28

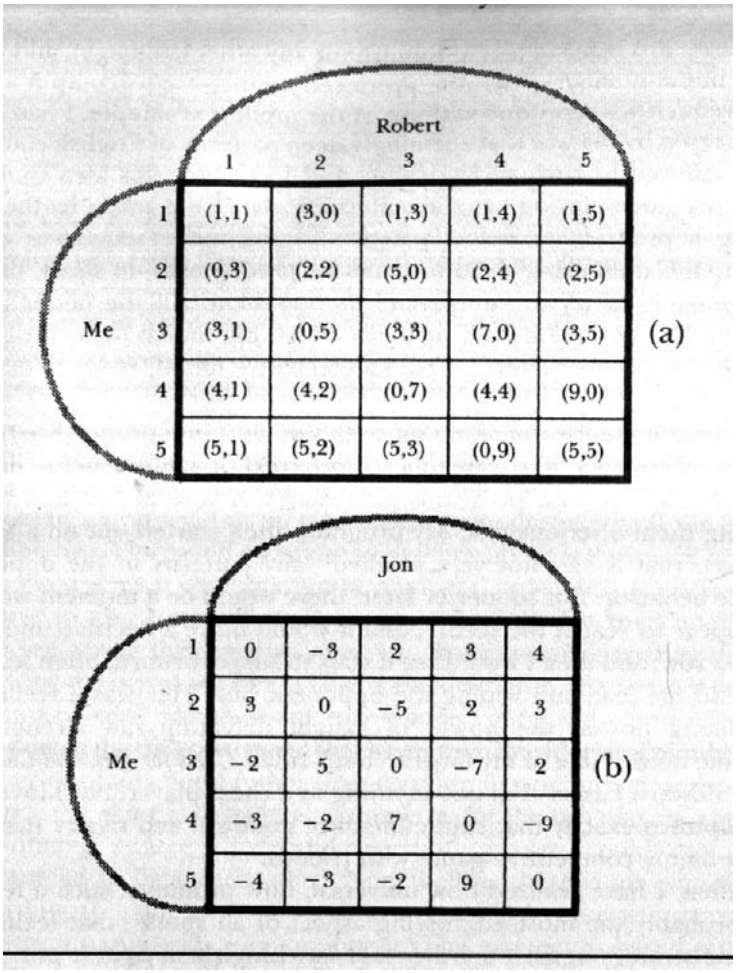
Undercut, Flaunt, Pounce, and Mediocrity: Psychological Games with Numbers

August, 1982

IN the summer of 1962, Robert Boeninger and I, both young mathematics students at Stanford, were riding in a bus somewhere in southern Germany on the way back from a brief trip to Prague, when we got bored. Out of the blue, we invented a curious game with numbers. Though the rules of our game were very simple, it was nonetheless very tricky to play, for it involved trying to "psych each other out" in devious ways. The rules we initially made up went like this: The game would consist of ten turns. On each turn, we'd each choose a number in secret, and then we'd compare them. One of us would choose a number-an integer-in the range 1-5, the other one an integer in the range 2-6. Each of us would get to "keep" his own number, that is, to add it to his score-provided they did not differ by 1. But in the case of two successive integers, the player with the lower of the two numbers collected both of them. So if I said 2 but Robert said 3, well then, I'd get 5 points, and poor Robert, none. Very jolly! At least until I said 5 and Robert said 4. Then not so jolly.

It seemed amusing to have the ranges not quite coincide, since it's hard to sort out who really has the advantage. One's intuitive first impression might be that the 2-6, or "larger", player has an advantage, but that is nicely counterbalanced by the fact that if you name 6, you're running the risk of being undercut by your opponent's 5, whereas your 6 itself can undercut nothing!. Moreover, the "small" player can always name 1 safely, without any risk of being undercut.

Although the asymmetry seemed charming, we soon decided that having equal ranges-both 1-5-was probably preferable. And that was the way we played the game, which I shall here call "Undercut". A table showing how much both of us stand to lose or gain for each possible pair of choices is shown in Figure 28-la. Such a table is known as a *payoff matrix*.



Robert

Me	1	2	3	4	5
1	(1,1)	(3,0)	(1,3)	(1,4)	(1,5)
2	(0,3)	(2,2)	(5,0)	(2,4)	(2,5)
3	(3,1)	(0,5)	(3,3)	(7,0)	(3,5)
4	(4,1)	(4,2)	(0,7)	(4,4)	(9,0)
5	(5,1)	(5,2)	(5,3)	(0,9)	(5,5)

(a)

Jon

Me	1	2	3	4	5
1	0	−3	2	3	4
2	3	0	−5	2	3
3	−2	5	0	−7	2
4	−3	−2	7	0	−9
5	−4	−3	−2	9	0

(b)

FIGURE 28-1. *In (a), the payoff matrix for the game of Undercut, as Robert Boeninger and I originally invented it. in each parenthesis pair, my payoff is on the left and Robert's is on the right. In (b), Jon Peterson's way of looking at things. This matrix exhibits the difference between Jon's profit and my profit-in other words, his net gain over me-for each choice of moves we might make. Looked at this way, Undercut is a zero-sum game*.

Competition was pretty fierce. The lovely thing about this game was how level upon level of "outpsyching" could pile up in our minds. For instance, could "tease" Robert by choosing 4 a few times in a row, trying to lure ,dim into naming 3, and just at that moment plan to switch my move on him, jumping to 2 and outfoxing him. But Robert of course would be most keenly aware of my ploy, and would have his own way of playing naive, leading me on, making me think I could get away with such tricks, and then pulling a higher-order one on me just when I least expected it.

When I returned to Stanford from Europe that fall, I was eager to get a computer to play this game. My friend Charles Brenner had recently written a program that compiled frequencies of letters and letter groups (trigrams,

to be precise) in a piece of text in English (or any other language), and then, using a random-number generator, produced pseudo-English output whose trigram frequencies reproduced those of the input text sample. I had been very impressed by the way that seemingly deep patterns of English could be so aptly captured by such an algorithm, and I saw how this idea could be adapted to a game-playing program. In particular, I was taken by the idea of getting a program to detect patterns in the move sequences of its opponent, and then using them to generate predictions-in short, having the computer itself try to "outpsych" its opponent. All the better if the opponent program were also trying to do something similar to my program! The more vicious, the better! I was in a combative mood, ready to take on all comers.

I have vivid memories of standing over the loud line-printer where the output would spill out, and watching the progress of games emerge line by line. We would have our programs play games of several hundred turns, thus giving them a serious test. My program often started out on a losing track, given that it had not yet "smelled" any patterns in the opposing program's behavior. But sooner or later, there would be a moment when it would appear to "catch the scent", and it would make a decisive undercut or two in a row, and then I would see it start to surge forward, often leaping quickly into the lead and wiping the opponent, but. This was a feeling of overwhelming power, the power of insight defeating raw strength. It reminds me now of one of my favorite book titles: Chess for Fun and Chess for Blood, by Edward Lasker. I'm not anything as a chess player, but I love that title. It captures exactly that subtle blend of goodwill and rivalry that one feels in a highly competitive game with friends.

Since then, I have realized how universal, how primitive, such a feeling is. It is probably the most engrossing aspect of all sports, that feeling of pitting one strategy against another and watching them fight it out. Dogs certainly seem to experience this feeling with pleasure. When I play a chasing game with my friend Shandy the Airedale, I detect in him a precise sense for how well I can anticipate his moves: in his dodging tactics, he always stays one ply-one level of trickery-ahead of me. Whenever I think I have caught on to his pattern, he somehow senses it, and just at that moment shifts his strategy so that I wind up lunging for a dog that is not there.

Oh, to be sure, he lets me win sometimes just to keep me interested. He even has the instinct of teasing, dropping his prized stick right in front of me, acting nonchalant about it and coolly tempting me to make a move for it. But he has it all calculated out. He knows how quick I am, how quick he is, and what my patterns of trying to fake him out are.

What's more, Shandy often seems to come up with new ways of shifting his strategy, so that I cannot simply catch onto the "meta-pattern" of his strategy-shifts and thereby outwit him. No, there is something extremely cunning in his dog's mind, and clearly the joyous exercise of that native

intelligence reflects a deeper quality of dogs and people in general, namely the enormous evolutionary advantage that intelligence seems to confer on beings that have it, in this dog-eat-dog, people-eat-people world.

* * *

But back to Undercut. One day, a math graduate student named Jon Peterson who used to hang around the Stanford "comp center", as it was known back then, challenged my program with his own. He said he had used game theory in his program. I wasn't worried. But when I pitted my program against his, I soon saw that there was good cause for worry. It's not that my program got trounced by his; rather that it just never caught on to any patterns, and simply wound up more or less tying him each time. This was baffling. Jon explained that he had computed appropriate weights for each different choice, 1 through 5, weights that had nothing to do with the opponent's strategy, but merely with the amount of payoff for each set of possibilities. The payoff matrix he was talking about is shown in Figure 28-lb. It shows, for each combination of numbers, how much Jon stands to gain relative to me. Notice the antisymmetry-the fact that each number, when reflected across the diagonal of zeros, changes sign, signaling that what is good for me is bad for him. (That is the definition of a zero-sum game: when the two players' scores in any turn always cancel out.) And of course the zeros down the diagonal mean that when we name identical numbers, it does neither of us any good (other than carrying us one turn closer to the end).

Since the game is completely symmetric for the two players, there can be no winning strategy, for otherwise both players could use it and be guaranteed to beat each other. Nonetheless, there is an optimal strategy, according to game theory, which in a statistical sense will guarantee you long-term parity with your opponent. This strategy is based on assigning statistical weights to the five numbers. To find those weights, you have to solve five simultaneous homogeneous linear equations. Each equation is based on making your expectation equal to zero. If Jon's weights for playing 1, 2, 3, 4, 5 are, respectively, a, b, c, d, e, then my expectation, when I choose, say, 3, will be - 2a + 5b - 7d + 2e (read straight off the third row of the payoff matrix). Set this expectation to zero and you have one of the five equations. The other four arise analogously. The system to solve is thus:

(1) -3b+2c+3d+4e=0
(2) 3a -5c+2d+3e=0
(3) -2a+5b -7d+2e=0
(4) -3a-2b+7c -9e=0
(5) -4a-3b-2c+9d =0

This amounts to inverting a 4 X 4 matrix. Jon had done so, and came up with the following weights: 10, 26, 13, 16, and 1, for choosing 1, 2, 3, 4, and

5 respectively. Thus, according to game theory, an optimal player should play 5 very seldom: one time out of 66. And 2 should be the most common choice. However, it would do little good to play ten l's in a row, followed by 26 2's, 13 3's, and so on. One must choose completely randomly, given these weights.

Imagine a 66-sided die on ten of whose faces the number 1 appears, only one face having 5 written on it, and so on. Each move, you must throw such a die (or a suitable computational simulation thereof). In other words, you must avoid any and all patterned behavior when you play according to this strategy. No matter how tempting it might be, you must not yield! Even if your opponent plays 5 a dozen times in a row, you must totally ignore it and merely keep on throwing your 66-sided die obliviously. That's the way Jon's program played, and it's why my program found nothing to pick up on. Had Jon's program ever given in to temptation and tried to outguess me, my program would likely have picked up some pattern and twisted it back to work against his. But his program knew nothing of temptations or teasing. It just played blindly on, and the longer the game, the more surely it would break even. If it won, so much the better, but it had only a fifty-fifty chance of that. That's the "optimum strategy" for you!

It was a humiliating and infuriating experience for me to watch my program, with all its "intelligence", struggle in vain to overcome the blind randomness of Jon's program. But there was no way out. I was most disappointed to learn that, in some sense, the "most intelligent" strategy of all not only was dumb-it even paid no attention whatsoever to the enemy's moves! Something about this seemed directly opposite to the original aim of Undercut, which was to have players trying to psych each other out to ever deeper levels.

* * *

When I saw the game so completely demolished by game theory, I abandoned it. Recently, however, I have returned to thinking about such games in which patterns in one's play can be taken advantage of, even if game theory in some theoretical sense can find the optimal strategy. There is still something curiously compelling and fascinating about the teasing and flirting and other ploys that arise in these games, something that vividly recalls strategies in evolution, and even seems relevant to many political situations today.

