MATHEMATICAL GAMES

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

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 paper.

		BEING	
		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

Payoff matrix for Newcomb's paradox

Two closed boxes, B1 and B2, are on a table. B1 contains $1,000. B2 contains either nothing or $1 million. You do not know which. You have an irrevocable choice between two actions:

1. Take what is in both boxes.
2. Take only what is in B2.

At some time before the test a superior Being has made a prediction about what you will decide. It is not necessary to assume determinism, only that you are persuaded that the Being's predictions are "almost certainly" correct. If you like, you can think of the Being as being God, but the paradox is just as strong if you regard the Being as a superior intelligence from another planet, or a supercomputer capable of probing your brain and making highly accurate predictions about your decisions. If the Being expects you to choose both boxes, he has left B2 empty. If he expects you to take only B2, he has put $1 million in it. (If he expects you to randomize your choice by, say, flipping a coin, he has left B2 empty.) In all cases B1 contains $1,000. You understand the situation fully, the Being knows you understand, you know that he knows and so on.

What should you do? Clearly it is not to your advantage to flip a coin, so that you must decide on your own. The paradox lies in the disturbing fact that a strong argument can be made for either decision. Both arguments cannot be right. The problem is to explain why one is wrong.

Let us look first at the argument for taking only B2. You believe the Being is an excellent predictor. If you take both boxes, the Being almost certainly will have anticipated your action and have left B2 empty. You will get only the $1,000 in B1. Contrariwise, if you take only B2, the Being, expecting that, almost certainly will have placed $1 million in it. Clearly it is to your advantage to take only B2.

Convincing? Yes, but the Being made his prediction, say a week ago, and then left. Either he put the $1 million in B2 or he did not. "If the money is already there, it will stay there whatever you choose. It is not going to disappear. If it is not already there, it is not going to suddenly appear if you choose only what is in the second box." It is assumed that no "backward causality" is operating, that is, your present actions cannot influence what the Being did last week. So why not take both boxes and get everything that is there? If B2 is filled, you get $1,001,000. If it is empty, you get at least $1,000. If you are so foolish as to take only B2, you know you cannot get more than $1 million, and there is even a slight possibility of getting nothing. Clearly it is to your advantage to take both boxes!

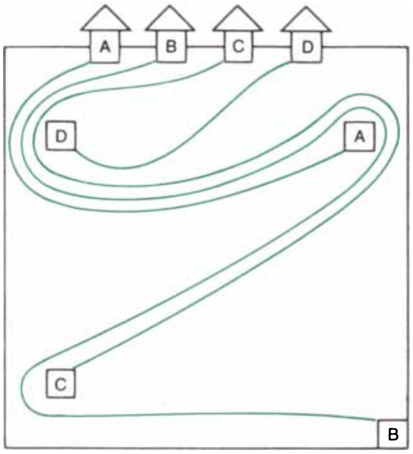
"I have put this problem to a large number of people, both friends and students in class," writes Nozick. "To almost everyone it is perfectly clear and obvious what should be done. The difficulty is that these people seem to divide almost evenly on the problem, with large numbers thinking that the opposing half is just being silly.

"Given two such compelling opposing arguments, it will not do to rest content with one's belief that one knows what to do. Nor will it do to just repeat one of the arguments, loudly and slowly. One must also disarm the opposing argument; explain away its force while showing it due respect."

Nozick sharpens the "pull" of the two arguments as follows. Suppose the experiment has been done many times before. In every case the Being predicted correctly. Those who took both boxes always got only $1,000, those who took only B2 got $1 million. You have no reason to suppose your case will be different. If a friend were observing the scene, it would be completely rational for him to bet, giving high odds, that if you take both boxes, you will get only $1,000. Indeed, if there is a time delay after your choice of both boxes, you know it would be rational for you yourself to bet, offering high odds, that you will get only $1,000. Knowing this, would you not be a fool to take both boxes?

Alas, the other argument makes you out to be just as big a fool if you do not. Assume that B1 is transparent. You see the $1,000 inside. You cannot see into B2, but the far side is transparent and your friend is sitting opposite. He knows whether the box is empty or contains $1 million. Although he says nothing, you realize that whatever the state of B2 is he wants you to take both boxes. He wants you to because, regardless of the state of B2, you are sure to come out ahead by $1,000. Why not take advantage of the fact that the Being played first and cannot alter his move?

