[28] It is the G. P.'s estimate of the irrationality of their characters to which one might most readily ascribe their reward, and is not due to their actions (subject to provisos yet to be made explicit).

To prove this let us make a careful sequential analysis. Our earliest time marker I will call t_1. The G. P. can be assigned as much time as we like before t_1 to collect data upon behavior, actions, genetic make-up, environments, etc., of all the ten thousand so that at time t_1 he has an enormous, but finite, I-file (I for information) for every one of the ten thousand. This he feeds into a super-program of a super-computer (which can, of course, be his own brain) and then he makes his predictions and places the results, in the form of money, in the boxes at a time t_2 *which must be later than* t_1. At some *still later time* t_3 the ten thousand choose, an event we can call T (for taking?). It is important to recognize that this sequence of times t_1, t_2 and t_3 with $t_1 < t_2 < t_3$, must exist, and that the time that the ten thousand are *informed* of what is to occur is only restricted to be earlier than t_3, but can in no way alter the inequalities $t_1 < t_2 < t_3$. If a world view with a time direction be accepted then in no sense can any event at t_3 cause any event at t_2. The I-file at t_1 can be considered to have direct causal link with event P (placing of money predictively) at t_2. Presumably the I-files at t_1 bear some sort of correlation to

the "total personalities" of the persons of the ten thousand at time t_1. Presumably also the "total personalities" of the persons of the t_1 will bear some relation to their total personalities at t_3 and thus have a link with T. We can illustrate this diagramatically



Examining the boxes at t_3, and later, serves as evidence for the existence of the postulated wavy lines. For subsequent experiments we can now state experimental probabilities (see Section V) for the predictions of the G. P. E-choosers mistake this type of probabilistic forward link between P and E as a backward semi-causal link. I say mistake because I do not think the world view that the last three words of the last sentence imply would be acceptable to any but a few of them.

I will repeat this another way. ("Anything I say three times must be true.") At time t_2 the G. P. separates the ten thousand into sheep and goats; that is into 7,495 E-types and 2,505 D-types. At t_3, three E-types make a D-choice, eight D-types make an E-choice. These 11 persons have "fooled" the G.P. Have they gained or lost by their actions? The three D-choosers have gained $1,000 since they receive

$1,001,000 instead of $1,000,000 (even if they feel insulted). The eight E-choosers have lost $1,000 since they get nothing instead of $1,000. What about the 9,989 who at t_3 act compatibly with their typing at t_2? The E-choosers loose $1,000 since they acquire $1,000,000 instead of $1,001,000. The D-choosers gain because they acquire $1,000 instead of nothing. Thus a D choice must always be preferred.

There are a number of dubious implicit, as well as explicit, assumptions involved, however, in setting up the paradox. It is necessary to examine these in order to blow away any last miasma of confusion, otherwise this argument will remain inconclusive to many readers.

The first two such assumptions that I shall deal with I will call A_4 and A_5. They are related to A_3 (see the third paragraph). A_4 is

> "You have absolute and complete trust that the G. P. will do as he says, and will actually put the money in the boxes at t_2." (A_4)

A_5 is

> "The G. P. does put the money in the box at t_2, and also does not interfere with it at any subsequent time." (A_5)

it is not possible to make any sensible or sensible-seeming argument

whether it be for D or for E, or *at all*, without A_4. It is not possible to make my preceding analysis without A_5. In this respect A_3 is misleading. You cannot know A_5 (see Section VI) at the time you make your argument. An outside observer who is not part of the ten thousand may know at t_2 that the G. P. does put the money in the boxes and by continual watching can assure participants that A_5 was true up to the time he makes the assurance. Thus with a team of trusty, but poker faced, observers we *might* make A_4 reasonable to the participants (especially if we have facilities to lock up the G. P. after he has deposited his money, and have some trusted scientists examine the boxes prior to the deposit). But let us instead momentarily suppose that I, as a participant, have previously observed only from the participating side, a similar sequence with results that were perfect, or more nearly perfect, than those I analyzed. Might I not say to myself "This Being is so superior to me in intellect and knowledge, why should I not assume that he has an enormously superior technology? If he has, then at t_2 he could fill both boxes for every person. At t_3 a destruct mechanism comes into play if you reach for both boxes, so that by the time you open them the $1,000,000 is destroyed." If you find this easier to believe than his fantastic predictive abilities you will, rationally now, make an E-choice!

Suppose instead that you had previously watched *not* just from the participating side but instead had been allowed to observe everything. Does it then follow that you accept A_4 since the stronger

A_5 appears true about the past? Not necessarily, because even if A_5 was true in the past, you can't "know" that it will hold true in the future. Additionally you might say "This Being is so superior to me in his understanding and manipulation of psychology, that it is possible that he brain-washed me into believing I saw him at t_2 put in exactly what was taken out at t_3. Further, any Being who both is so superior, and appears to desire by his system of rewards to encourage irrationality in humanity is not one I can trust!" This might even apply to the parenthesis about making A_4 reasonable.

This does not exhaust the possible reaction induced by more and more observations of perfect results, especially if, perhaps for other overwhelming reasons, you *do* trust the G. P. One person might go mad; another might re-investigate, and experiment on concepts of the nature of time. Those who believe in an all seeing, all knowing, omnipotent God, (the truly Great "G.P."), or who like to study theology will recognize the (in my opinion) insoluble problems we could now approach: problems of Free-will and Determinism, problems like those produced by Plato's Ideals (which both must and cannot contain themselves -- like Russell's Paradox), problems of Good and Evil. As long as perfectionist beliefs are practically applied only to problems of ethical behavior their paradoxes remain, as it were, in obeyance. But almost all words that imply perfection for example "know," "certain" even "true" and "false" can be used in a way that subtly distorts our view of the world we live in -- which is so full of uncertainties, lack of knowledge, broken promises, unfulfilled contracts, and elusive

probabilities.

