Noesis

The Journal of the Mega Society
Number 130
April 1997

EDITORIAL
Chris Cole
P O Box 10119
Newport Beach, CA 92658

Well, we made it to 130 issues!

CONTENTS

ROBERT NOZICK

NEWCOMB'S PROBLEM AND TWO PRINCIPLES OF CHOICE*

> Both it and its opposite must involve no mere artificial illusion such as at once vanishes upon detection, but a natural and unavoidable illusion, which even after it has ceased to beguile still continues to delude though not to deceive us, and which though thus capable of being rendered harmless can never be eradicated.
>
> IMMANUEL KANT, *Critique of Pure Reason*, A422, B450

I

Suppose a being in whose power to predict your choices you have enormous confidence. (One might tell a science-fiction story about a being from another planet, with an advanced technology and science, who you know to be friendly, etc.) You know that this being has often correctly predicted your choices in the past (and has never, so far as you know, made an incorrect prediction about your choices), and furthermore you know that this being has often correctly predicted the choices of other people, many of whom are similar to you, in the particular situation to be described below. One might tell a longer story, but all this leads you to believe that almost certainly this being's prediction about your choice in the situation to be discussed will be correct.

There are two boxes, (B1) and (B2). (B1) contains $1000. (B2) contains either $1000000 ($M), or nothing. What the content of (B2) depends upon will be described in a moment.

(B1) {$1000} (B2) {$M or $0}

You have a choice between two actions:

(1) taking what is in both boxes

(2) taking only what is in the second box.

Furthermore, and you know this, the being knows that you know this, and so on:

(I) If the being predicts you will take what is in both boxes, he does not put the $M in the second box.

(II) If the being predicts you will take only what is in the second box, he does put the $M in the second box.[1]

The situation is as follows. First the being makes its prediction. Then it puts the $M in the second box, or does not, depending upon what it has predicted. Then you make your choice. What do you do?

There are two plausible looking and highly intuitive arguments which require different decisions. The problem is to explain why one of them is not legitimately applied to this choice situation. You might reason as follows:

First Argument: If I take what is in both boxes, the being, almost certainly, will have predicted this and will not have put the $M in the second box, and so I will, almost certainly, get only $1000. If I take only what is in the second box, the being, almost certainly, will have predicted this and will have put the $M in the second box, and so I will, almost certainly, get $M. Thus, if I take what is in both boxes, I, almost certainly, will get $1000. If I take only what is in the second box, I, almost certainly, will get $M. Therefore I should take only what is in the second box.

Second Argument: The being has already made his prediction, and has already either put the $M in the second box, or has not. The $M is either already sitting in the second box, or it is not, and which situation obtains is already fixed and determined. If the being has already put the $M in the second box, and I take what is in both boxes I get $M + $1000, whereas if I take only what is in the second box, I get only $M. If the being has not put the $M in the second box, and I take what is in both boxes I get $1000, whereas if I take only what is in the second box, I get no money. Therefore, whether the money is there or not, and which it is already fixed and determined, I get $1000 more by taking what is in both boxes rather than taking only what is in the second box. So I should take what is in both boxes.

Let me say a bit more to emphasize the pull of each of these arguments:

The First: You know that many persons like yourself, philosophy

FREE WILL is not more Γ₀ NONDETERMINACY, but Γ₁ DETERMINACY. It takes a THEOREM DEMON to "BIND" IT (out-compute it via Γ-uncertainty).

MATHEMATICAL GAMES

Free will revisited, with a mind-bending prediction paradox by William Newcomb

FREE WILL involves RECESSION through UNDERLYING PARAMETERS of CHOICE of... OF CHOICE.

by Martin Gardner

A common opinion prevails that the juice has ages ago been pressed out of the free-will controversy, and that no new champion can do more than warm up stale arguments which every one has heard. This is a radical mistake. I know of no subject less worn out, or in which inventive genius has a better chance of breaking open new ground.

—WILLIAM JAMES

One of the perennial problems of philosophy is how to explain (or explain away) the nature of free will. If the concept is explicated within a framework of determinism, the will ceases to be free in any commonly understood sense and it is hard to see how fatalism can be avoided. *Che sarà, sarà.* Why work hard for a better future for yourself or for others if what you do must always be what you do do? And how can you blame anyone for anything if he could not have done otherwise?

On the other hand, attempts to explicate will in a framework of indeterminism seem equally futile. If an action is not caused by the previous states of oneself and the world, it is hard to see how to keep the action from being haphazard. The notion that decisions are made by some kind of randomizer in the mind does not provide much support for what is meant by free will either.

Philosophers have never agreed on how to avoid the horns of this dilemma. Even within a particular school there have been sharp disagreements. William James and John Dewey, America's two leading pragmatists, are a case in point. Although Dewey was a valiant defender of democratic freedoms, his metaphysics regarded human behavior as completely determined by what James called the total "push of the past." Free will for Dewey was as illusory as it is in the psychology of B. F. Skinner. In contrast James was a thoroughgoing indeterminist. He believed that minds had the power to inject genuine novelty into history, that not even God himself could know the future except partially. "*That,*" he wrote, "is what gives the palpitating reality to our moral life and makes it tingle... with so strange and elaborate an excitement."

A third approach, pursued in depth by Immanuel Kant, accepts both sides of the controversy as being equally true but incommensurable ways of viewing human behavior. For Kant the situation is something like that pictured in one of Piet Hein's "grooks":

> A bit beyond perception's reach
> I sometimes believe I see
> That Life is two locked boxes, each
> Containing the other's key.

Free will is neither fate nor chance. In some unfathomable way it partakes of both. Each is the key to the other. It is not a contradictory concept, like a square triangle, but a paradox that our experience forces on us and whose resolution transcends human thought. That was how Niels Bohr saw it. He found the situation similar to his "principle of complementarity" in quantum mechanics. It is a viewpoint that Einstein, a Spinozist, found distasteful, but many other physicists, J. Robert Oppenheimer for one, found Bohr's viewpoint enormously attractive.

What has free will to do with mathematical games? The answer is that in recent decades philosophers of science have been wrestling with a variety of queer "prediction paradoxes" related to the problem of will. Some of them are best regarded as a game situation. One draws a payoff matrix and tries to determine a player's best strategy, only to find oneself trapped in a maze of bewildering ambiguities about time and causality.