Furthermore, there is something strikingly academic and bookish to adopting a purely game-theoretic strategy when playing against a human opponent, especially in the face of "teasing" strategies. Obviously, humans have more complex goals in life than merely winning the game, and this fact determines a lot about how they play a game. Impatience and audacity, for instance, are both important psychological elements in human game-playing, and an optimal strategy in the ordinary game-theoretic sense does not take those into account. Therefore I feel games of this sort are still

important models of how people and larger organizations tackle complex challenges and threats.

Let me describe, therefore, some more recent variations on Undercut that I have been experimenting with. They all involve extending the degree to which one can go out on a limb by "baiting" one's opponent. My purpose was to encourage teasing, which means that one player flaunts a pattern for a while, implicitly saying, "I dare you just try an undercut!" So to $W' encourage this kind of pattern-flaunting, it seemed reasonable to award patterns points whenever they are not picked up on by the enemy. Let's call this version "Flaunt".

Suppose that you and I are playing Flaunt. I say 4, you say 1. As in Undercut, I get 4 points, you get 1. Now on my next turn, suppose I again say 4, and you say 2. If we were playing Undercut, I would again get 4. But in Flaunt, repetitions are rewarded. Therefore, I am given the product of my two numbers: 4 X 4, or 16. Now suppose that on my next turn I again play 4, while you again play 2. My bravado now earns me 4 X 4 X 4, or 64, points, while you get 2 X 2, or 4. So in these three turns, I have gained 4+ 16+64, or 84 points, to your 1 +2+4, or 7. Of course, you have not been oblivious to my prancing-about in front of you-you have merely been biding your time. Now you make your move-a 3-hoping to undercut me. Too badI chose 2! I get 5 points, you get nothing. Sorry, sucker.

But what if I had been so dumb as to let you catch me at this? If I had indeed said 4 this time, I would have been hoping for 256 points. But as you successfully undercut my pattern with your 3, you get a high reward for this, namely 259 points (your 3 points plus "my" 256 points).

* * *

Now Flaunt, like haircuts, can come in various styles. The one I have just presented is the simplest. But more complex patterned behavior can also be rewarded, if you like. I am not sure of the best way to do this, so what follows -the game I call "Superflaunt"-is only one possible way to reward pattern-flaunting. Suppose that instead of playing 1-2-2 against my 4-4-4 moves, you had played 2, then 1, then 2. You might well have had a reason for doing so. Maybe it was the continuation of a pattern for you and was worth your while keeping up for the moment. If your previous four moves had been 2-1-2-1, then your recent three moves would have continued that pattern. Depending on how it's scored, extending your own established pattern might be more worthwhile than undermining my relatively new one. If 2-1-2-1-2-1-2 is worth the product of its elements, then that amounts to 16 points. (Actually, it's worth 16 only if it was preceded by another 2-1, but that's beside the point.) By the time you've picked up on my 4-4 pattern, maybe it seems worth it to you to let me have my third 4 while you name one more 2, thinking that that will lull me a bit and at that moment, you will suddenly strike, and undercut me.

So what constitutes a pattern in this game of Superflaunt? At the moment, I'm inclined to limit it to one fairly simple definition, although it might be possible to have more complex definitions. The main idea is that a pattern exists when, in a given situation, you do what you did last time you were in "that situation". So it all hinges on what we mean by the notion of "same situation". Let's say that you have just played x, and are about to play y. We'll say that you are making a pattern provided that the most recent time you played x you also followed it with y. If, for instance, your last seven moves had been 3-4-1-5-3-4-1, then to make your pattern continue, you must play 5, and after that, 3, 4, 1, 5, 3, 4, 1, and so on. When you first establish the sequence 3-4-1-5, you of course get no bonus points, because until the repetitions start, there is not any pattern. Thus only when the second 4 is played has a pattern started, and it nets you 12 (3 X 4) points. The next patterned move, 1, nets you another 12 points, and then saying 5 gives you 60 points (as long as it is not undercut)! But as soon as you break the pattern, your cumulative product must start out again from scratch.

If you had played 3-4-1-5-3-4, and were worried about the obviousness of playing 1 now, you might choose to play 4, which, although it breaks one pattern, establishes another pattern (viz., 4-4). Now in ordinary Flaunt, this in itself would already be worth 16 points, but in Superflaunt only on your next 4 would you begin to reap the benefits of your patterned playing, since only then would you have made "the same choice" in "the same situation" twice in a row.

A limitation of Undercut and Flaunt is that both confine your moves to a small range. I wanted a game in which numbers of arbitrary size were permitted. It was not too hard to come up with the following-game, which I call "Underwhelm". You and I both think of positive integers. Now, if they are unequal and do not differ by 1, then whoever named the lower one gets that number (the other player getting, of course, nothing). If they differ by 1, then the namer of the upper of the two is awarded both numbers. In that respect, Underwhelm is like a tipped-over version of Undercut (another name for it was "Overcut"). If our two numbers are equal, then neither of us gets anything for this turn.

The goal can be a fixed number of points-any number. For example, 1,000 seems a good choice, although 100 or even a million will do just fine. Think about what this does to the game. Clearly it is not useful for you to name huge numbers, because I am likely to name a lower number and then you will get nothing while I will get something. So there is pressure on both of us-it seems-to play fairly small numbers. But if we stick to very small numbers, then the likelihood of being "overcut" is fairly high. Furthermore, the scores will advance very, very slowly. If we are progressing toward the goal of 1,000 points at a snail's pace, someone will want to speed things up. And so someone will go out on a limb, naming a big number like 81. Of course, doing so just once is not useful, because the other player will not have known in advance that that 81 was coming.

But suppose that I say 81 several times in a row. (Pattern-flaunting is not rewarded in Underwhelm, by the way-at least not in this version.) You will soon catch on, and may well be tempted to say 82, to overcut me. Or perhaps you will want to make points more conservatively off my foolishness, by simply choosing numbers close to but below 81, such as 70. Aha! Once I've managed to lure you up into my vicinity, then of course I can start trying =to jump below you. And maybe I can even anticipate just when you'll "bite". If so, then I can really take you to the cleaners.

The interesting thing about Underwhelm is that by using obvious patterns as bait to lure the opponent, either one of us can in essence establish one or more Undercut-like games at various positions along the number line. I can set one up in the vicinity of 81, trying to coax you into saying 82 just when I anticipate it. Meanwhile, you may be playing a baiting game down around 30, getting 30 points each time I extravagantly bait you with my 81, and you know that sooner or later I am bound to try to catch you there, either going below you or overcutting you.

What I find fascinating is how many parallel subgames of this type can arise spontaneously in a game of Underwhelm. Particularly interesting is what happens toward the end, when one player has a significant lead. At that point, the trailing player will tend to play very conservatively, naming very small numbers. This means that the possibilities for overcutting are much enhanced. Moreover, there is an entirely psychological element to this game having to do with human impatience. Nobody wants to dawdle to victory by choosing smallish numbers over and over again several hundred times. Therefore, the simple quest for some variety will inevitably lead to some quirky, daring play -every once in a while, and that will of course be exploitable.

Much of the spontaneous and creative teasing behavior that tends to occur in these games has its parallels in evolution. The most picturesque and vivid portrayal that I know of the uncanny patterns and counter patterns that are set up by living beings competing against each other is provided by Richard Dawkins in his book The Seth Gene. The discussion centers around the notion of an evolutionary stable strategy, or ESS-a term due to J. Maynard Smith. An ESS is defined as: "a strategy which, if most members of a population adopt it, cannot be bettered by an alternative strategy". However, here, "adoption of a strategy by an individual" really means that that individual has genes for that behavioral policy. It's not a question of choice.

Dawkins' first example of this concept involves rival genes for two types of aggressive behavior in a given species. The two strategies are named "hawk" and "dove", and have the recent political connotations of those terms. If x positive points are assigned for winning a fight, y negative points for wasting time, and z negative points for getting injured, one can calculate,

as a function of x, y, and z, the eventual optimal balance of hawks and doves in the population. This may be an average over time, involving swings back and forth between mostly having hawks and mostly having doves, or it may represent an eventual equilibrium in which the ratio is stable.

Dawkins considers a wide variety of colorful everyday examples in human life, carefully comparing them to strategies in the world of nonhuman evolution. Such things as gas wars, with their price-fixing and treacherous undercutting, fall very neatly into line with the game-theoretic analysis that he brings to bear. Some other strategies considered are: "retaliator" (an individual who, when attacked by a hawk, behaves like a hawk, and when attacked by a dove, behaves like a dove); "bully" (who goes around behaving like a hawk until somebody hits back, then immediately runs away); "proper-retaliator" (who is like a retaliator, but who occasionally tries a brief experimental escalation of the contest). These five strategies can all be activated simultaneously in a computer simulation of a large population, just as the strategies in Undercut could fight each other on a computer. From such simulations, one can learn about the optimum strategies without doing the game theory. In essence, Dawkins maintains, this is what nature has done over eons: Vast numbers of strategies have fought each other, nature's profligacy paying off in the long run in the development of species with optimal strategies, in some sense of the term.

Dawkins uses this concept to show how group selection can seem to be taking place in a population, when in fact mere gene selection can account for what is observed. He says:

> Maynard Smith's concept of the ESS will enable us, for the first time, to see clearly how a collection of independent selfish entities can come to resemble a single organized whole .... Selection at the low level of the single gene can give the impression of selection at some higher level.

The book contains many other provocative examples of peculiar strategies that offer sometimes frightening parallels to situations in the world of human politics, often reminding me of the dangers of the current arms race. In fact, the connection is made explicitly by Dawkins more than once. He refers to "evolutionary arms races" and the survival value of deception of one species by another.

One of the funnier parts of Dawkins' book, although it is dead serious, is concerned with the evolution of sexuality. To show how sexuality might have evolved, he invents "sneaky" versus "honest" gametes (fertilized eggs) and shows how, over many generations, the former will slowly evolve into males, the latter into females. Along the way, such amusingly named strategies are discussed as the "domestic-bliss strategy", the "he-man strategy", the "coy" and "fast" strategies (limited to females), and the "faithful" and "philanderer" strategies (limited to males). Dawkins emphasizes that these are only metaphors, and are not to be taken literally (and certainly not

anthropomorphically). When one takes them with the proper grain of salt, however, they can enormously illuminate the mechanisms of evolution. And many of these strategies find their counterparts in such number games as we have discussed above.

* * *

As I was preparing this article, I had a long phone conversation with Robert Boeninger in which we tried out various versions of these games. One idea that intrigued me was to play Underwhelm, but with no specific target number of points, such as 1,000, in mind. Instead, a convention of another sort would terminate play. My candidate for that convention was "Stop when the two players' numbers differ by 2." Thus if I say 10 and you say 8, that marks the game's end (and neither of us gets any points on that turn).

Robert and I tried this version out, and quickly discovered that whenever somebody started losing, they would have no option but to go for a stalemate-a nonterminating game. One way for the losing player to do this is to name huge arbitrary numbers, so that they cannot be anticipated and so the condition for termination is never met. The player who is ahead, having nothing to lose, will cooperate by naming small numbers all the time, thereby gaining even more points and building up even more of a lead. So you get a kind of vicious circle in which both players wind up cooperating in a stalemate.

Robert suggested that one way to prevent this is to add the condition that if either player wins five turns in a row (i.e., gets a positive number of points five times in a row), then the game is over. This prevents the player who is trailing from going for the stalemate, because such behavior will now ensure loss. As Robert amusingly pointed out, even if you are behind, you can start to "wind things up" by trying to win five turns in a row, for by the time those five turns have passed, you may be in the lead! My name for this game is "Pounce", since it made me feel like a tiger hunting down a giraffe in the savannah, bringing down my prey in one swift, sudden move.