Nozick, a specialist in decision theory, approaches the paradox by considering analogous game situations in which, as here, there is a conflict between two respected principles of choice: the "expected-utility principle" and the "dominance principle." To see how the principles apply, consider the payoff matrix for Newcomb's game [*see illustration on opposite page*]. The argument for taking only B2 derives from the principle that you should choose so as to maximize the expected utility (value to you) of the outcome. Game theory calculates the expected utility of each action by multiplying each of its mutually exclusive outcomes by the probability of the outcome, given the action. We have assumed that the Being predicts with near-certainty, but let us be conservative and make the probability a mere .9. The expected utility of taking both boxes is (.1 × $1,001,000) + (.9 × $1,000) = $101,000. The expected utility of taking only B2 is (.9 × $1,000,000) + (.1 × $0) = $900,000. Guided by this principle, your best strategy is to take only the second box.

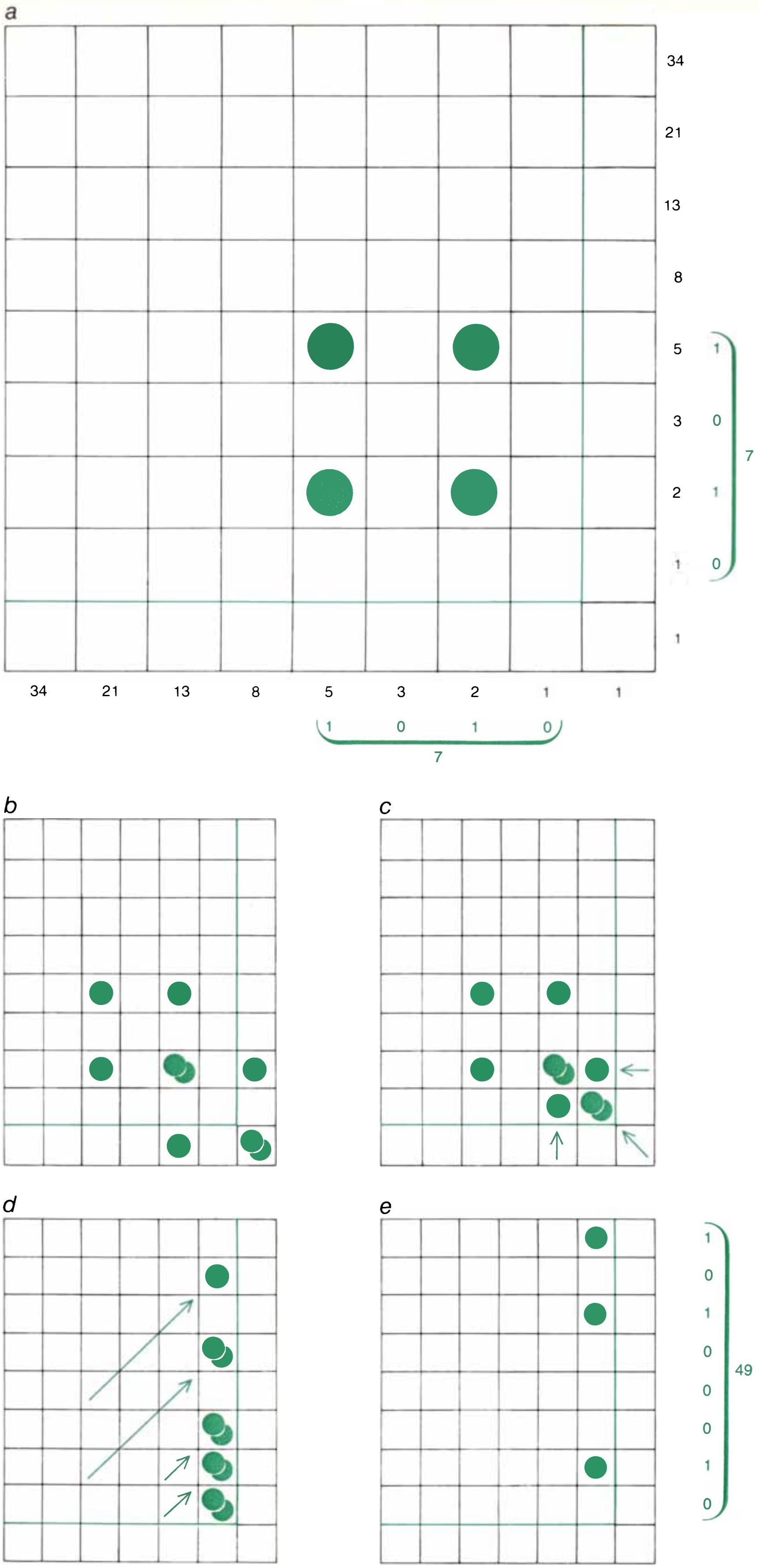


Solution to the schoolhouses problem

The dominance principle, however, is just as intuitively sound. Suppose the world divided into *n* different states. For each state *k* mutually exclusive actions are open to you. If in at least one state you are better off choosing *a*, and in all other states either *a* is the best choice or the choices are equal, then the dominance principle asserts that you should choose *a*. Look again at the payoff matrix on the opposite page. The states are the outcomes of the Being's two moves. Taking both boxes is strongly dominant. For each state it gives you $1,000 more than you would get by taking only the second box.