A marvelous example of such a paradox came to light in 1970 in a paper, "Newcomb's Problem and Two Principles of Choice," by Robert Nozick, a philosopher at Harvard University. The paradox is so profound, so amusing, so mind-bending, with thinkers so evenly divided into two warring camps, that it bids fair to produce a literature vaster than that dealing with the prediction paradox of the unexpected hanging. (See this department for March, 1963, or the reprinted version of that piece in *The Unexpected Hanging and Other Mathematical Diversions*, Simon and Schuster, 1969.)

Newcomb's paradox is named after its originator, William A. Newcomb, a theoretical physicist at the University of California's Lawrence Livermore Laboratory. (His great-grandfather was the brother of Simon Newcomb, the astronomer.) Newcomb thought of the problem in 1960 while meditating on a famous paradox of game theory called the prisoner's dilemma [see "Escape from Paradox," by Anatol Rapoport; SCIENTIFIC AMERICAN, July, 1967]. A few years later Newcomb's problem reached Nozick by way of their mutual friend Martin David Kruskal, a Princeton University mathematician. "It is not clear that I am entitled to present this paper," Nozick writes. "It is a beautiful problem. I wish it were mine." Although Nozick could not resolve it, he decided to write it up anyway. His paper appears in *Essays in Honor of Carl G. Hempel*, edited by Nicholas Rescher [and published by D. Reidel in 1970. "What follows is largely a paraphrase of Nozick's

		MOVE 1 (PREDICTS YOU TAKE ONLY BOX 2)	MOVE 2 (PREDICTS YOU TAKE BOTH BOXES)
YOU	MOVE 1 (TAKE ONLY BOX 2)	$1,000,000	$0
	MOVE 2 (TAKE BOTH BOXES)	$1,001,000	$1,000

MATHEMATICAL GAMES

Reflections on Newcomb's problem: a prediction and free-will dilemma

by Martin Gardner

This department's topic for July, 1973, Newcomb's paradox, produced an enormous outpouring of letters. Robert Nozick, who first wrote about the paradox in a paper published in 1970, agreed to look over the correspondence and put down his reactions. Nozick is a philosopher at Harvard University and the author of *Anarchy, State and Utopia*, a book that will be published this summer by Basic Books. William A. Newcomb, the man who discovered the paradox, is a theoretical physicist at the Lawrence Livermore Laboratory of the University of California.

What follows is the communication I received from Nozick in October. May I urge readers who wish to write again not to do so until they have read Nozick's original paper? It goes into considerably more technical detail than my first article or Nozick's present comments.

Newcomb's problem involves a Being who has the ability to predict the choices you will make. You have enormous confidence in the Being's predictive ability. He has already correctly predicted your choices in many other situations and the choices of many other people in the situation to be described. We may imagine that the Being is a graduate student from another planet, checking a theory of terrestrial psychology, who first takes measurements of the state of our brains before making his predictions. (Or we may imagine that the Being is God.) There are two boxes. Box 1 contains $1,000. Box 2 contains either $1 million or no money.

You have a choice between two actions: taking what is in both boxes or taking only what is in the second box. If the Being predicts you will take what is in both boxes, he does not put the $1 million in the second box. If he predicts you will take only what is in the second box, he puts the million in the second box. (If he predicts you will base your choice on some random external event, he does not put the money in the second box.) You know these facts, he knows you know them and so on. The Being makes his prediction of your choice, puts the $1 million in the second box or not, and then you choose. What do you do?

There are plausible arguments for reaching two different decisions:

1. The expected-utility argument. If you take what is in both boxes, the Being almost certainly will have predicted this and will not have put the $1 million in the second box. Almost certainly you will get only $1,000. If you take only what is in the second box, the Being almost certainly will have predicted this and put the money there. Almost certainly you will get $1 million. Therefore (on plausible assumptions about the utility of the money for you) you should take only what is in the second box [*see illustration on opposite page*].

2. The dominance argument. The Being has already made his prediction and has either put the $1 million in the second box or has not. The money is either sitting in the second box or it is not. The situation, whichever it is, is fixed and determined. If the Being put the million in the second box, you will get $1,001,000 if you take both boxes and only $1 million if you take only the second box. If the Being did not put the money in the second box, you will get $1,000 if you take both boxes and no money if you take only the second box. In either case you will do better by $1,000 if you take what is in both boxes rather than only what is in the second box [*see illustration on page 104*].

Each argument is powerful. The problem is to explain why one is defective. Of the first 148 letters to *Scientific American* from readers who tried to resolve the paradox, a large majority accepted the problem as being meaningful and favored one of the two alternatives. Eighty-nine believed one should take only what is in the second box, 37 believed one should take what is in both boxes—a proportion of about 2.5 to one. Five people recommended cheating in one way or another, 13 believed the problem's conditions to be impossible or inconsistent and four maintained that the predictor cannot exist because the assumption that he does leads to a logical contradiction.

Those who favored taking only the second box tried in various ways to undercut the force of the dominance argument. Many pointed out that if you thought of that argument and were convinced by it, the predictor would (almost certainly) have predicted it and you would end up with only $1,000. They interpreted the dominance argument as an attempt to outwit the predictor. This position makes things too simple. The proponent of the dominance argument does believe he will end up with only $1,000, yet nevertheless he thinks it is best to take both boxes. Several proponents of the dominance principle bemoaned the fact that rational individuals would do worse than irrational ones, but that did not sway them.

Stephen E. Weiss of Morgantown, W.Va., tried to reconcile the two views. He suggested that following the expected-utility argument maximizes expectation, whereas following the dominance argument maximizes correct decision. Unfortunately that leaves unexplained why the correct decision is not the one that maximizes expectation.

The assumptions underlying the dominance argument, that the $1 million is already in the second box or it is not and that the situation is fixed and determined, were questioned by Mohan S. Kaleikar, a physicist at the Nevis Laboratories of Columbia University, who wrote: "Perhaps it is false to say that the Being has definitely made one choice or the other, just as it is false to say that the electron [in the two-slit experiment] went through one slit or the other. Perhaps we can only say that there is some amplitude that B2 [second box] has $1 million and some other amplitude that it is empty. These amplitudes interfere unless and until we make our move and open up the box. . . . To assert that 'either B2 contains $1 million or else it is empty' is an intuitive argument for which there is no evidence unless we open the box. Admittedly the intuitive evidence is strong, but as in the case of the double-slit electron diffraction our intuition can sometimes prove to be wrong."