Let us return to Newcomb's Paradox. I want to discuss another dubious and potentially misleading assumption which I have already touched on indirectly. This is the assumption A_2 which apparently allows you to state a probability for the G.P.'s prediction, because the G.P. has previously predicted the choices of many people in the *same* situation that you are now in. But some group of people must have faced the choice with only A_1 to guide them. The "experimental probabilities" derived from this experiment will now provide an A_2 for the next group. As we discussed at length in Section V, this means that the second group is not in the same situation. Experimental probabilities from the second experiment can be added to readjust the A_2 for a third group. The situation is then again different. It is not clear that we can assume a convergence sufficiently adequate that any group after a certain point in time does, essentially, fact the same situation as the last one. *Only this would allow us the assumption* A_3. We can assume it does, if we wish; but certainly our actual decision as to the present estimated accuracy of the G. P. and the type of mechanism he was employing (for example collection of data and prediction, or some sort of trickery or something unknown) would depend on the analysis of the whole previous sequence of results.

I hope that it is now apparent both why the story of the fatal gene seems like this, and yet much simpler. Mr. Smith makes a decision on probability arguments to save himself from death at a time when he is either already doomed or already safe (from this particular cause of death

only, of course). Similarly in Newcomb's Paradox our choice is made when it cannot (subject to provisos which I have discussed) affect what is in the boxes. It is more confusing simply because the assumptions tend to push the mind more strongly to ascribe to the predictions a self-fulfilling power. A known prediction, or the knowledge that a prediction has been made can certainly affect human behavior subsequent to it. But that subsequent behavior cannot, in any present scientific world view, affect the precedent prediction.

REFERENCES

* Work supported by the U. S. Atomic Energy Commission.

1. "Meet the disease at its first-stage."

2. Harriet Beecher Stowe, Uncle Tom's Cabin, Chapter 20.

3. Martin Gardner, Sci. Amer. 229, No. 1, 104 (July 1973).

4. Martin Gardner, Sci. Amer. 230, No. 3, 102 (March 1974).

5. Robert Nozick, "Newcomb's Problem and the Two Principles of Choice," continued in Essays in Honor of Carl G. Hempel, ed. Nicholas Rescher, Pub. D. Reidel (1970).

6. Gottlob Frege, Grundgesetze der Arithmetik (Breslau 1894).

7. Gottlob Frege, Grundgesetze der Arithmetik begriftsschriftheh abgeleit, Zweiter Band, Jena 1903.

8. Translations from the Philosophical Writings of Gottlob Frege, ed. Peter Geach and Max Black (Basil Blackwell, Oxford, 1952).

9. Kurt Gödel, "Uber formal unentscheidbare Sätze der Principia Mathematica und ver wandter System I." ("On formally Undecided Propositions of Principia Mathematica and Related Systems," Mh. f. Math. 38, 173-198 (1931).[23]

10. Ernest Nagel and James R. Newman, Sci. Amer. 194, No. 6, 71 (1956).

11. Alfred Tarski, "Der Wahrheilsbegriff in dien Sprachen der Deduktwen Disziplinen." ("The Concept of Truth in Formalized Languages") I. Anz. Akad. Wiss., Wier Nr. 2, 23-25 (1932).

12. W. W. Bartley III, Sci. Amer. 227, No. 1, 38 (July 1972).

13. Charles, Baron Bowen (1835-1894).

14. T. Hobbes, English Works of Thomas Hobbes, ed. W. Molesworth (Bohn: London, 1839-1840).

15. G. W. Leibniz, "Reflections on Knowledge, Truth and Ideas" (1684) contained in The Monadology etc., tr. and ed. A. Latter (O.V.P.: Oxford, 1898).

16. Anthony Flew, Western Philosophy (The Bobbs-Merrill Company, Inc., Indianapolis, N.Y., 1971).

17. Wesley S. Salmon, Sci. Amer. 228, No. 5, 75 (May 1973).

18. W. V. Quine, Sci. Amer., 206, No. 4, 84 (April 1962).

19. See page 40 of reference 10.

20. Note the assumptions hidden in this statement!

21. C. E. M. Hansel in ESP: A Scientific Evaluation makes some interesting comments on the mostly likely conclusions to be drawn when the results of an experiment that are by analysis unlikely to be due to chance (Scribner, New York, 1966).

22. Psychology again! To most of us a one millionth chance is something that can't happen to us. Many would lay an even bet in a bar containing twenty-five people, that there will be two with the same birthday -- and expect to make money after visits to only a few bars. (The odds are about 54:43 there will be two such people).

23. Alfred North Whitehead and Bertrand Russell, Principia Mathematica, (orig. 1910), (Cambridge University Press, 1925).

24. G. Ryle, for example The Concept of Mind. (Hutchinson; Random, 1949).

25. A. J. Ayer, for example, Language, Truth and Logic (Second Edit. Gollancz: London, 1946).

26. Rudyard Kipling, "Wee Willie Winkie."

27. This can be put more precisely, that is in a less resoluble form. See reference 18.

28. John Webster, "The White Devil."

LEGAL NOTICE

This report was prepared as an account of work sponsored by the United States Government. Neither the United States nor the United States Atomic Energy Commission, nor any of their employees, nor any of their contractors, subcontractors, or their employees, makes any warranty, express or implied, or assumes any legal liability or responsibility for the accuracy, completeness or usefulness of any information, apparatus, product or process disclosed, or represents that its use would not infringe privately owned rights.

TECHNICAL INFORMATION DIVISION
LAWRENCE BERKELEY LABORATORY
UNIVERSITY OF CALIFORNIA
BERKELEY, CALIFORNIA 94720