* * *

One day, several years after the Undercut episode, my sister Laura and our friend Michael Goldhaber and I were having lunch in the Peninsula Creamery and jotting down various trivia on napkins, as was our wont, and somehow it came to us to play a number game involving three persons. We decided that on each turn, each of us would choose a number in a certain range, and, since it seemed too boring to let the biggest number win, and equally boring if the littlest number won, it became obvious that the middlemost number should be rewarded. So we decided that on each turn, only the "most mediocre" player's score would be allowed to increase. It

would increase, /of course, by the mediocre number; the other players' scores would stay fixed. (A bit of a problem was posed when two players chose tying numbers, but we found a makeshift way of handling that case.)

Thus at the end of, say, five turns, we would all compare our scores, and the highest one ... No, wait a minute. Why should we let the highest score win? To do that would be, after all, contrary to the spirit of each turn. We saw quite clearly that, if the spirit of the whole was to be consistent with the spirit of its parts, then the player whose score was in the middle should win! We called our game "Mediocrity", but occasionally I like to refer to it as "Hruska".

This name was inspired by a famous remark by the then senator from Nebraska, Roman Hruska. In those days (the early 1970's), President Nixon was attempting to get G. Harrold Carswell appointed to the Supreme Court, against the vehement opposition of Indiana senator Birch Bayh and others. In a radio interview defending Carswell against his critics, Senator Hruska came out with the following profundity:

> Even if he were mediocre, there are a lot of mediocre judges and people and lawyers. They are entitled to a little representation, aren't they, and a little chance? We can't have all Brandeises and Frankfurters and Cardozos and stuff like that there.

Alas for mediocrity, Carswell's nomination was defeated. But it worked out fine for Hruska, who shall forevermore be known as a champion of mediocrity-and stuff like that there.

Speaking of champs, after eating our sandwiches and drinking our thick, rich, old-fashioned milkshakes (served in metal containers, to boot!), the three of us sat in our booth and played a few rounds of this quirky game, and what came to our minds but the inspired idea of determining the World Champion of Mediocrity! So we totaled up our scores over several games, to see who had come out highest. Highest?! Again something seemed wrong. The pervasive spirit of mediocrity that had settled on us that day like a heavy smog urged us to deem Champion not the player who had won the most games, not the player who had won the fewest games, but the player who had won the middlemost number of games. Which we did, and I forget who it was. This may be appropriate.

At that point, a general principle seemed to be emerging, which created a hierarchy of levels of Mediocrity. To win at Level Two (that is, our "Championship" level), it's best to be a mediocre player at Level One (the single-game level). This means that whereas before it was desirable to be extremely mediocre at choosing mediocre numbers, now it's desirable to be mediocrely mediocre at choosing mediocre numbers. How perverse! How wonderful! How wonderfully perverse! It fits in with a general principle of perversity, a Zen-flavored principle, that applies to many aspects of life: Try too hard, and you wind up a loser.

* * *

After the initial session at the Peninsula Creamery in which the game of Mediocrity was born, I worked on a number of versions of it, trying to polish it and make it into an elegant game. I am not sure if I succeeded, but I would like to present the rules as they presently stand.

The major issue is how to avoid ties-not only ties at Level Zero, but at all higher levels. My current best solution is the following: Let each player have a slightly shifted range, relative to the other two. More concretely, let player A pick integers from, say, 1 to 5. Then players B and C will have staggered ranges: B picks numbers of the form n + 1/3, and C picks numbers of the form n +2/3, where n runs from 1 to 5. Clearly, then, there can be no ties at Level Zero.

Now what happens at Level One? Recall that a Level One game consists of five Level Zero games, in each of which the middlemost number is awarded to the player who chose it, with the other two players getting zero. Well, the first part of this scoring scheme is fine, but the second part has to be modified very slightly in order to avoid ties at higher levels. Suppose that the numbers chosen are as follows: A: 3; B: 2 3; C: 4 3. Having the middle number, A receives 3 points. B and C, however, do not receive zero points each, but the closest positive approximation to zero that they can, given their staggered ranges. Thus, 1/3 of a point goes to B, and 2/3 of a point to C.

The reasoning behind this goes as follows: After five turns, each player has received five numbers of the same form. Player A's five pure integers will add up to a pure integer. Player' B's five numbers of the form n + 1/3 will add up to a number of the form n +2/3, and player C's five numbers of the form n +2/3 will add up to a number of the form n + 1/3. Thus at the next level up, B and C have exchanged roles in terms of the form of their numbers. Consequently, the three total scores at the new level are all of different form and cannot tie, hence there will always be a most mediocre Level One score: a winner.

If we now go on to consider a game at Level Two, we must award points to each Level One game. The "winner" of a Level One game gets, of course, that middling number of points, while the other two players once again receive the closest positive approximations to zero possible, in their respective forms. For player A, this means exactly zero points, as before. However, for.B it now means 2/3, and for C, 1/3. Five games at Level One constitute one game at Level Two. The heretofore tacit "Principle of Uniformity of Levels" compels us to sum up the five Level One numbers to produce Level Two scores. Needless to say, the same reasons as before will prevent tie scores from arising, and so there will always be a Level Two winner.

The same general principle will of course allow us to extend the game of Mediocrity to any number of levels. One game of Level N+ 1 Mediocrity

consists of five Level N games. The winner of each Level N game is awarded their Level N score, and the other two players get the minimum amount (i.e., 0, 1/3, or 2/3) of the form of their scores at that level. These five Level N numbers are added up to yield totals for the three players, and the middlemost one wins.

Actually, there is nothing sacred about always having five Level N turns in a Level N+ 1 game; the "width" could as well be four, or even two. (Multiples of three must be avoided, since after three moves, all three players have scores that are perfect integers, thus allowing ties.) With a width as narrow as two, this allows a very deep (i.e., many-leveled) game to be played much more easily. For instance, with width two, a five-level game of Mediocrity requires only 32 Level Zero turns-whereas with the standard width of five, merely three levels of depth will require 125 Level Zero turns. Moreover, there is nothing sacred about the Level Zero choices being confined to numbers bounded by 5; they could run from 1 to infinity! This is just one of the many possible variants of Mediocrity.

*** * ***

I can testify that the strategy for playing even Level Two Mediocrity gets mighty confusing very quickly. I have played Level Three Mediocrity oil a couple of occasions, and found it completely beyond my reach. I find this both fascinating and frustrating. And think what it implies about world politics, if such simple games as the ones described in this article are so baffling. How much more complex are the "games" of international bargaining, bluffing, and war-making! All of the conceptual messes that we have discussed above have their counterparts (only "squared", so to speak) in international politics. As one watches these huge themes being played out on the world stage, one can hardly help feeling like a single cell in some vast organism whose strategy was set long ago, the consequences of which one can only watch, hoping all will turn out for the best.

Post Scriptum.

Suppose you are playing a very, very short game of Undercut: one turn long. The number of points you receive will be multiplied by 1,000 and then paid to you in dollars. What would you play? The answer must depend on what your goal is. Which interests you more: beating your opponent, or amassing as much money as possible? If the former is your priority, then a score of 9 to 0 (your favor) is no better than a score of 3 to 1: Either way, you win just as fully. But if money is your desire, the former is $6,000 more favorable than the latter. Even more striking: If you both name 5, you have

a tie game-a big disappointment for someone out to win, but for someone out for money, $5,000 is a fine take.

There is thus a big difference between the original payoff matrix for Undercut and Jon Peterson's modified matrix (both shown in Figure 28-1). Jon's matrix looks at the game solely from the point of view of someone who wants to beat the other player. By taking, the difference between payoffs, Jon managed to convert Undercut into a zero-sum game, which he knew to be tractable by the methods of game theory. But if he had left it as Robert and I had formulated it to begin with, it would not have been so easy.

In fact, the original (non-zero-sum) formulation of Undercut subsumes the most famous of all non-zero-sum games: the Prisoner's Dilemma (a treacherous Gordian knot with which the next few chapters deal). I have extracted, in Figure 28-2, just a small fragment of the Undercut payoff

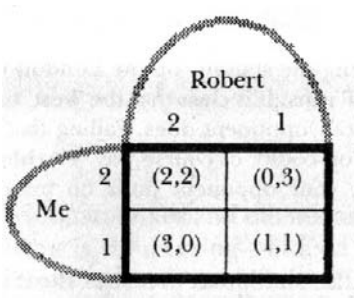


FIGURE 28-2. *A portion of the original Undercut payoff matrix, showing how Undercut contains a Prisoner's Dilemma matrix. (In fact, it contains several.)*

matrix, geometrically rearranged but otherwise intact. This miniature payoff matrix has virtually all the same mathematical qualities as does the standard Prisoner's Dilemma payoff matrix (Figure 29-1). Thus Undercut actually poses a more severe problem than Jon Peterson said. His trick of subtracting one person's payoff from the other's will turn any symmetric game into a zero-sum game, which is tractable by standard techniques of game theory. But that ignores significant aspects of the original game; in particular, for any normal person, losing by 3 to 5 (and getting $3,000) is precisely as good as winning by 3 to 1 (also getting $3,000). But in Jon's matrix, these two events are as opposite as night and day-as opposite as -2 and +2.

* * *

One real-life counterpart to Undercut is given by the following amusing observation. The long-distance telephone rates get much cheaper at 11:00 at night, and so as 11 approaches, the lines get less and less busy, until suddenly, when the hour strikes, the lines get very crowded. In some parts of the country, this "rush hour" prevents you from being able to get a line

at all, which is molt annoying. So you have the option of calling just before 11 and getting an expensive line, or calling just after 11 and taking your chances. Maybe you decide that it's better to call just before 11, and pay the extra amount for the security of getting through. But if everybody thinks of this strategy, then calling just before 11 is self-defeating! So then you have to start pushing your calls back earlier, perhaps even into a more expensive time period ... I guess this is just a new variant on the old "Nobody ever goes there any more because it's always too crowded" joke.

Games of this sort and jokes do indeed have a lot in common. In an article in the British journal Manifold titled "A Pandora's Box of non-Games", Anatole Beck and David Fowler set forth a panoply of rather silly games that are halfway between true games and pure jokes. The tragedy is that so many of them resemble current global political behavior. For instance, consider the game they call Finchley Central:

> Two players alternate naming the stations on the London Underground. The first to say 'Finchley Central' wins. It is clear that the `best' time to say 'Finchley Central' is exactly before your opponent does. Failing that, it is good that he should be considering it. You could, of course, say 'Finchley Central' on your second turn. In that case, your opponent puffs on his cigarette and says, 'Well.... ' Shame on you.

Another amusing game, quite similar to the ones described in the column, is called Penny Pot

> Players alternate turns. At each turn, a player either adds a penny to the pot or takes the pot. Winning player makes first move in next game. Like Finchley Central, this game defies analysis. There is; of course, the stable situation in which each player takes the pot whenever it is not empty. This is a solution?

At the end of their article, Beck and Fowler add:

> M. Henton of New Addington noted with horror that there is an isomorphism between Finchley Central- and the game commonly known as 'Nuclear Deterrent'. 'It occurs to me that we should work very fast to analyse the non-games, before we are left with a non-world.'

Several readers wrote in to tell me that they had worked out by game theory the optimal strategy for playing my game of Underwhelm, and that they had found it involves playing only the numbers between 1 and 5, in the ratios 25:19:27:16:14. Numbers higher than 5 should never be played at all! This was a surprise to me, taking away most of the interest of the game. Oh, well ... as they say in game theory, "You win some, you lose some."

The Prisoner's Dilemma Computer Tournaments and the Evolution of Cooperation

May, 1983

LIFE is filled with paradoxes and dilemmas. Sometimes it even feels as if the essence of living is the sensing-indeed, the savoring-of paradox. Although all paradoxes seem somehow related, some paradoxes seem abstract and philosophical, while others touch on life very directly. A very lifelike paradox is the so-called "Prisoner's Dilemma", discovered in 1950 by Melvin Dresher and Merrill Flood of the RAND Corporation. Albert W. Tucker wrote the first article on it, and in that article he gave it its now-famous name. I shall here present the Prisoner's Dilemma-first as a metaphor, then as a formal problem.