Kaleikar's argument makes a version of the problem, in which the second box is transparent on the other side and someone has been staring into it for a week before we make our choice, a significantly different decision problem. It seems not to be. Erwin Schrödinger, in a

SOME Q & A ON THE RESOLUTION OF NEWCOMB'S PARADOX

The following questions were suggested by the content of a telephone conversation which occurred on Sunday, February 9, 1997 between Rick Rosner and me. By the time this conversation was over, Rick (who, as I recall, found my first paper on the resolution "pretty unreadable") seemed to have a much better grasp of the resolution than before. I hope this will prove true for other readers who were stymied by the mathematical style of the original paper. To better orient yourself regarding this dialogue, please read Nozick's original definition of Newcomb's problem in this issue.

Newcomb's problem: There are two boxes on a table. Box A is transparent and contains $1,000. Box B is opaque. You are allowed to choose either A and B, or B alone. However, you have been told told by a reputedly omniscient being that he has put $1,000,000 in box B if and only if he has predicted that you will take B alone. If, on the other hand, he has predicted that you will take both A and B, or that you will make your choice on the basis of some random event, then B is empty. On the basis of your own past experience, you have full confidence in this being's predictive abilities. You know that he has correctly predicted the outcome of this game on many previous occasions with many other players. Furthermore, he has correctly predicted your own behavior in many other situations of various kinds. What should you do?

QUESTION: What is the difference between Newcomb's problem and Newcomb's paradox?

ANSWER: Newcomb's *problem* is to specifiy which of two distinct alternatives should be chosen to maximize gain in a certain decision-theoretic context. Unfortunately, since each choice has an apparently sound justification, there are two equally valid solutions. Newcomb's *paradox* arises from the mutual contradiction of these solutions. The object is to show why one justification, and therefore one solution, is actually better than the other, and thus to simultaneously resolve the paradox and solve the problem.

One choice, taking both boxes, seems justified because if the money is already in the boxes, then one can lose nothing (and may gain the extra $1,000 in box A) by taking both. This argument is called the **dominance argument,** and it relies on the idea that time is confined to a one-way, cause-to-effect linear sequence from past to future. The other choice, taking only one box, seems justified because there is a virtually unlimited amount of inductive empirical evidence to the effect that a million-dollar gain will result. This argument is called the **expected utility argument,** and it relies on our ability to make inferences from observation...i.e., to infer the future from the past.

Thus, Newcomb's paradox pits the idea that time is linear and one-way - that the past affects the future, but never vice versa - against the idea that we can infer the future

from the past, which in this case requires the existence of multiple levels of time. It therefore centers on the nature of time and its interface with human cognition.

QUESTION: The nature of time? But isn't the crux of the paradox often perceived as the existence or nonexistence of free will?

ANSWER: Yes, but this perception is erroneous. The paradox can be resolved, and the choice can be made, irrespective of free will. That is, the model which resolves the paradox - the NST - logically supports the existence *or* nonexistence of free will. What really counts is the *relationship* between free will and time, as developed in the CTMU.

QUESTION: What is the NST, and how does it resolve Newcomb's paradox?

ANSWER: The NST, or Nested Simulation Tableau, can for present purposes be described as a computational hierarchy of virtual realities, each one nested within a higher reality surrounding it. The NST resolves the paradox by building an inclusive computative reality around the physical reality we take for granted. In this extended universe, time is directionally unrestricted, but can still be treated as a one-way linear dimension within the physical level. Thus, it is a *model* in which expected utility and dominance may coexist in the Newcomb context.

An easy way to envision the NST is to imagine that physical reality is a "program" running on a vast, ultrahigh-resolution 3-dimensional "monitor" consisting of one or more sub-monitors corresponding to physical cognitive agents. To whatever extent the sub-monitors interact on a physical level, their contents intersect at a mathematical interface. Thus, submonitors represent subjective frames of reference, whereas the interface represents objective Minkowski spacetime. The setup allows arbitrary localized access to the physical level, letting the higher-level programmer-controller avail himself of distributed and nondistributed programming at his convenience.

The NST model contains at least two possible mechanisms for creating the Newcomb scenario, each one supporting a distinct kind of "omniscience". In one, the programmer simply controls the thought and behavior of the physical subject, effecting omniscience by means of omnipotence. In the other, the programmer allows the physical subject to make all his own decisions, but "random-accesses" certain key physical-time junctures in order to control the effects of those decisions once they have been made. In the latter case, the predicted subject has free will; in the former, he does not.

QUESTION: How should the evidence posited in Newcomb's problem be interpreted?

ANSWER: When reading the above definition of the problem, the thing to notice is that the evidence for omniscience is hard in quality and unlimited in quantity. Regarding quality, put yourself in the subject's position and consider the meaning of the word

know. Since you "know" that ND (the programmer's simulated self-image) has made many true predictions, you must be an eyewitness (as opposed to someone who got his information at second hand and may have been been misled). And since in many cases it was your own behavior that was predicted, you are in a position to rule out trickery.

Regarding quantity, consider Nozick's use of the word **many.** "Many" could mean, say, twenty; it could also mean twenty thousand. Because this information is not explicit, we must address the most extreme case: an unbroken string of as many successful predictions as can be observed in an entire human lifetime. That is, we must address the case in which there is so much evidence for omniscience that it cannot be rationally discounted. Thus, the definition of Newcomb's problem brings us so close to hypothetical certainty regarding ND's omniscience that no "wriggle-room" remains. The probability that you have not been tricked, lied to, or deceived by an improbable random sequence is at least as high as the probability that time is linear, especially given your inability to prove the latter assumption.

It follows that if you cannot prove rigorously that Newcomb's problem constitutes a logical absurdity - if you cannot, in apparent violation of modern physics, prove that time is strictly unidirectional within a reality which has only one level - then only one possible conclusion remains: you occupy a "virtual reality" whose programmer is virtually omniscient with respect to your thought and behavior.