That is as far as we can go into Nozick's analysis, but interested readers should look it up for its mind-boggling conflict situations related to Newcomb's problem. Nozick finally arrives at the following tentative conclusions:

If you believe in absolute determinism, and that the Being has in truth predicted your behavior with unswerving accuracy, you should "choose" (whatever that can mean!) to take only B2. For example, suppose the Being is God and you are a devout Calvinist, con-



Fibonacci notation for 7 × 7

vinced that God knows every detail of your future. Or assume that the Being has a time-traveling device he can launch into the future and bring back with a motion picture of what you in fact did on that future occasion when you made your choice. Believing that, you should take only B2, firmly persuaded that your feeling of having made a genuine choice is sheer illusion.

Nozick reminds us, however, that Newcomb's paradox does *not* assume that the Being has perfect predictive power. If you believe that you possess a tiny bit of free will (or alternatively that the Being is sometimes wrong, say once in every 20 billion cases), then this may be one of the times the Being has erred. Your wisest decision is to take both boxes.

Nozick is not happy with this conclusion. "Could the difference between one in *n* and none in *n*, for arbitrarily large finite *n*, make this difference? And how exactly does the fact that the predictor is certain to have been correct dissolve the force of the dominance argument?" Both questions are left unanswered. Nozick hopes that publishing the problem "may call forth a solution which will enable me to stop returning, periodically, to it."

One such solution, "to restore [Nozick's] peace of mind," was attempted by Maya Bar-Hillel and Avishai Margalit of Hebrew University in Jerusalem in their paper "Newcomb's Paradox Revisited," in *British Journal for the Philosophy of Science*, Volume 23 (1972), pages 295–304. They adopt the same game-theory approach taken by Nozick but come to an opposite conclusion. Even though the Being is not a perfect predictor, they recommend taking only the second box. You must, they argue, resign yourself to the fact that your best strategy is to behave *as if* the Being has made a correct prediction, even though you know there is a slight chance he has erred. You know he has played before you, but you cannot do better than to play as if he is going to play after you. "For you cannot outwit the Being except by knowing what he predicted, but you cannot know, or even meaningfully guess, at what he predicted before actually making your final choice."

It may seem to you, Bar-Hillel and Margalit write, that backward causality is operating—that somehow your choice makes the $1 million more likely to be in the second box—but this is pure flimflam. You choose only B2 "because it is inductively known to correlate remarkably with the existence of this sum in

Other airlines fly as fast as we do. And their aircraft have the same range as ours. How can we then offer faster services and fewer stops between Europe and

S.E. Asia, Japan, Australia, N. America, West.

And between S.E. Asia and N. America, East.

How, indeed?

The answer is simple: *Copenhagen*. Or, more precisely, the geographical position of Copenhagen, the main gateway of our intercontinental routes.

Look at a globe. Or, better still, stretch a string over its surface between points within the areas above.

In many cases you will find that the string runs straight over Copenhagen (try, for instance, Paris-Tokyo, or London-Sydney, or Berlin-Seattle, or Glasgow-Singapore).

In other cases, Copenhagen will at least be closer to the stretched string than any other major airport in Europe (like Bangkok-New York, which makes the string run as far north as Spitsbergen).

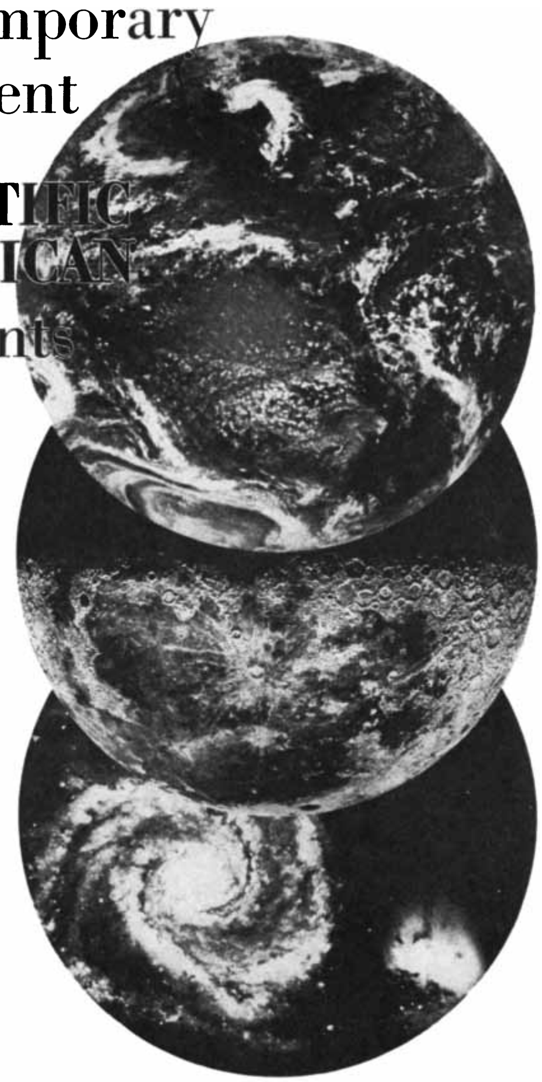
So it's as simple as that: it's shorter via Copenhagen. And flying via Copenhagen is like running the inner track: a way to save distance and time.