The original formulation in terms of prisoners is a little less clear to the uninitiated, in my experience, than the following one. Assume you possess copious quantities of some item (money, for example), and wish to obtain some amount of another item (perhaps stamps, groceries, diamonds). You arrange a mutually agreeable trade with the only dealer of that item known to you. You are both satisfied with the amounts you will be giving and getting. For some reason, though, your trade must take place in secret. Each of you agrees to leave a bag at a designated place in the forest, and to pick up the other's bag at the other's designated place. Suppose it is clear to both of you that the two of you will never meet or have further dealings with each other again.

Clearly, there is something for each of you to fear: namely, that the other one will leave an empty bag. Obviously, if you both leave full bags, you will both be satisfied; but equally obviously, getting something for clothing is even more satisfying. So you are tempted to leave an empty bag. In fact, you can even reason it through quite rigorously this way: "If the dealer brings a full bag, I'll be better off having left an empty bag, because I'll have gotten

all that I wanted and given away nothing. If the dealer brings an empty bag, I'll be better off having left an empty bag, because I'll not have been cheated I'll have gained nothing but lost nothing either. Thus it seems that no matte what the dealer chooses to do, I'm better off leaving an empty bag. So I'll lease an empty bag."
The dealer, meanwhile, being in more or less the same boat (though at the other end of it), thinks analogous thoughts and comes to the parallel conclusion that it is best to leave an empty bag. And so both of you, with your impeccable (or impeccable-seeming(logic, leave empty bags, and go away empty-handed. How sad, for if you had both just cooperated, you could have each gained something you wanted to have. Does logic prevent cooperation? This is the issue of the Prisoner's Dilemma.

* * *

In case you're wondering why it is called "Prisoner's Dilemma", here's the reason. Imagine that you and an accomplice (someone you have no feelings for one way or the other) committed a crime, and now you've both been apprehended and thrown in jail, and are fearfully awaiting trials. You are being held in separate cells with no way to communicate. The prosecutor offers each of you the following deal (and informs you both that the identical deal is being offered to each of you-and that you both know that as well'): "We hake a lot of circumstantial evidence on you both. So if you both claim innocence, we will convict you anyway and you'll both get two years in jail. But if you will help us out by admitting your guilt and making it easier for us to convict your accomplice-oh, pardon me, your *alleged* accomplice why, then, we'll let you out free. And don't worry- about revenge-your accomplice will be in for five years! How about it?" Warily you ask, "But what if we *both* say we're guilty?" "Ah, well, my, friend-I'm afraid you'll both get four-year sentences, then."

Now you're in a pickle! Clearly, you don't want to claim innocence if your partner has sung, for then you're in for five long years. Better you should both have sung-then you'll only get four. On the other hand, if your partner claims innocence, then the best possible thing for you to do is sing. since then you're out scot-free! So at first sight, it seems obvious what you should do: Sing! But what is obvious to you is equally obvious to your opposite number, so now it looks like you both ought to sing, which means -Sing Sing for four years! At least that's what logic tells you to do. Funny% since if both of you had just been illogical and maintained innocence, you d both be in for only half as long! Ah, logic does it again.

* * *

Let us now go back to the original metaphor and slightly alter its conditions. Suppose that both you and your partner very much want to have

a regular supply of what the other has to offer, and so, before conducting your first exchange, you agree to carry on a lifelong exchange, once a month. You still expect never to meet face to face. In fact, neither of you has any idea how old the other one is, so you can't be very sure of how long this lifelong agreement may go on, but it seems safe to assume it'll go on for a few months anyway, and very likely for years.

Now, what do you do on your first exchange? Taking an empty bag seems fairly nasty as the opening of a relationship-hardly an effective way to build up trust. So suppose you take a full bag. and the dealer brings one as well. Bliss-for a month. Then you both must go back. Empty, or full? Each month, you have to decide whether to *defect* (take an empty bag) or to *cooperate* (take a full one). Suppose that one month, unexpectedly, your dealer defects. Now what do you do? Will you suddenly decide that the dealer can never be trusted again, and from now on always bring empty bags, in effect totally giving up on the whole project forever? Or will you pretend you didn't notice, and continue being friendly? Or-will you try to punish the dealer by some number of defections of your own? One? Two? A random number? An increasing number, depending on how many defections you have experienced? Just how mad will you get?

This is the so-called *iterated* Prisoner's Dilemma. It is a very difficult problem. It can be, and has been, rendered more quantitative and in that form studied with the methods of game theory and computer simulation. How does one quantify it? One builds a *payoff matrix* presenting point values for the various alternatives. A typical one is shown in Figure 29-1a. In this matrix, mutual cooperation earns both parties 2 points (the subjective value of receiving a full hag of what you need while giving up a full hag of what you have). 'Mutual defection earns you both 0 points (the subjective value of gaining nothing and losing nothing, aside from making a vain trip out to the forest that month). Cooperating while the other defects stings: you get -1 point while the rat gets 4 points! Why so many? Because it is so pleasurable to get something for nothing. And of course, should you happen to be a rat some month when the dealer has cooperated, then you get 4 points and the dealer loses 1.

It is obvious that in a *collective* sense, it would be best for both of you to always cooperate. But suppose you have no regard whatsoever for the other. There is no "collective good" you are both working for. You are both supreme egoists. Then what? The meaning of this term, "egoist", can perhaps be made clear by the following. Suppose you and your dealer have developed a trusting relationship of Mutual cooperation over the years, when one day you receive secret and reliable information that the dealer is quite sick and will soon die, probably within a month or two. The dealer has no reason to suspect that you have heard this. Aren't you highly tempted to defect, all of a sudden, despite all your years of cooperating? You are, after all, out for yourself and no one else in this cruel, cruel world. And since it seems that this may very well be the dealer’s last month, why not profit

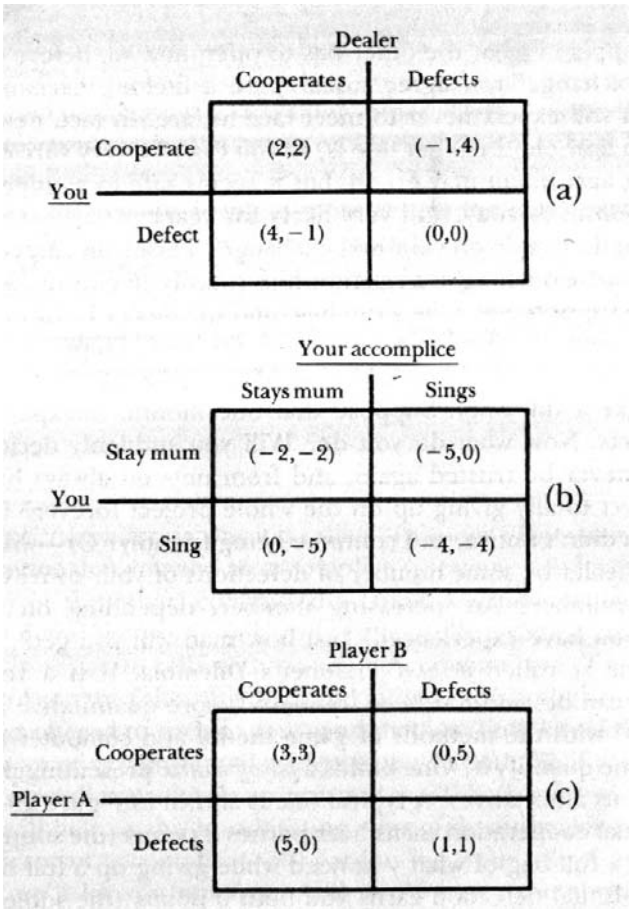


FIGURE 29-1. The Prisoner's Dilemma.

In (a), a Prisoner's Dilemma payoff matrix in the case of a dealer and a buyer of commodities or services, in which both participants have a choice: to cooperate (i.e., to deliver the goods or the payment) or to defect (i.e., to deliver nothing). The numbers attempt to represent the degree of satisfaction of each partner in the transaction.

In (b), the formulation of the Prisoner's Dilemma to which it owes its name: in terms of prisoners and their opportunities for double-crossing or collusion. The numbers are negative because they represent punishments: the length of both prisoners' prospective jail sentences, in years. This metaphor is due to Albert W Tucker.

In (c), a Prisoner's Dilemma formulation where all payoffs are nonnegative numbers. This is my canonical version, following the usage in Robert Axelrod's book, The Evolution of Cooperation.

as much as possible from your secret knowledge? Your defection may never be punished, and at the worst, it will be punished by one last-gasp defection by the dying dealer.

The surer you are that this next turn is to be the very last one, the more you feel you must defect. Either of you would feel that way, of course, on learning that the other one was nearing the end of the rope. This is what is meant by “egoism”. It means you have no feeling of friendliness or

goodwill or compassion for the other player; you have no conscience; all you care about is amassing points, more and more and more of them.

What does the payoff matrix for the other metaphor, the one involving prisoners, look like? It is shown in Figure 29-1b. The equivalence of this matrix to the previous matrix is clear if you add a constant-namely, 4-to all terms in this one. Indeed, we could add any constant to either matrix and the dilemma would remain essentially unchanged. So let us add 5 to this one so as to get rid of all negative payoffs. We get the canonical Prisoner's Dilemma payoff matrix, shown in Figure 29-1c. The number 3 is called the reward for mutual cooperation, or R for short. The number 1 is called the punishment, or P. The number 5 is T, the temptation, and 0 is S, the sucker's payoff. The two conditions that make a matrix represent a Prisoner's Dilemma situation are these:

$$(I)\ T>R>P>S$$

$$(2)\ (T+S)/2 < R$$

The first one simply makes the argument go through for each of you, that "it is better for me to defect no matter what my counterpart does". The second one simply guarantees that if you two somehow get locked into out-of-phase alternations (that is, "you cooperate, I defect" one month and "you defect, I cooperate" the next), you will not do better-in fact, you will do worse-than if you were cooperating with each other each month.

Well, what would be your best strategy? It can be shown quite easily that there is no universal answer to this question. That is, there is no strategy that is better than all other strategies under all circumstances. For consider the case where the other player is playing ALL D-the strategy of defecting each round. In that case, the best you can possibly do is to defect each time as well, including the first. On the other hand, suppose the other player is using the Massive Retaliatory Strike strategy, which means "I'll cooperate until you defect and thereafter I'll defect forever." Now if you defect on the very first move, then you'll get one T and all P's thereafter until one of you dies. But if you had waited to defect, you could have benefited from a relationship of mutual cooperation, amassing many R's beforehand. Clearly that bunch of R's will add up to more than the single T if the game goes on for more than a few moves. This means that against the ALL D strategy, ALL D is the best counterstrategy, whereas "Always cooperate unless you learn that you or the other player is just about to die, in which case defect" is the best counterstrategy against Massive Retaliatory Strike. This simple argument shows that how you should play depends on who you're playing.

The whole concept of the "quality" of a strategy takes on a decidedly more operational and empirical meaning if one imagines an ocean populated by dozens of little beings swimming around and playing Prisoner's Dilemma over and over with each other. Suppose that each time two such beings encounter each other, they recognize each other and

remember how previous encounters have gone. This enables each one to decide what it wishes to do this tine. Now if each organism is continually swimming around and bumping into the others, eventually, each one will have met even other one numerous times, and thus all strategies will have been given the opportunity to interact with each other. By "interact", what is meant here is certainly not that anyone knocks anyone else out of the ocean, as in an elimination tournament. The idea is simply that each organism gains zero or more points in each meeting, and if sufficient time is allowed to elapse, everybody will have met with everybody else about the same number of times. and now the only question is: Who has amassed the most points? Amassing points is truly the name of the game.

It doesn't matter if you have "beaten" anyone, in the sense of haying gained more Front interacting with them than they gained from interacting with you. That kind of "victory" is totally irrelevant here. What matters is not the number of "victories" rung tip by any individual. but the individual's total point count-a number that measures the individual's overall viability in this particular "sea" of many strategies. It sounds nearly paradoxical. but you could lose many-indeed, all-of your individual skirmishes, and yet still come out the overall winner.