QUESTION: In a way, virtual reality seems like an obvious solution...too obvious to have been missed by all the professional philosophers who wrote papers on Newcomb's paradox. What could possibly account for everyone having missed it?

ANSWER: In academia, reputation is at least as highly valued as innovation. If a radical conceptual innovation has no immediate payoff in patents or grant money, its cost to the innovator can be great. Academic standards of conformity tend to preclude flights of imagination, especially the kind that might attract the derision of one's peers. Though philosophy is the mother of all disciplines, the success of modern science has left most philosophers with some amount of "science envy"...i.e., inflated expectations from scientific methodology and aspirations to a scientific mindset, including an inveterate distrust of anything departing too sharply from scientifically acceptable ideology. Academic philosophy has thus sacrificed its duty as the most general and fundamental of disciplines - an open mind - to the coin of the academic realm, grants and tenure. Nowhere is this clearer than the history of Newcomb's paradox.

QUESTION: Many conventional analyses of Newcomb's paradox treat it with a mixture of game theory and subjective probability - i.e., degree of belief or confirmation. Why is that not sufficient?

ANSWER: The evidence in this problem - an arbitrarily long, unbroken string of

predictive successes - is hypothetically factual and logically implies omniscience. This evidence does not arise from opinion, but from some aspect of factual reality. Therefore, the question of its validity, and that of the problem itself, comes down to whether or not there exists a model of reality that incorporates a mechanism for omniscience. But again, opinion has nothing whatsoever to do with this question.

Because people are often irrational and subjectively motivated, practical game theory must allow for these features of human psychology. Professional philosophers are sensitive to this need and would like to construct a bridge between the rational and irrational sides of human nature. Being without an example with which to work, they fixated upon the supposed "irrational component" of Newcomb's problem. But if the Newcomb scenario *were* irrational - if omniscience were logically inconceivable - then its evidence would have to be dismissed, and the irrational component would vanish. Indeed, if the problem turns out to have a logically identifiable irrational component, then it will be revealed as an irrational problem which does not admit of rational analysis. But until then, it must be interpreted and solved on a purely rational basis, and Newcomb's wager must be handled accordingly by any rational subject.

Unfortunately for professional philosophers who want to use rationality to bridge the chasm between rationality and irrationality, this is a general situation. As soon as their rational bridge touches down on the irrational side of the gap, it is doomed by its very nature to collapse. It is the chasm itself that is irrational, and the best we can do is construct our bridges over and around the irrational features of our psychological terrain. The *real* paradox is that so few of us, even after decades of pointless academic wrangling over Newcomb's paradox, seem to have come to this realization.

QUESTION: The philosophical community would seem to be right about one thing at least: the idea that physical reality is "virtual" is pretty wild. VR technology is still in its infancy, and there is an obvious problem of scale. Would simulating an entire universe not seem to be prohibitively difficult and costly in a number of ways?

ANSWER: Logically speaking, the opinion that VR is "pretty wild" is irrelevant. It depends on your subjective expectations. Questions of timing are also irrelevant, given a VR programmer's technological independence from the world he is simulating and his ability to create virtual memories and perceptions in the minds of simulated beings (including their own technological inferiority). And as far as scale is concerned, the programmer need not explicitly simulate a whole universe. He need only simulate the information directly perceived, recognized, inferred or imagined within the minds of his software homunculi.

In fact, he need only attend to the cognition of one homunculus at a time! Once you realize that you may inhabit a simulation, you also realize that your reality may be a solipsistic one in which just you, and only secondarily the objects of your perception,

"exist". Solipsism may be anathema in philosophical and scientific circles, but unless you can personally construct an airtight logical proof against it, you must admit it as a logical possibility subject to confirmation. Philosophy is the one area where plausibility alone is *never* a sufficient argument. Only logical implication goes the distance.

QUESTION: Are you sure that nobody thought of the virtual reality scenario first with regard to Newcomb's problem?

ANSWER: I first encountered Newcomb's paradox in the early-to-mid-1980's, and as I recall, thought up the NST (Nested Simulation Tableau) shortly thereafter. For all I know, somebody else may have considered it even before that. But if so, it is still far from certain that he developed it to the required extent or was able to explain in logical terms why it constituted a resolution of the paradox. For example, did he describe an actual mechanism for omniscience, define the NST in logical terms, or recognize it as an extended hierarchy analogous to the theory of metalanguages and thus as a model for all relevant arguments including dominance and expected utility? Probably not. Did he pursue its logical implications, as I did with the CTMU? Again, probably not. Thus, even with credentials to ease his way, he may not have been able to convince the editor of any academic journal that his idea was publishable. And even if such a suggestion were published, then judging from the speed and thoroughness with which it seems to have been buried and forgotten, it was never properly justified or defended.

Fortunately, I'm not an academic, and *Noesis* is not an academic journal.

QUESTION: VR has been big in Hollywood lately. Isn't the NST suspiciously trendy?

ANSWER: Using Hollywood-style VR in a movie plot - "Johnny Mnemonic slaps on his 3-d data goggles, dons a pair of electronic gloves, and mixes it up with the bad guys at high noon in Cyberspace" - is not quite the same as applying it mathematically to resolve an intractable philosophical paradox. Furthermore, when the paper was first written nearly a decade ago, VR wasn't quite as fashionable. Had the whole thing been handled properly, the Mega Society could have used the resolution to "ride the wave" and generate some publicity for itself. Much to my chagrin, this opportunity was squelched. But I did my part by delivering the goods, and it wasn't me who decided to diddle around until VR became "suspiciously trendy".

QUESTION: Haven't you been accused of "making metaphysical assumptions" in your resolution of Newcomb's paradox?

ANSWER: Such an accusation has been made, but it is baseless. The NST is amenable to a number of possible constructions. The one used to resolve Newcomb's problem - a computer within a computer... running a simulation within a simulation... - is closely analogous to the *theory of metalanguages*, already an essential ingredient of

logic. Experimenting with various possible interpretations of this ingredient is not the same thing as making factual assumptions.