And Copenhagen is the only major airport in Europe, from which nonstop flights to Bangkok, Los Angeles, Tokyo and Seattle are all within the reach of long-distance aircraft.

When you're going far east or far west you have two alternatives:

FLY VIA COPENHAGEN–OR FACE A DETOUR

SAS SCANDINAVIAN AIRLINES
General Agent for Thai International



the box, and though we do not assume a causal relationship, there is no better alternative strategy than to behave as if the relationship was, in fact, causal."

For those who argue for taking only B2 on the grounds that causality is independent of the direction of time—that your decision actually "causes" the second box to be either empty or filled with $1 million—Newcomb proposed the following variant of his paradox. Both boxes are transparent. B1 contains the usual $1,000. B2 contains a piece of paper with a fairly large integer written on it. You do not know whether the number is prime or composite. If it proves to be prime (you must not test it, of course, until after you have made your choice), then you get $1 million. The Being has chosen a prime number if he predicts you will take only B2 but has picked a composite number if he predicts you will take both boxes.

Obviously you cannot by an act of will make the large number change from prime to composite or vice versa. The nature of the number is fixed for eternity. So why not take both boxes? If it is prime, you get $1,001,000. If it is not, you get at least $1,000. (Instead of a number B2 could contain any statement of a decidable mathematical fact that you do not investigate until after your choice.)

It is easy to think of other variations. For example, there are 100 little boxes each holding a $10 bill. If the Being expects you to take all of them, he has put nothing else in them. But if he expects you to take only one box—perhaps you pick it at random—he has added to that box a large diamond. There have been thousands of previous tests, half of them involving you as a player. Each time, with possibly a few exceptions, the player who took a single box got the diamond and the player who took all the boxes got only the money. Acting pragmatically, on the basis of past experience, you should take only one box. But then how can you refute the logic of the argument that says you have everything to gain and nothing to lose if the next time you play you take all the boxes?

These variants add nothing essentially new. With reference to the original version Nozick halfheartedly recommends taking both boxes. Bar-Hillel and Margalit strongly urge you to "join the millionaire's club" by taking only B2. That is also the view of Kruskal and Newcomb. But has either side really done more than just repeat its case "loudly and slowly"? Can it be that Newcomb's paradox validates free will by invalidating the possibility, in principle, of a predictor capable of guessing a person's choice between two equally rational actions with better than 50 percent accuracy?

What does the reader think? I cannot answer letters, but in a later piece I shall report on which side got the largest vote and comment on letters of particular interest.

The first of last month's questions asked for a formula giving the maximum number of noncrossing edges that can be drawn as part of a complete graph for n points. It is $3n - 6$, for n greater than 2. The corresponding formula for complete bipartite graphs of m,n points is $2(m + n) - 4$, for m and n each greater than 1. "Odd," a friend once remarked of this second formula, "that the number is always even." Proofs of both cases are not difficult. These formulas for noncrossing edges are of no help in finding formulas for crossing numbers because there is no known way to predict the minimum number of crossings produced by the edges not drawn.

One solution to the four-schoolhouses puzzle, in which four boys have to reach their respective schools without any of their paths crossing one another or going outside the boundary, is shown in the illustration on page 105.

John Harris of Santa Barbara, Calif., discovered an ingenious way to multiply numbers in Fibonacci notation, using the Napier counting board described in April. He added an extra 1-row and 1-column outside the heavy line to the counting board [*see illustration on page 106*]. Suppose you want to multiply 7 by 7. Place the counters according to Napier's rules [*see "a" in the illustration*]. More counters are now positioned according to the following rule: On the diagonal that extends down and to the right from each counter, n, put a counter on every alternate cell, starting with the cell two cells away from counter n [*b*].

Each counter outside the heavy line is moved to the nearest cell inside the line [*c*]. Now move all counters up and to the right along their diagonals to the heavy line [*d*]. Clear the column according to the Fibonacci clearing rules given in April [*e*]. The counters, reading from the top down, give the correct product in Fibonacci notation. Readers familiar with the Fibonacci series will enjoy proving that Harris' algorithm works. Division by this method, however, seems to be hopelessly complicated.