As the image suggests very strongly, this whole situation is highly relevant to questions in evolutionary biology. Can totally selfish and unconscious organisms living in a common environment come to evolve reliable cooperative strategies Can cooperation emerge in a world of pure egoists? In a nutshell, can cooperation evolve out of noncooperation? If so, this has revolutionary import for the theory of evolution, for many of its critics have claimed that this was one place that it was hopelessly snagged.

Well, as it happens, it has now been demonstrated rigorously and definitively that such cooperation can emerge, and it was done through a computer tournament conducted by political scientist Robert Axelrod of the Political Science Department and the Institute for Public Policy Studies of the University of Michigan in Ann Arbor. More accurately, Axelrod first studied the ways that cooperation evolved by means of a computer tournament, and when general trends emerged, he was able to spot the underlying principles and prove theorems that established the facts and conditions of cooperation's rise from nowhere. Axelrod has written a fascinating and remarkably thought-provoking book on his findings, called The Evolution of Cooperation, published in 1984 by Basic Books, Inc. (Quoted sections below are taken front an early draft of that book.) Furthermore, he and evolutionary biologist William 1), Hamilton have worked out and published many of the implications of these discoveries for evolutionary theory. Their work has won much acclaim-including the 198 1 Newcomb Cleveland prize, a prize awarded annually by the American Association for the Advancement of Science for “an outstanding paper published in Science”.

There are really three aspects of the question "Can cooperation emerge in a world of egoists?" The first is: How can it get started at all? the second is: Can cooperative strategies survive better than their non-cooperative rivals? The third one is: Which cooperative strategies will do the best, and how will they come to predominate?

To make these issues vivid, let me describe Axelrod's tournament and its somewhat astonishing results. In 1979, Axelrod sent out invitations to a number of professional game theorists, including people who had published articles on the Prisoner's Dilemma. telling them that he wished to pit many strategies against one another in a round-robin Prisoner's Dilemma tournament, with the overall goal being to amass as many points as possible. He asked for strategies to be encoded as computer programs that could respond to the 'C' or 'D' of another player, taking into account the remembered history of previous interactions with that same player. A program should always reply with a 'C' or a 'I)', of course, but its choice need not be deterministic. That is. consultation of a random-number generator was allowed at any point in a strategy.

Fourteen entries were submitted to Axelrod, and he introduced into the field one more program called RANDOM, which in effect flipped a coin (computationally simulated, to be sure) each move, cooperating if heads came up, defecting otherwise. The field was a rather variegated one, consisting of programs ranging from as few as four lines to as many as 77 lines (of Basic). Every program was made to engage each other program (and a clone of itself) 200 times. No program was penalized for running slowly. The tournament was actually run five times in a row, so that ,pseudo-effects caused by statistical fluctuations in the random-number generator would be smoothed out by averaging.

The program that won was submitted by the old Prisoner's Dilemma hand, Anatol Rapoport, a psychologist and philosopher from the University of Toronto. His was the shortest of all submitted programs, and is called TIT FOR TAT'.. TIT FOR TAT uses a very simple tactic:

Cooperate on move 1;
Thereafter, do whatever the other player did the previous move.

That is all. It sounds outrageously simple. How in the world could such a program defeat the complex stratagems devised by other experts?

Well, Axelrod claims that the game theorists in general did not go far enough in their analysis. They looked "only two levels deep", when in fact they should have looked three levels deep to do better. What precisely does this mean? He takes a specific case to illustrate his point. Consider the entry called JOSS (submitted by Johann Joss, a mathematician from Zurich, Switzerland). JOSS´s strategy is very similar to TIT FOR TAT´s, in that it

begins by cooperating, always responds to defection by defecting and nearly always responds to cooperation by cooperating. The hitch is that JOSS uses a random-number generator to help it decide when to pull a "surprise defection" on the other player. JOSS is set up so that it has a 10 percent probability of defecting right after the other player has cooperated.

In playing TIT FOR TAT, JOSS will do fine until it tries to catch TIT FOR TIT off guard. When it defects, TIT FOR TAT retaliates with a single defection, while JOSS "innocently" goes hack to cooperating. Thus we have a "DC" pair. On the next move, the 'C' and 'D' will switch places since each program in essence echoes the other's latest move, and so it will go: CD, then DC, CD, DC, and so on. There may ensue a long reverberation set off by JOSS's D, but sooner or later, JOSS will randomly, throw in another unexpected D after a C from TIT FOR TAT. At this point, there will he a "DD" pair, and that determines the entire rest of the match. Both will defect forever, now. The "echo" effect resulting from JOSS's first attempt at exploitation and TIT FOR TAT's simple punitive act lead ultimately to complete distrust and lack of cooperation.

This may seem to imply that both strategies are at fault and will suffer for it at the hands of others, but in fact the one that suffers from it most is JOSS, since JOSS tries out the same trick on partner after partner, and in mans cases this leads to the same type of breakdown of trust, whereas TIT FOR TA T, never defecting first, will never be the initial cause of a breakdown of trust. Axelrod's technical term for a strategy that never defects before its opponent does is nice. TIT FOR TAT is a nice strategy,JOSS is not. Note that "nice" does not mean that a strategy never defects! TIT FOR TAT defects when provoked, but that is still considered being "nice".

Axelrod summarizes the first tournament this way:

> A major lesson of this tournament is the importance of minimizing echo effects in an environment of mutual power. A sophisticated analysis must go at least three levels deep. First is the direct effect of a choice. This is easy, since a defection always earns more than a cooperation. Second are the indirect effects, taking into account that the other side may or may not punish a defection. This much was certainly appreciated by many of the entrants. But third is the fact that in responding to the defections of the other side, one may be repeating or even amplifying one's own previous exploitative choice. Thus a single defection may be successful when analyzed for its direct effects, and perhaps even when its secondary effects are taken into account. But the real costs may be in the tertiary effects when one's own isolated defections turn into unending mutual recriminations. Without their realizing it, many of these rules actually wound up punishing themselves. With the other player serving as a mechanism to delay the self-punishment by a few moves, this aspect of self-punishment was not perceived by the decision rules ....
>
> The analysis of the tournament results indicates that there is a lot to be learned about coping in an environment of mutual power. Even expert strategists from political science, sociology, economics, psychology, and mathematics made the systematic errors of being too competitive for their own

good. not forgiving enough, and too pessimistic about the responsiveness of the other side.

Axelrod not only analyzed the first tournament, he even performed a number of "subjunctive replays" of it, that is, replays with different sets of entries. He found, for instance, that the strategy called TIT FOR TWO T.ATS, which tolerates two defections before getting mad (but still only- strikes back once), would have won, had it been in the line-up. Likewise, two other strategies he discovered, one called REVISED DOWNING and one called LOOK-AHEAD, would have come in first had they been in the tournament.

In summary, the lesson of the first tournament seems to have been that it is important to be nice ("don't be the first to defect") and forgiving, ("don't hold a grudge once you've vented your anger"). TIT FOR TAT possesses both these qualities, quite obviously.

* * *

After this careful analysis, Axelrod felt that significant lessons had been unearthed, and he felt convinced that more sophisticated strategies could be concocted, based on the new information. Therefore he decided too hold a second, larger computer tournament. For this tournament, he not only invited all the participants in the first round, but also advertised in computer hobbyist magazines, hoping to attract people who were addicted to programming and who would be willing to devote a good deal of time to working out and perfecting their strategies. To each person who entered, Axelrod sent a full and detailed analysis of the first tournament, along with a discussion of the "subjunctive replays" and the strategies that would have won. He described the strategic concepts of "niceness" and "forgiveness" that seemed to capture the lessons of the tournament so well, as well as strategic pitfalls to avoid. Naturally, each entrant realized that all the other entrants had received the same mailing, so that everyone knew that everyone knew that everyone knew that ...

There was a large response to Axelrod's call for entries. Entries were received from six countries, from people of all ages, and from eight different academic disciplines. Anatol Rapoport entered again, resubmitting TIT FOR TAT (and was the only one to do so, even though it was explicitly stated that anyone could enter any program written by anybody). A ten-year-old entered, as did one of the world's experts on game theory and evolution, John Maynard Smith, professor of biology at the University of Sussex in England, who submitted TIT FOR TWO TATS. Two people separately submitted REVISED DOWNING.

Altogether, 62 entries were received, and generally speaking, they were of a considerably higher degree of sophistication than those in the first tournament. The shortest was again TIT FOR TAT, and the longest was a program from New Zealand, consisting of 152 lines of Fortran. Once again,

RANDOM was added to the field, and with a flourish and a final carriage return, the horses were off' Several hours of computer time later, the results came in.

The outcome was nothing short of stunning: TIT FOR 7:I T, the simplest program submitted, won again. What's more, the two programs submitted that had won the subjunctive replays of the first tournament now turned up way down in the list: TIT FOR TWO TATS came in 24th, and REVISED DOWNING ended up buried in the bottom half of the field.

This may seem horribly nonintuitive, but remember that a program's success depends entirely on the environment in which it is swimming. There is no single "best strategy" for all environments, so that winning in one tournament is no guarantee of success in another. TIT FOR TAT has the advantage of being able to "get along well" with a great variety of strategies, while other programs are more limited in their ability to evoke cooperation. Axelrod puts it this way:

> What seems to have happened is an interesting interaction between people who drew one lesson and people who drew another lesson from the first round. Lesson One was "Be nice and forgiving." Lesson Two was more exploitative: "If others are going to be nice and forgiving, it pays to try to take advantage of them." The people who drew Lesson One suffered in the second round from those who drew Lesson Two.

Thus the majority of participants in the second tournament really had not grasped the central lesson of the first tournament: the importance of being willing to initiate and reciprocate cooperation. Axelrod feels so strongly about this that he is reluctant to call two strategies playing against each other "opponents"; in his book he always uses neutral terms such as "strategies" or "players". He even does not like saying they are playing against each other, preferring "with". In this article, I have tried to follow his usage, with occasional departures. One very striking fact about the second tournament is the success of "nice" rules: of the top fifteen finishers. only one (placing eighth) was not nice. Amusingly, a sort of mirror image held: of the bottom fifteen finishers, only one was nice!

Several non-nice strategies featured rather tricky probes of the opponent (sorry!), sounding it out to see how much it "minded" being defected against. Although this kind of probing by a program might fool occasional opponents, more often than not it backfired, causing severe breakdowns of trust. Altogether, it turned out to be very costly to try to use defections to "flush out" the other player's weak spots. It turned out to be more profitable to have a policy of cooperation as often as possible, together with a willingness to retaliate swiftly against any attempted undercutting. Note, however, that strategies featuring massive retaliation were less successful than TIT FOR TAT with its more gentle policy of restrained retaliation.

Forgiveness is the key here, for it helps to restore the proverbial "atmosphere of mutual cooperation" (to use the phrase of international diplomacy) after a small skirmish.

"Be nice and forgiving" was in essence the overall lesson of the first tournament. Apparently, though, many people just couldn't get themselves to believe it, and were convinced that with cleverer trickery and scheming, they could win the day. It took the second tournament to prove them dead wrong. And out of the second tournament, a third key strategic concept emerged: that of provocability-the notion that one should "get mad" quickly at defectors. and retaliate. Thus a more general lesson is: "Be nice, provocable, and forgiving."

* * *

Strategies that do well in a wide variety of environments are called by Axelrod robust. and it seems that ones with "good personality traits"-that is, nice. provocable. and forgiving strategies-are sure to be robust. TIT FOR TAT is by no means the only possible strategy with these traits, but it is the canonical example of such a strategy, and it is astonishingly robust.

Perhaps the most vivid demonstrations of TIT FOR TAT's robustness were provided by various subjunctive replays of the second tournament. The principle behind any replay involving a different environment is quite simple. From the actual playing of the tournament, you have a 63X63 matrix documenting how well each program did against each other program. Now, the effective "population" of a program in the environment can be manipulated mathematically by attaching a weight factor to all that program's interactions, then just retotaling all the columns. This way you can get subjunctive instant replays without having to rerun the tournament.