After all, physics already has a bidirectional conception of time. Granted, time usually behaves in what seems to be a directional way, and scientists have had a great deal of success in treating time as a one-way linear dimension. In the branch of physics called thermodynamics, the one-way linearity of time is known as *entropy*. On the other hand, particle physics treats time as a two-way symmetry, tachyonic time reversal emerges from relativity theory, and in quantum theory, measurements can have implications that may affect the past. Times have changed in physics, and so has time's directionality.

Looking at the matter from a cybernetic viewpoint, physics has already prescinded from one-way time and introduced a degree of time-directional *freedom*. The dominance argument then reduces this freedom by introducing a *constraint* to the effect that time is exclusively one-way...a constraint that introduces *information* with respect to an extant pair of directional possibilities. Since only the constraint embodies new information, only the constraint qualifies as an "assumption". In other words, by prescinding to the NST and thereby suspending the one-way constraint, I avoided the risk of introducing spurious information to the Newcomb context and thus did the diametric *opposite* of "making an assumption". And in the process - hard though it may be to grasp - I remained truer to the spirit of modern science than did my critic(s).

A good case can be made that this accusation, coming from the source from which it came - the publisher of *Noesis*, the journal in which it appeared - has robbed the solution of the attention it deserved for the last seven years. In any case, the fact that it was made so carelessly, and has been maintained so obstinately, represents a singular injustice not only to me, but to the Mega Society and the readers of this journal.

QUESTION: Didn't the famous recreational mathematician and *Mathematical Games* columnist Martin Gardner believe that Newcomb's paradox was irresolvable, and didn't he use another kind of paradox to show why this might be so?

ANSWER: Yes. The paradox that Gardner used in support of his opinion was as follows. "A supercomputer is asked to predict if a certain event will occur in the next three minutes. If the prediction is no, it turns on a green light. If yes, it turns on a red light. The computer is now asked to predict whether the green light will go on. By making the event part of the prediction, the computer is rendered logically impotent."

This is one of a class of "prediction paradoxes" illustrating the pitfalls of foretelling the future. However, unlike a physical computer, a higher-level NST programmer cannot be physically compelled to make logically self-contradictory predictions. And even if he were carelessly to make such a prediction, tripping himself up in the process, his ability to successfully make other kinds of prediction would remain intact.

Our own Ronald K. Hoeflin once tried to use an older but similar paradox against the resolution. "Can an omnipotent God create a stone too heavy for even Himself to lift? If He is omnipotent, then He can do anything and the answer is yes. But then He cannot lift the stone and is therefore not omnipotent." Again, this paradox is resolved by NST stratification. God as a Self-simulated physical entity cannot lift the stone, but God the NST Programmer *can*. By letting us distinguish between these two aspects of God, the NST resolves the paradox.

Gardner's use of a computer to illustrate his position on the paradox is not without irony. His book *Gotcha: Paradoxes to Puzzle and Delight* contained not just an account of Newcomb's paradox, but a mere 7 pages away, a brief description of the theory of metalanguages...a theory of which the NST is a close computational analogue. Thus, far from presenting a logical invalidation of Newcomb's paradox, Gardner presented - albeit in two separate pieces - the logical means to resolve it!

QUESTION: What is the theory of metalanguages, why is the NST a metaphysical analogue of it, and what does the analogy imply?

ANSWER: Any language is used in reference to something of an abstract or concrete nature. With respect to this relationship, the language is called an *object language* and the referent is called its *object universe*. We can now define a higher-level language expressing the details of this relationship, a so-called *metalanguage*. The metalanguage can be used to selectively posit, derive and compare various possible rules of object-language syntax and semantics (something which cannot be done in the object-language itself, which begins with and can never deviate from its own syntax). Given such a metalanguage, we can then define a yet-higher metalanguage expressing the relationship of the first metalanguage to *its* universe, i.e., to the relationship of object-language to object-universe, including the rules of object-level syntax and semantics. Inductively extended, this logical relationship constitutes the "theory of metalanguages". This theory is an established ingredient of logic.

It has been realized by a long line of great minds, from ancient Ionian philosophers to scientists from Lavoisier to Freud to Chomsky, that cognition is a form of language. This is true not only in terms of content - rational cognition is generally reducible to a sequence of conventionally linguistic expressions - but in terms of form, as becomes evident when we represent thought as a mathematical sequence of neural state-transitions. If we accordingly replace the terms "object universe", "object-language" and "metalanguage" with "physical universe", "object-level cognition" and "higher-level cognition" respectively, and then add a cybernetic control parameter, the NST is the result. The NST is thus a well-defined mathematical object homomorphic to a necessary and well-studied ingredient of logic, the theory of metalanguages.

The analogy between metalanguages and NST "stratified virtual reality" implies that for

logical purposes, any TOE (Theory of Everything), considered as a language, is required (a) to be its own regressive metalanguage, i.e., to explain how and why its own axioms and theorems apply to reality, how and why the explanation applies to reality, and so on *ad infinitum;* and (b) in placing reality within the framework of logic for analytical purposes, to explicitly incorporate the entire metalinguistic component of logic and a mapping of reality thereto. In other words, no NST, no TOE. Just as in Aristotle's day, cosmology is a branch of metaphysics and requires a logical framework of metaphysical scope. That framework begins with the NST and "ends" with the CTMU, which terminates the most general metaphysical stage of theorization and functions as a new beginning with respect to science.

QUESTION: What about feelings, emotions, and the subjective impression of consciousness? Can you account for the subjective and emotional dimensions of human existence in a computational setting?

ANSWER: I can, but I don't yet have to. Feelings, emotions and the impression of consciousness have not yet been accounted for in <u>any</u> setting. For all you know, your feelings and emotions are already mere computational artifacts of physiological programming. By demanding that a computational mechanism be produced for them, you are presuming that a noncomputational explanation already exists. But as far as you know, it doesn't, and you cannot ask more of the NST than you do of science.

QUESTION: No matter how well the NST can be logically justified as a conceptual tool, VR still seems highly unnatural as an explanation for reality. Does it become more "natural" as its implications are developed?

ANSWER: Yes. The CTMU, which is what the NST ultimately becomes, is a very natural theory well in keeping with the history of world philosophy and science. The NST is merely a starting point for working cognition into reality theory, which turns out to be a necessary step towards theoretical unification. This is unsurprising when we realize that the dependency of physical reality on observation and measurement, both of which are cognitive in a general sense, respectively characterize relativity and quantum theory, the dominant theories of physics on large and small scales.