This simple observation means that the results of a huge number of potential subjunctive tournaments are concealed in, but potentially extractable from, the 63X63 matrix of program-z s.-program totals. For instance, Axelrod discovered, using statistical analysis, that there were essentially six classes of strategies in the second tournament. For each of these classes, he conducted a subjunctive instant replay of the, tournament by quintupling the importance (the weight factor) of that class alone, thus artificially inflating a certain strategic style's population in the environment. When the scores were retotaled, TIT FOR TAT emerged victorious in five out of six of those hypothetical tournaments, and in the sixth it placed second.

Undoubtedly the most significant and ingenious type of subjunctive replay that Axelrod tried was the ecological tournament. Such a tournament consists not merely of a single subjunctive replay, but of -a whole cascade of hypothetical replays, each one's environment determined by the results of the previous one. In particular, if you take a program's score in a tournament as a measure of its "fitness", and if you interpret "fitness" as meaning "number of progeny in the next generation", and finally, if you let

"next generation" mean "next tournament", then what you get is that each tournament's results determine the environment of the next one-arid in particular, successful programs become more copious in the next tournament. This type of iterated tournament is called "ecological" because it simulates ecological adaptation (the shifting of a fixed set of species' populations according to their mutually defined and dynamically developing environment), as contrasted with evolution via mutation (where new species can come into existence).

As one carries an ecological tournament through generation after generation, the environment gradually changes. In a paraphrase of how Axelrod puts it, here is what happens. At the very beginning, poor programs and good programs alike are equally represented. As time passes, the poorer ones begin to drop out while the good ones flourish. But the tank order of the good ones may now change. because their "goodness" is no longer being measured against the same field of competitors as initially. Thus success breeds ever more success-but only provided that the success derives from interaction with other similarly successful programs. If, by contrast, some program's success is due mostly to its ability to milk "dumber" programs for all they're worth, then as those programs are gradually squeezed out of the picture, the exploiter's base of support will be eroded and it will suffer a similar fate.

A concrete example of ecological extinction is provided by HARRINGTON. the only non-nice program among the top fifteen finishers in the second tournament. In the first 200 generations of the ecological tournament, while TIT FOR TAT and other successful nice programs were gradually increasing their percentage of the population, HARRINGTON ' too was increasing its percentage. This was a direct result of HARRINGTON´S exploitative strategy. However, by the 200th generation, things began to take a noticeable turn. Weaker programs were beginning to go extinct, which meant fewer and fewer dupes for HARRINGTON to profit from. Soon the trend became apparent: HARRINGTON could not keep up with its nice rivals. By the 1,000th generation, HARRINGTON was as extinct as the dodos it had exploited. Axelrod summarizes:

> Doing well with rules that do not score well themselves is eventually a self'-defeating process. Not being nice may look promising at first, but in the long run it can destroy the very environment it needs for its own success.

Needless to say, TIT FOR TAT fared spectacularly well in the ecological tournament, increasing its lead ever more. After 1,000 generations, not only was TIT FOR TAT ahead, but its rate of growth was greater than that of any other program. This is an almost unbelievable success story,, all the more so because of the absurd simplicity of the "hero". One amusing aspect of it is that 7'IT FOR TAT did not defeat a single one of its rivals in their encounters. This is not a quirk; it is in the nature of TIT FOR TAT. TIT FOR TAT simply cannot defeat anyone; the best it can achieve is a tie, and often it loses (though not by much).

Axelrod makes this point very clear:

> TIT FOR TAT won the tournament, not by beating the other placer, but by eliciting behavior from the other player which allowed both to do well. TIT FOR TAT was so consistent at eliciting mutually rewarding outcomes that it attained a higher overall score than any other strategy in the tournament.
>
> So in a non-zero-sum world you do not have to do better than the other player to do well for yourself. This is especially true when you are interacting with many different players. Letting each of them do the same or a little better than you is fine, as long as you tend to do well yourself. There is no point in being envious of the success of the other player, since in an iterated Prisoner's Dilemma of long duration the other's success is virtually a prerequisite of your doing well for yourself.

He gives examples from everyday life in which this principle holds. Here is one:

> A firm that buys from a supplier can expect that a successful relationship will earn profit for the supplier as well as the buyer. There is no point in being envious of the supplier's profit. Any attempt to reduce it through an uncooperative practice, such as by not paying your bills on time, ,will only encourage the supplier to take retaliatory action. Retaliatory action could take many forms, often without being explicitly labeled as punishment. It could be less prompt deliveries, lower quality control, less forthcoming attitudes on volume discounts, or less timely news of anticipated market conditions. The retaliation could make the envy quite expensive. Instead of worrying about the relative profits of the seller, the buyer should worry about whether another buying strategy would be better.

Like a business partner who never cheats anyone, TIT FOR TAT never beats anyone-yet both do very well for themselves.

One idea that is amazingly counterintuitive at first in the Prisoner's Dilemma is that the best possible strategy to follow is ALL D if the other player is unresponsive. It might seem that some form of random strategy might do better, but that is completely wrong. If I have laid out all my moves in advance, then playing TIT FOR TAT will do you no good, nor will flipping a coin. You should simply defect every move. It matters not what pattern I have chosen. Only if I can be influenced by your play will it ever do you any good to cooperate.

Fortunately, in an environment where there are programs that cooperate (and whose cooperation is based on reciprocity), being unresponsive is a very poor strategy, which in turn means that ALL D is a very poor, strategy. The single unresponsive competitor in the second tournament was RAND0m and it finished next to last. The last-place finishers strategy was responsive, but its behavior was so inscrutable that it looked unresponsive

And in a more recent computer tournament conducted by Marek Lugowsl, and myself in the Computer Science Department at Indiana L-university three ALL-D's came in at the very bottom (out of 53), with a couple of RA.VDO.11's giving them a tough fight for the honor.

One way to explain TIT FOR TIT's success is simply to say that it elicits cooperation, via friendly persuasion. Axelrod spells this out as follows:

Part of its success might be that other rules anticipate its presence and are designed to do well with it. Doing well with TIT FOR TAT requires cooperating with it, and this in turn helps TIT FOR TIT Even rules that were designed to see what then could get away with quickly apologize to TIT FOR F-IT. Am rule that tries to take advantage of TIT FOR T.AT will simple hurt itself. TIT FOR TAT benefits from its own nonexploitability because three conditions are satisfied:

1. The possibility of encountering TIT FOR TAT is salient;
2. Once encountered. TIT FOR TAT is easy to recognize: and
3. Once recognized, TIT FOR TAT 's nonexploitability is easy to appreciate

This brings out a fourth "personality trait" (in addition to niceness. provocability, and forgiveness) that may play an important role in success: recognizability. or straightforwardness. Axelrod chooses to call this trait clarity, and argues for it with clarity:

Too much complexity can appear to be total chaos. If you are using a strategy that appears random, then you also appear unresponsive to the other player. If you are unresponsive, then the other player has no incentive to cooperate with you. So being so complex as to be incomprehensible is yen dangerous.

How rife this is with morals for social and political behavior! It is rich food for thought.

* * *

Anatol Rapoport cautions against overstating the advantages of TIT FOR TIT; in particular, he believes that TIT FOR LIT is too harshly retaliatory on occasion. It can also be persuasively argued that TIT FOR TAT is too lenient on other occasions. Certainly there is no evidence that TI T FOR TIT is the ultimate or best possible strategy. Indeed, as has been emphasized repeatedly, the very concept of "best possible" is incoherent, since all depends on environment. In the tournament at Indiana University mentioned earlier, several TIT-FOR-TAT- like strategies did better than pure TIT FOR LIT did. They all shared, however, the three critical "character traits" whose desirability had been so clearly delineated by Axelrod's prior analysis of the important properties of TIT FOR TAT They were simple a little better than TIT FOR Liz' at detecting nonresponsiveness, and when they were convinced the other player was unresponsive, they switched over to an ALL-D mode.

In his book, Axelrod takes pains to spell out the answers to three fundamental questions concerning the temporal evolution of cooperation in a world of raw egoism. The first concerns initial viability: How can cooperation get started in a world of unconditional defection-a "primordial sea" swarming with unresponsive .ILL-D creatures? The answer (whose proof I omit here) is that invasion by small clusters of conditionally cooperating organisms, even if they form a tiny minority, is enough to give cooperation a toehold. One cooperator alone will die, but small clusters of cooperators can arrive (via mutation or migration, say) and propagate even in a hostile environment, provided they are defensive like TIT FOR TAT. Complete pacifists-Quaker-like programs-will not survive, however, in this harsh environment.

The second fundamental question concerns robustness: What type of strategy does well in unpredictable and shifting environments? We have already seen that the answer to this question is: Any strategy possessing the four fundamental "personality traits" of niceness, provocability, forgiveness, and clarity. This means that such strategies, once established. will tend to flourish, especially in an ecologically evolving world.

The final question concerns stability: Can cooperation protect itself from invasion? Axelrod proved that it can indeed. In fact, there is a gratifying asymmetry to his findings: Although a world of "meanies" (beings using the inflexible ALL-D strategy) is penetrable by cooperators in clusters, a world of cooperators is not penetrable by meanies, even if they arrive in clusters of any size. Once cooperation has established itself, it is permanent. As Axelrod puts it, "The gear wheels of social evolution have a ratchet."

The term "social" here does not mean that these results necessarily apply only to higher animals that can think. Clearly, four-line computer programs do not think-and yet, it is in a world of just such "organisms" that cooperation has been shown to evolve. The only "cognitive abilities" needed by TIT FOR T.-IT are: (1) recognition of previous partners, and (2) memory of what happened last time with this partner. Even bacteria can do this, by interacting with only one other organism (so that recognition is automatic) and by responding only to the most recent action of their "partner" (so that memory requirements are minimal). The point is that the entities involved can be on the scale of bacteria, small animals, large animals, or nations. There is no need for "reflective rationality"; indeed, TIT FOR LIT could be called "reflexive" (iii the sense of being as simple as a knee-jerk reflex) rather than "reflective".

For people who think that moral behavior toward others can emerge only when there is imposed some totally external and horrendous threat (say, of the fire-and-brimstone sort) or soothing promise of heavenly reward (such as eternal salvation), the results of this research must give pause for thought. In one sentence, Axelrod captures the whole idea: Mutual cooperation can

emerge in a world of egoists without central control, by starting with a cluster ol individuals who rely on reciprocity.

There are so man situations in the world today where these ideas seem of extreme relevance-indeed, urgency-that it is very tempting to draw morals all over the place. In the later chapters of his book, Axelrod offers advice about how to promote cooperation in human affairs, and at the very end the political scientist in him cautiously ventures some broad conclusions concerning global issues, which are a fitting way for me to conclude as well

> Today, the most important problems facing humanity are in the arena of international relations where independent, egoistic nations face each other in a state of near anarchy. Many of these problems take the form of an iterated Prisoner's Dilemma. Examples can include arms races, nuclear proliferation. crisis bargaining, and military escalation. Of course, a realistic understanding of these problems would have to take into account mane factors not incorporated into the simple Prisoner's Dilemma formulation, such as ideology. bureaucratic politics, commitments, coalitions, mediation, and leadership. Nevertheless, we can use all the insights we can get.
>
> Robert Gilpin [in his book War and Change in World Politics] points out that from the ancient Greeks to contemporary scholarship all political theory addresses one fundamental question: "How can the human race. whether for selfish or more cosmopolitan ends, understand and control the seemingly blind forces of history?" In the contemporary world this question has become especially acute because of the development of nuclear weapons.
>
> The advice given in this book to players of the Prisoner's Dilemma might also serve as good advice to national leaders as well: Don't be envious, don't he the first to defect, reciprocate both cooperation and defection, and don't be too clever. Likewise, the techniques discussed in this book for promoting cooperation in the Prisoner's Dilemma might also be useful in promoting cooperation in international politics.
>
> The core of the problem is that trial-and-error learning is slow and painful. The conditions may all be favorable for long-run developments, but we may not have the time to wait for blind processes to move us slowly towards mutually rewarding strategies based upon reciprocity. Perhaps if we understand the process better, we can use our foresight to speed up the evolution of cooperation.

Post Scriptum.

In the course of writing this column and thinking the ideas through. I was forced to confront over and over again the paradox that the Prisoner's Dilemma presents. I found that I simply could not accept the seemingly flawless logical conclusion that says that a rational player in a noniterated situation will always defect. In turning this over in my mind and trying to articulate my objections clearly, I found myself inventing variation after

variation after variation on the basic situation. I would like to describe just a few here.

A version of the dealer-and-buyer scenario involving bags exchanged in a forest actually occurs in a more familiar context. Suppose I take my car in to get the oil changed. I know little about auto mechanics, so when I come in to pick it up. I really have no way to verify if they've done the job. For all I know, it's been sitting untouched in their parking lot all day, and as I drive off they may be snickering behind my back. On the other hand, maybe I've got the last laugh, for how do they know if that check I gave them will bounce?

This is a perfect example of how either of us could defect, but because the situation is iterated, neither of us is likely to do so. On the other hand, suppose I'm on my way across the country and have some radiator trouble near Gillette. Wyoming, and stop in town to get my radiator repaired there. There is a decent chance now that one party or the other will attempt to defect, because this kind of situation is not an iterated one. I'll probably never again need the services of this garage, and they'll never get another check from me. In the most crude sense, then, it's not in my interest to give them a good check, nor is it in theirs to fix my car. But do I really defect? Do I give out bad checks? No. Why not?

Consider this related situation. Late at night, I bang into someone's car in a deserted parking lot. It's apparent to me that nobody witnessed the incident. I have the choice of leaving a note, telling the owner who's to blame, or scurrying off scot-free. Which do I do? Similarly, suppose I-have given a lecture in a classroom in a university I am visiting for one day, and have covered the board with chalk. Do I take the trouble of erasing the board so that whoever comes in the next morning won't have to go to that trouble? Or do I just leave it?

I was recently waiting to board an airplane when a voice announced: "Passengers holding seats in rows 24 to 36 may now board." Well, my seat was in row 4, so I waited. A few minutes later, the voice said that passengers in rows 18 to 36 were free to board. A group of people got up and went in. Then after a couple of minutes, rows 10 to 36 were told they could board. A dozen people or so remained in the waiting area. For a while, we were all patient, waiting for the final announcement allowing us to board, but after about five minutes, people started fidgeting a bit and edging up toward the gate. Then, after another two or three minutes, a couple of people just went right on. And then the rest of us wondered, "Should we get on, too? Will we be left behind?" For most of the people, the answer was obvious: they rushed to board. And once they had boarded, then the rest of us felt kind of like suckers, and we just got on too. In effect, there was a stampede that converted cooperators into defectors. Even the people who triggered the stampede had originally been cooperating, but after a while the temptation

got to be too great, and they broke down. At that point, some sort of phase transition, or collective shift, took place, and the stable state of patient cooperation collapsed into a chaotic scrambling for places. Actually, it wasn't that bad, and there was a good reason for the relatively polite way we did board, defectors though we were: we all had seat assignments, so it didn't matter who got on first. But imagine if the earliest defectors were sure to get the best remaining seats! The contemporary aphorist .Ashleigh Brilliant has found just the right *bons mots* to describe this sort of dilemma:

Should I abide by the rules until they're changed, or help speed tht change by breaking them?

Better start rushing before the rush begins!

In pondering the Prisoner's Dilemma, I could not help but be reminded of horrible scenarios in Nazi concentration camps, where large herds of unarmed people would be led to their deaths by small herds of armed people. It seems that a stampede by the masses could quickly have overcome a small number of guards, at least in certain critical narrow passageways here and there. The trouble is, it would require certain death on the part of a few ultra-cooperators. in exchange for the liberation of a large number of other people. Generally speaking, individuals are not willing to perform such an exchange. Nobody wants to be in the front lines of a protest demonstration facing troops with machine guns. Everyone wants to be in the rear. But not everyone can be in the rear! If nobody is willing to be in the front lines, then there will be no front lines, and consequently no demonstration at all.

Driving a car has a certain primitive quality to it that brings out the animal in us all, and probably that's why it confronts us with Prisoner's-Dilemma like situations so often-more often than any other activity I can think of. How about those annoying drivers who, when there's a long line at a freeway exit, zoom by all the politely lined-up cars and then butt in at the very last moment, getting off 50 cars ahead of you? Are you angry t such people. or do you do it too? Or, worse-do you do it and yet resent others who have such gall?

I have been struck by the relative savagery of the driving environment in the Boston area. I know of no other city in which people are so willing to take the law into their own hands, and to create complete anarchy. There seems to be less respect for such things as red lights. stop signs, lines in the street, speed limits, other people's cars, and so forth, than in any other city or state, or country that I have ever driven in. This incessant "me-first" attitude seems to be a vicious, self-reinforcing circle. Since there are so many people who do whatever they want, nobody can afford to be polite and let other

people in ahead of them (say), for then they will be taken advantage of repeatedly and will wind up losing totally. You simply must assert yourself in many situations, and that means you must defect. Of course, just one defection does not an .ALL-D player make. In fact, a retaliatory defection is just good old TIT-FOR-TAT playing. However, very often in Boston driving, there is no way you can get back at a nasty driver who cuts in front of you and then takes off screeching around the corner. That person is gone forever. You can take out your frustrations only on the rest of the people near you, who are not to blame for that driver. You can cut in ahead of them. Does this do any good? That is, does it teach anybody a lesson? Obviously it will teach them only that it pays to defect. And thus the spiral starts.

Is there any way to put a halt to the descending spiral, the vortex towards oblivion? Is there any point at which the people of Boston will collectively come to realize that it has gotten so bad that they will all suddenly "flip" and begin to cooperate in situations where they formerly would have defected? Can there be a stampede toward cooperation, just as there can be a stampede toward defection?

Clearly, if large numbers of people were to start driving much less aggressively and nastily, everybody would benefit. Huge snarls would unsnarl-in fact would never form. Traffic would flow smoothly and regularly. The shoulders-those favorite illegal passing lanes for defectors-would be completely clear. So clear, in fact, that just think-you and I could make sensational progress by swerving onto an empty shoulder and passing everybody. Wheee! Isn't this fun? Aren't those other people suckers, staying in the slow lane and glaring at us? Say, how come other people are barging in on us? This is our lane. Oh, so that person in the yellow car wants to play dirty, eh? Okay. I'll show them what playing dirty's really like!

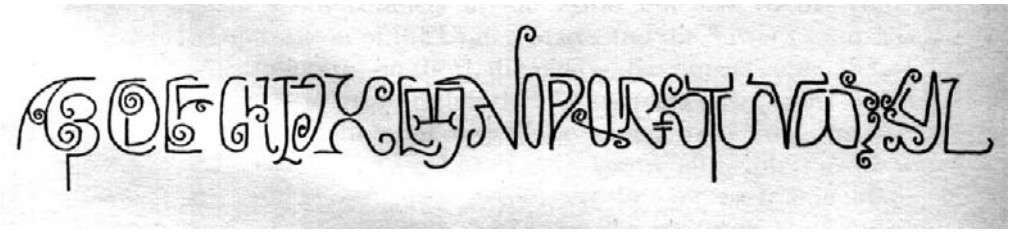
Sound familiar? Is there any solution to such terrible spirals? Sometimes I am very pessimistic on that subject. Anatol Rapoport and I exchanged letters concerning this matter, and he related a frightening anecdote. I quote from his letter:

> Do you know of the experiment performed by Martin Shubik, in which a dollar bill was auctioned off for S3.40? This was a consequence of a rule (the implications of which dawned on the subjects only when they were already hooked) specifying that while the highest bidder got the dollar, the second-highest bidder would also have to pay what he last bid. It thus became imperative to keep going, since the second-highest bidder (whoever he was at each stage) had progressively more to lose as the bids went up. Are Reagan and Andropov too stupid to see the point? ....
>
> I believe the "technological imperative" is driving our species to extinction. Ever more horrendous weapons must be produced, simply because it is possible to produce them. Eventually the\, must be used, to justify the insane waste. It thus becomes imperative to seal off' the "logic" of the paradigm based on "deterrence", "balance of power", and similar metaphors-to make it in-assailable.

> I don't think intelligence plans a part in the vicious cycle of the arms race. The rulers only think they make the decisions. If they were C players. they would not be where they are. If they started to play C while in office, they would he impeached. overthrown, or assassinated. Does this mean that D players are selected for? Possibly in the short run, but not on the time scale of evolution. H. sapiens is apparently not the last word, but for me, a homocentric, this is no consolation.

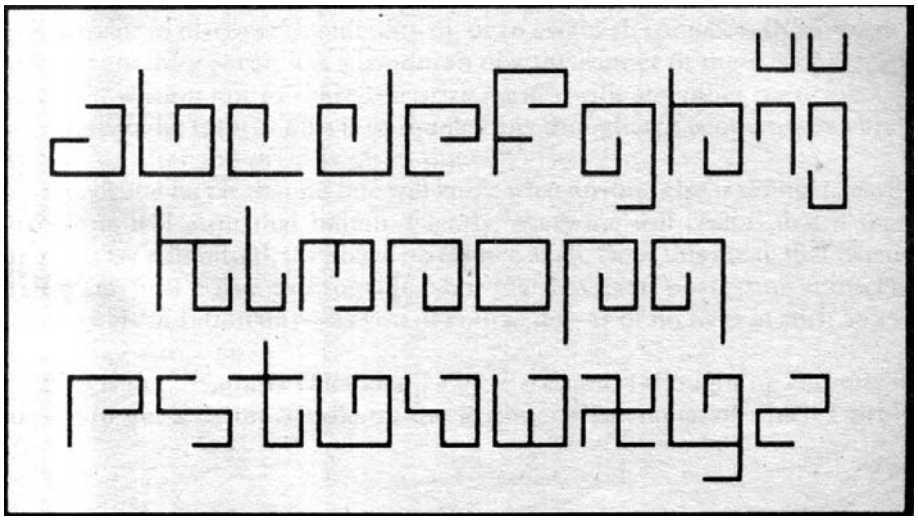
Pretty- sobering words from one of the leading rational thinkers of our era

Section VII:
Sanity and Survival



Section VII:
Sanity and Survival

In the four chapters of this concluding section, themes of the previous section are carried further and brought into contact with common social dilemmas and, eventually, the current world situation. On a small scale, we are constantly faced with dilemmas like the Prisoner's Dilemma, where personal greed conflicts with social gain. For any two persons, the dilemma is virtually identical. What would be sane behavior in such situations? For true sanity, the key element is that each individual must be able to recognize both that the dilemma is symmetric and that the other individuals facing it are equally able. Such individuals-individuals who will cooperate with one another despite all temptations toward crude egoism-are more than just rational; they are superrational, or for short, sane. But there are dilemmas and "egos" on a suprahuman level as well. We live in a world filled with opposing belief systems so similar as to be nearly interchangeable, yet whose adherents are blind to that symmetry. This description applies not only to myriad small ,conflicts in the world but also to the colossally blockheaded opposition of the United States and the Soviet Union. Yet the recognition of symmetry-in short, the sanity-has not yet come. In fact, the insanity seems only to grow, rather than be supplanted by sanity. What has an intelligent species like our own done to get itself into this horrible dilemma? What can it do to get itself out? Are we all helpless as we watch this spectacle unfold, or does the answer lie, for each one of us, in recognition of our own typicality, and in small steps taken on an individual level toward sanity?



30

Dilemmas for Superrational Thinkers, Leading Up to a Luring Lottery

June, 1983

AND then one fine day, out of the blue, you get a letter from S. N. Platonia, well-known Oklahoma oil trillionaire, mentioning that twenty leading rational thinkers have been selected to participate in a little game. "You are one of them!" it says. "Each of you has a chance at winning one billion dollars, put up by the Platonia Institute for the Study of Human Irrationality. Here's how. If you wish, you may send a telegram with just your name on it to the Platonia Institute in downtown Frogville, Oklahoma (pop. 2). You may reverse the charges. If you reply within 48 hours, the billion is yours-unless there are two or more replies, in which case the prize is awarded to no one. And if no one replies, nothing will be awarded to anyone."