QUESTION: Applying the metaphysics of virtual reality to a longstanding philosophical conundrum seems like a radical and possibly momentous departure in the field of analytical philosophy. How important is it?

ANSWER: Very important indeed. The field of metaphysics *qua* philosophy has been stagnating for years. Professional philosophers, cowed by the success of the physical sciences, have all but abandoned cosmology to the physicists. Some philosophers still hope to advance the field, but - let's face it - the physicists are walking all over them. Regarding the ultimate nature of reality, a Hawking, Lederman, Penrose or Tipler can

outsell books by factors of many, and if one actually does bother to look at what the philosophers are putting out, disappointment generally follows. Part of the reason: logic and mathematics are our most powerful tools for organizing abstractions, and most philosophers lack the mathematical creativity and discipline required by modern cosmology. Instead, they indulge in physics envy, exegesis, pointless criticism, or fussy games of pick-and-choose from the dusty scrolls of philosophical patriarchs.

What philosophy needs is a new paradigm. The NST and its logical development, the CTMU, are ideas whose times have come.

QUESTION: This whole matter seems to verge on the topic of religion. Can we draw any clear parallels between the two?

ANSWER: Yes. Western religion in particular has tended to separate God from the physical universe in a way requiring the full stratified NST scenario. And oriental religions, by regarding physical reality as an "illusion" concealing a deeper level of being, tacitly embrace the potential for multiple levels of ontological simulation. Once two distinct NST strata are posited, the rest follow by induction. Any attempt to unify science and religion, or the physical and spiritual universe, must therefore construct a mapping between the NST and scientific theory. Given the pressing social, political, and psychological need for such a unification, the importance and conceptual utility of the NST - and its full logical development, the CTMU - are undeniable. That's why the CTMU doubles as a "logical religion" of its own. This was first pointed out in my original paper, *The Resolution of Newcomb's Paradox.*

QUESTION: The VR resolution of Newcomb's paradox seems to be *at least* as cogent and well-founded as the many other "resolutions" that have been published in standard philosophical journals. Why haven't you just sent it to the editor of an accredited academic periodical?

ANSWER: For the following reasons. (1) For years, philosophical journals have been printing so-called "resolutions" and criticisms of the paradox that are not only dead wrong, but completely miss its thrust. The correct resolution might well be rejected because it makes these journals, and the academics who read and write for them, look bad. (2) Although I entered college with a full academic scholarship, I soon fell victim to personal misfortune and bureaucratic pettiness. In the years since, tuition has only gone up while I've remained poor, and the terms of my departure left me all but ineligible for financial aid. The net result: my academic credentials amount to a high school diploma, period. For many "respectable" scholastic journals, that's an instant (and asinine) excuse to dismiss anything I send them out of hand...especially given my evident ability to make their contributors appear to have been asleep. (3) Even if my unsponsored paper were subject to double-blind refereed review, there is *a fortiori* a considerable likelihood that during the months of wrangling that would follow, some

publish-or-perish hack would deduce my weak academic position, paraphrase my paper, and attempt to retrodictively author it himself, using his superior credentials to get it rushed into print under his own name. And even if the paper were miraculously to be published with proper attribution, my lack of credentials would dog me still. To an academic outsider, journal credit is about as negotiable as confetti.

None of these reasons alone is decisive. But taken together, they militate strongly against the idea that someone in my situation, and with my history of unfair treatment by institutionalized higher education, can blindly repose his trust in academia.

But then again, that's what *Noesis* is supposed to be for, isn't it?

SOME Q & A ON THE 10-MARBLES PROBLEM

In view of Publisher Cole's recent resurrection of the controversy over the notorious 10-marbles problem, some further clarification might be in order. Again, thanks to Rick Rosner for suggesting some of the following questions.

The 10-marbles problem: From a box containing exactly 10 marbles, one marble at a time may be randomly extracted, examined and replaced. If you do this exactly 10 times and each time observe a white marble, then on this basis alone, what is the probability that all 10 of the marbles in the box are white?

This problem originally appeared in *Noesis,* in roughly the above form, as an item in one of Ron Hoeflin's "Trial Tests". While it has since been modified, this is the version over which the present controversy exists.

The controversy, mainly between Publisher Cole and myself, is about whether or not the problem is solvable as given. I say it is; Cole says it isn't. Specifically, I say that a solution follows directly from a simple initialization of Bayes' rule with all possible combinations of white and nonwhite marbles. Cole, on the other hand, claims that we need further information on (a) the prior distribution from which the box was filled (including the specific colors it contains) and (b) the rule by which the marbles in the box were chosen from the prior distribution. Cole's name for his position is "Bayesian Regression". (Chris can correct me if he thinks I've misinterpreted his position; in fact, I wouldn't mind if he did, provided that in the process he finally spells out precisely what his position actually is.)

The following experiment has been proposed to test the hypothesis that the probability in question is approximately .67, as computed from a straightforward initialization of Bayes' theorem with every possible proportion of white-to-nonwhite marbles.

1. Compose a large number of statistical rules for constructing 10-element sets of 1-10

colors each. Some of these rules should permit the construction of all-white sets.
2. Feed these rules into a computer programmed to construct sets accordingly and construct equal numbers of sets using each rule. Continue until a significant number of all-white sets have been constructed.
3. Program the computer to randomly sample the elements of these sets at 10 samples per set. Loop this procedure until each set has been sampled numerous times and a large number of all-white runs have been generated.
4. Tabulate every all-white run according to the composition of the corresponding set.

QUESTION: What is the relationship between the data - the run of 10 white marbles in a row - and the prior distribution from which the contents of the box were chosen?

ANSWER: Since the data come from the box only *after* the box has been filled, they tell us not about the prior distribution, but only about what actually made it into the box. The solid walls of the box constitute a logical barrier between the prior distribution and the data. Otherwise, we could simply dispense with the box altogether.

Obviously, the contents of the box may reflect the prior distribution. In this case, if the data accurately reflect the contents of the box, then they reflect the prior distribution as well. But what if the contents of the box are "improbable" relative to the prior distribution? Then the prior distribution is nothing but "disinformation" relative to the contents of the box, and mixing it with perfectly good data is absurd.

For example, suppose that the method of filling the box was chosen deliberately to conceal the nature of the prior distribution. E.g., suppose that the prior distribution consisted of 10 white and 10 million nonwhite marbles of various specific colors, but that the 10 white marbles were deliberately sought out and put in the box. Then virtually all continuity between the prior distribution and subsequent observations has been destroyed, and knowledge of the prior distribution - in which nonwhite marbles were a million times more numerous than white ones -can only interfere with accuracy. Since we cannot assume that the contents of the box reflect the prior distribution, knowledge of the prior distribution cannot be necessary.

We can sum it up like this. In some cases, knowledge of the prior distribution can help; in others, it can hurt. If it were "necessary", it would help all of the time. But it doesn't, and so it isn't.

QUESTION: What about the rule according to which the contents of the box were chosen, including the composition of the prior distribution?

ANSWER: The contents of the box could have been selected by either a statistical or deterministic rule. Suppose that the rule was deterministic; e.g., that someone deliberately put certain numbers of marbles of certain colors in the box. In this case,

knowing the rule amounts to knowing the exact contents of the box. But if we have this information, then no probability need be calculated!

If, before calculating the probability of an event, we can demand to be told whether or not it will occur or has occurred, then there is no reason to calculate any probability whatsoever. We can simply talk, trick, or tickle the information we want out of whomever is "responsible". Unfortunately, it is impossible to pre-assign responsibility for random events. It follows that we do not need to know the rule to calculate the required probability.

QUESTION: Obviously, if we need not know the prior distribution, then we need not know the specific colors it contains. But what about the specific colors in the box? Don't we at least have to know *them* in order to properly initialize Bayes' theorem?

ANSWER: No. The frequency with which a white marble is observed depends only on the fraction of marbles in the box that are white. Since it makes no difference how we refer to the rest, "nonwhite" is sufficient. To verify this, perform the following thought experiment. Imagine that the box contains white and gray marbles only. Now ask yourself whether the frequency of white observations will change if you paint the gray marbles various nonwhite colors. Since the answer is obviously "no", you need no information on any specific color but the one whose measure is to be inferred ("white").

QUESTION: What should we assume as initial probabilities for the 10 possible 10-marble distributions of white and nonwhite marbles in Bayes' theorem?

ANSWER: We don't have to assume anything. We have data on frequency - a 10-trial run - from which to determine their probabilities. Assumption always defers to data.

QUESTION: An experiment was recently proposed to confirm the hypothesis that 67% of all-white 10-trial runs from random 10-marble sets will be found to have come from all-white sets (see description above). Isn't setting up an experiment like this one rather difficult, provided the hypothesis is to be properly evaluated?

ANSWER: To be fair, yes. The most important criterion for setting up such an experiment is that the data - the run of 10 white observations - be allowed to "select" the most probable prior distributions from the universe of all possible prior distributions. That way, prior distributions favoring 10-element sets likely to yield all-white runs will predominate, whereas those which do not will tend to drop out. This criterion may be problematic, but it's how nature works, and anything else is too presumptive. The whole idea is to let the data self-select *without interference from assumptions built into the experiment,* and screening out all such assumptions requires extreme care.

Unfortunately, probability theory in its current state offers little guidance. There are

several important issues here which have not yet been fully explored, and anybody tackling this experiment at the present time is pretty much on his own. In fact, the only thing he knows for certain is that if it fails to confirm the given hypothesis, which is implied by the convergent equivalence of frequency and probability established by the Law of Large Numbers, then it is flawed. That's because the efficacy of the experiment relies as heavily on this equivalence as does the hypothesis itself.

To some extent, calling for an experiment was deceptive of me (okay, I was annoyed!) If the experiment fails to yield the hypothetical result, then it must have been bungled. We know this because the relationship between frequency and proportion would be violated by any other outcome. If an experiment contradicts a mathematical fact, which is what this relationship is, then the experiment must yield to the fact and not vice versa. So if anybody runs an experiment which yields any result but .67, he'll simply have to publish the exact setup so that we can try to determine how he botched the job. Then we can finally confront the *real* problem of "Bayesian Regression"!

QUESTION: Wasn't this controversy supposed to be about Bayes' theorem, as opposed to the Law of Large Numbers?

ANSWER: Like all probabilistic theorems, Bayes' theorem tacitly relies on the relationship between frequency and proportion (or probability). The Law of Large Numbers is just a precise numerical statement of that relationship, and its mathematical derivation a justification. It is the numerical basis of probability theory and statistics.

QUESTION: Aren't there certain problems with the initialization of Bayes' theorem?

ANSWER: Only in the absence of data. Without data to provide qualitative and quantitative information about initial possibilities, one has no idea what possibilities to set equal. But this problem applies to *any* probabilistic rule, not just Bayes'.

QUESTION: Chris Cole may not be a professional probability theorist, but he is obviously a very intelligent person. He must have glimpsed *something* about this problem that made him think it was unsolvable (the primary thesis of his supposed theory of "Bayesian Regression"). What do you think he meant?

ANSWER: What Chris Cole may have meant was this. Suppose we begin with a badly skewed prior distribution containing 2 white marbles and 2,000,000 nonwhite ones of various colors. Say we choose two marbles at random for insertion in a box. Now say we sample marbles randomly from the box and get a "run" of 2 white marbles. What is the probability that both of the marbles in the box are white?

Obviously, the nature of the prior distribution makes it extremely unlikely that both of the marbles in the box are white. If we do not know the prior distribution, we do not

know this, and will compute an unrealistically high probability.

However, there are several things to note about this example. First, it assumes the prior application of a statistical rule involving a highly asymmetric prior distribution (as defined relative to a spatiotemporally-inclusive "proto-distribution" consisting of all possible prior distributions). Since this assumption is anything but general, neither is any criticism based on it. In fact, Bayes' theorem is more likely to be right in the long term if it completely ignores pathological examples of this kind, which are themselves too improbable to serve as reliable probabilistic criteria. In general, the more improbable the assumed prior distribution, the less it is worth as a counterexample.

Second, what if the prior distribution had contained only 1 white marble instead of 2? Then the probability we seek is "really" 0. But as we explained above, demanding to know this information is to demand that a probability be replaced with a fact. If this is allowed, then we need never compute another probability of any kind.

Third, prior to sampling, the box was highly unlikely to contain even one white marble. Once a white marble shows up, however, it can be treated as a "given", and we may compute a *conditional probability* predicated on the datum. This ability - to restrict the context to known conditions and compute a probability *relative to* them - lets any probabilistic calculation be isolated from the external context in which every condition has a prior probability. In other words, even when we lack the information to compute the probability (or measure) of a given datum within the overall context - whatever that may be! - we can still compute a subsequent conditional probability based on that datum. Because Chris' current version of "Bayesian Regression" peremptorily forbids the calculation of such a conditional probability, it is certainly erroneous.

Indeed, if we deny the existence of relative (or conditional) probability, we are either (a) denying the existence of probability, or (b) saying that only "absolute probability" exists. If (a), then we have to dump probability theory in its entirety. If (b), then a "probability" can only be one of two things: a deterministic constraint affecting the elements of a set, in which case it is a fact rather than a probability, or a statistic affecting only a whole set, in which case it amounts to a logical quantifier and precludes any inductive use of probability theory whatsoever. Again, probability theory as we know it bites the dust.

This brings up a very basic distinction between logic and probability, or deterministic and probabilistic reasoning. Probability does not have to be perfect; it only has to be valid in "most cases". Unlike a deterministic constraint, which can be factually invalidated by counterexample, probability is invulnerable to occasional bursts of improbable short-term data. Such deviations are inevitable, and we cannot require probabilistic theorems to forecast every one of them specifically.

QUESTION: So is Bayesian Regression all wrong about probabilistic unsolvability?

ANSWER: Not in a wider probabilistic context. Suppose we have a coin, toss it twice, and come up with two heads in a row. Say that we now wish to compute the probability that both sides of the coin are heads. If we don't know that the coin is fair, then it may be "loaded" so that one side always lands up. How do we separate the possibility that it is loaded from the possibility that it has two heads?

Although this problem is superficially analogous to the "2-marbles problem" considered in the last answer, there is a subtle difference. In the 2-marbles problem, we have a basis for temporally isolating the "loading stage", or skewed prior distribution, from the "two-heads" stage, or contents of the box. This lets us isolate the box and data from the prior distribution and compute an appropriate conditional probability. However, the "loaded 2-headed coin" problem is not so readily decomposed into distinct conditional probabilities. Instead of computing each probability separately, we must instead compute a joint probability for the coin being loaded AND/OR two-headed!

In the loaded 2-headed coin problem, the lack of a spatiotemporal boundary between two independent parameters, balance and stamping, severs the relationship between frequency and proportion. Balance affects frequency, but frequency of what? And stamping, which affects proportion, obviously affects frequency as well, but only if given a fair chance...i.e., only if the coin is not so unbalanced that only one face can show up.

Is there any way to apply conditional probability to this problem? Let's try it. By making two independent symmetrizing assumptions - "the coin is fair" and "the coin is normally stamped" (with a head on one side and tail on the other) - we can calculate conditional probabilities for two-headedness and loading repectively. But since one or both of these assumptions may be false, we must again calculate P(coin is loaded AND/OR two-headed). In the 2-marbles problem, no falsifiable assumptions are required.

Ironically, it is the temporal *dependence* of the 2-marble distribution on the prior distribution that lets the problem be decomposed into *distinct* conditional probabilities, whereas it is the *independence* of stamping and balancing that *prevents* an analogous decomposition in the coin problem. In the 2-marbles problem, dependence has created a convenient boundary between cause and effect, whereas in the coin problem no such boundary exists. If this is what Chris Cole was trying to say, then he almost had a point.

QUESTION: Are the problems we've been discussing - Newcomb's paradox, the 10-marbles problem, and Marilyn's 2-boys paradox - related?

ANSWER: Yes. We've already seen that the 10-marbles and 2-boys problems can both be viewed in terms of conditional probability. And in Newcomb's paradox, the NST model (which corresponds to the expected utility argument) inductively embeds the classical-physics model (which corresponds to the dominance argument) much as a conditional probability P(B|A) embeds the component probability P(B).

More than anything else, this is what has confused me about Chris Cole's opinions regarding these problems. In Newcomb's paradox, he summarily restricts the context to one-way time and excludes logical regression to the NST, burning the upward bridge from dominance to expected utility. But in the 10-marbles problem, he starts at the prior distribution and burns the downward bridge from cause to effect by denying the existence, for Bayesian purposes, of an intermediate context - namely, the closed and finite box-data ensemble! He likes opposite sides of the bridge in each problem, but torches it with equal fire from either direction.

I can only hope that Chris, with whom I've conversed and whose intellect I found sharp and stimulating, will stop selling himself short by clinging to self-contradictory and mutually-contradictory positions on problems which have been solved within an inch of their lives directly under his nose.

QUESTION: What is the bearing of all of this on the Mega Society?

ANSWER: As long as IQ societies have existed, they have been bent on justifying their existence. One of the more profound justifications seems to go something like this. The world is in dire need of a combination Darwin, Einstein and Ghandi - a "secular saviour", so to speak. Unfortunately, the world tends to be markedly unfriendly to genius. So a budding Aristotle, Newton or Voltaire needs all the help he (or she) can get, and it may be up to the IQ societies to help him rise above the tide of mediocrity.

Obviously, if the people who run these societies - the ones who determine who gets recognized for what - lack the ability and fairness to see and acknowledge anyone else's achievements, then the above justification is nothing but a flimsy facade. That would not only denigrate the whole concept of these organizations, but vastly diminish their intrinsic value and potential standing in the intellectual world. And it would give their most promising members - the ones who might actually have what it takes to make a beneficial mark on civilization - far less to lose by walking away from them.

The kind and quantity of resistance I've encountered from the hi-Q politicos of the Mega Society cannot have been lost on most of our members. Regardless of how much animus anyone bears towards me personally - and as far as I'm concerned, I've done nothing to deserve any of it - my own treatment suggests that if a new Great Intellect *does* happen to appear among our ranks, recognition may not come. In fact, it might even be taken to suggest that this lifeboat in the heartless sea of mediocrity makes up for its lack of a rudder with just another good-old-boy political hit list.

I think we can all agree that the world deserves better.

****