You have no way of knowing who the other nineteen participants are; indeed, in its letter, the Platonia Institute states that the entire offer will be rescinded if it is detected that any attempt whatsoever has been made by any participant to discover the identity of, or to establish contact with, any other participant. Moreover, it is a condition that the winner (if there is one) must agree in writing not to share the prize money with any other participant at any time in the future. This is to squelch any thoughts of cooperation, either before or after the prize is given out.

The brutal fact is that no one will know what anyone else is doing. Clearly, everyone will want that billion. Clearly, everyone will realize that if their name is not submitted, they have no chance at all. Does this mean that twenty telegrams will arrive in Frogville, showing that even possessing transcendent levels of rationality-as you of course do-is of no help in such an excruciating situation?

This is the "Platonia Dilemma", a little scenario I thought up recently in trying to get a better handle on the Prisoner's Dilemma, of which I wrote

last month. The Prisoner's Dilemma can be formulated in terms resembling this dilemma, as follows. Imagine that you receive a letter from the Platonia Institute telling you that you and just one other anonymous leading rational thinker have been selected for a modest cash giveaway. As before, both of you are requested to reply by telegram within 48 hours to the Platonia Institute, charges reversed. Your telegram is to contain, aside from your name, just the word "cooperate" or the word "defect". If two "cooperate"s are received, both of you will get $3. If two "defect"s are received, you both will get $1. If one of each is received, then the cooperator gets nothing and the defector gets $5.

What choice would you make? It would be nice if you both cooperated, so you'd each get $3, but doesn't it seem a little unlikely? After all, who wants to get suckered by a nasty, low-down, rotten defector who gets $5 for being sneaky? Certainly not you! So you'd probably decide not to cooperate.

It seems a regrettable but necessary choice. Of course, both of you, reasoning alike, come to the same conclusion. So you'll both defect, and that way get a mere dollar apiece. And yet-if you'd just both been willing to risk a bit, you could have gotten $3 apiece. What a pity!

* * *

It was my discomfort with this seemingly logical analysis of the "one-round Prisoner's Dilemma" that led me to formulate the following letter, which I sent out to twenty friends after having cleared it with *Scientific American*

Dear X:

I am sending this letter out via Special Delivery to twenty of 'you' (namely, various friends of mine around the country). I am proposing to all of you a one-round Prisoner's Dilemma game, the payoffs to be monetary (provided by Scientific American). It's very simple. Here is how it goes.

Each of you is to give me a single letter: 'C' or 'D', standing for 'cooperate' or 'defect'. This will be used as your move in a Prisoner's Dilemma with each of the nineteen other players. Thepayoff matrix I am using for the Prisoner's Dilemma is given in the diagram [see Figure 29-1c].

Thus if everyone sends in 'C', everyone will get $57, while if everyone sends in 'D', everyone will get $19. You can't lose! And of course, anyone who sends in 'D' will get at least as much as everyone else will. If, for example, I 1 people send in 'C' and 9 send in 'D', then the 11 C-ers will get $3 apiece from each of the other C-ers (making $30), and zero from the D-ers. So C-ers will get $30 each. The D-ers, by contrast, will pick up $5 apiece from each of the C-ers, making $55, and $1 from each of the other D-ers, making $8, for a grand total of $63. No matter what the distribution is, D-ers always do better than C-ers. Of course, the more C-ers there are, the better everyone will do!

By the way, I should make it clear that in making your choice, you should not aim to be the winner, but simply to get as much money for yourself as possible. Thus you should be happier to get $30 (say, as a result of saying 'C' along with 10 others, even though the 9 D-sayers get more than you) than to get $19 (by

saying 'D' along with everybody else, so nobody `beats' you). Furthermore, you are not supposed to think that at some subsequent time you will meet with and be able to share the goods with your co-participants. You are not aiming at maximizing the total number of dollars Scientific American shells out, only at maximizing the number that come to you!

Of course, your hope is to be the unique defector, thus really cleaning up: with 19 C-ers, you'll get $95 and they'll each get 18 times $3, namely $541 But why am I doing the multiplication or any of this figuring for you? You're very bright. So are all of you! All about equally bright, I'd say, in fact. So all you need to do is tell me your choice. I want all answers by telephone (call collect, please) the day you receive this letter.

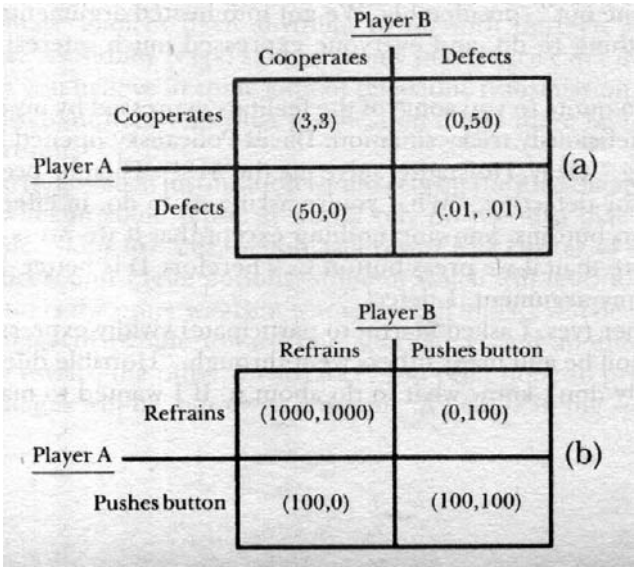
It is to be understood (it almost goes without saying, but not quite) that you are not to try to get in touch with and consult with others who you guess have been asked to participate. In fact, please consult with no one at all. The purpose is to see what people will do on their own, in isolation. Finally, I would very much appreciate a short statement to go along with your choice, telling me why you made this particular choice.

Yours....

P. S. -By the way, it may be helpful for you to imagine a related situation, the same as the present one except that you are told that all the other players have already submitted their choice (say, a week ago), and so you are the last. Now what do you do? Do you submit 'D', knowing full well that their answers are already committed to paper? Now suppose that, immediately after having submitted your 'D' (or your 'C') in that circumstance, you are informed that, in fact, the others really haven't submitted their answers yet, but that they are all doing it today. Would you retract your answer? Or what if you knew (or at least were told) that you were the first person being asked for an answer?

And-one last thing to ponder-what would you do if the payoff matrix looked as shown in Figure 30-la ?

FIGURE 30-1. *In (a), a modification of Figure 29-1(c). Here, the incentive to defect seems considerably stronger. In (b), the payoff matrix for a Wolf s-Dilemma situation involving just two participants. Compare it to that in Figure 29-1(c).*



* * *

I wish to stress that this situation is not an iterated Prisoner's Dilemma (discussed in last month's column). It is a one-shot, multi-person Prisoner's Dilemma. There is no possibility of learning, over time, anything about how the others are inclined to play. Therefore all lessons described last month are inapplicable here, since they depend on the situation's being iterated. All that each recipient of my letter could go on was the thought, "There are nineteen people out there, somewhat like me, all in the same boat, all grappling with the same issues as I am." In other words, there was nothing to rely on except pure reason.

I had much fun preparing this letter, deciding who to send it out to, anticipating the responses, and then receiving them. It was amusing to me, for instance, to send Special Delivery letters to two friends I was seeing every day, without forewarning them. It was also amusing to send identical letters to a wife and husband at the same address.

Before I reveal the results, I invite you to think how you would play in such a contest. I would particularly like you to take seriously the assertion "everyone is very bright". In fact, let me expand on that idea, since I felt that people perhaps did not really understand what I meant by it. Thus please consider the letter to contain the following clarifying paragraph:

> All of you are very rational people. Therefore, I hardly need to tell you that you are to make what you consider to be your maximally rational choice. In particular, feelings of morality, guilt, vague malaise, and so on, are to be disregarded. Reasoning alone (of course including reasoning about the others' reasoning) should be the basis of your decision. And please always remember that everyone is being told this (including this!)!

I was hoping for-and expecting-a particular outcome to this experiment. As I received the replies by phone over the next several days, I jotted down notes so that I had a record of what impelled various people to choose as they did. The result was not what I had expected-in fact, my friends "faked me out" considerably. We got into heated arguments about the "rational" thing to do, and everyone expressed much interest in the whole question.

I would like to quote to you some of the feelings expressed by my friends caught in this deliciously tricky situation. David Policansky opened his call tersely by saying, "Okay, Hofstadter, give me the $19!" Then he presented this argument for defecting: "What you're asking us to do, in effect, is to press one of two buttons, knowing nothing except that if we press button D, we'll get more than if we press button C. Therefore D is better. That is the essence of my argument. I defect."

Martin Gardner (yes, I asked Martin to participate) vividly expressed the emotional turmoil he and many others went through. "Horrible dilemma", he said. "I really don't know what to do about it. If I wanted to maximize

my money, I would choose D and expect that others would also; to maximize my satisfactions, I'd choose C, and hope other people would do the same (by the Kantian imperative). I don't know, though, how one should behave rationally. You get into endless regresses: 'If they all do X, then I should do Y, but then they'll anticipate that and do Z, and so . . .' You get trapped in an endless whirlpool. It's like Newcomb's paradox." So saying, Martin defected, with a sigh of regret.

In a way echoing Martin's feelings of confusion, Chris Morgan said, "More by intuition than by anything else, I'm coming to the conclusion that there's no way to deal with the paradoxes inherent in this situation. So I've decided to flip a coin, because I can't anticipate what the others are going to do. I think-but can't know-that they're all going to negate each other." So, while on the phone, Chris flipped a coin and "chose" to cooperate.

Sidney Nagel was very displeased with his conclusion. He expressed great regret: "I actually couldn't sleep last night because I was thinking about it. I wanted to be a cooperator, but I couldn't find any way of justifying it. The way I figured it, what I do isn't going to affect what anybody else does. I might as well consider that everything else is already fixed, in which case the best I can do for myself is to play a D."

Bob Axelrod, whose work proves the superiority of cooperative strategies in the iterated Prisoner's Dilemma, saw no reason whatsoever to cooperate in a one-shot game, and defected without any compunctions.

Dorothy Denning was brief: "I figure, if I defect, then I always do at least as well as I would have if I had cooperated. So I defect." She was one of the people who faked me out. Her husband, Peter, cooperated. I had predicted the reverse.

* * *

By now, you have probably been counting. So far, I've mentioned five D's and two C's. Suppose you had been me, and you'd gotten roughly a third of the calls, and they were 5-2 in favor of defection. Would you dare to extrapolate these statistics to roughly 14-6? How in the world can seven individuals' choices have anything to do with thirteen other individuals' choices? As Sidney Nagel said, certainly one choice can't influence another (unless you believe in some kind of telepathic transmission, a possibility we shall discount here). So what justification might there be for extrapolating these results?

Clearly, any such justification would rely on the idea that people are "like" each other in some sense. It would rely on the idea that in complex and tricky decisions like this, people will resort to a cluster of reasons, images, prejudices, and vague notions, some of which will tend to push them one way, others the other way, but whose overall impact will be to push a certain percentage of people toward one alternative, and another percentage of people toward the other. In advance, you can't hope to predict what those percentages will be, but given a sample of people in the situation, you can

hope that their decisions will be "typical". Thus the notion that early returns running 5-2 in favor of defection can be extrapolated to a final result of 14-6 (or so) would be based on assuming that the seven people are acting "typically" for people confronted with these conflicting mental pressures.

The snag is that the mental pressures are not completely explicit; they are evoked by, but not totally spelled out by, the wording of the letter. Each person brings a unique set of images and associations to each word and concept, and it is the set of those images and associations that will collectively create, in that person's mind, a set of mental pressures like the set of pressures inside the earth in an earthquake zone. When people decide, you find out how all those pressures pushing in different directions add up, like a set of force vectors pushing in various directions and with strengths influenced by private or unmeasurable factors. The assumption that it is valid to extrapolate has to be based on the idea that everybody is alike inside, only with somewhat different weights attached to certain notions.

This way, each person's decision can be likened to a "geophysics experiment" whose goal is to predict where an earthquake will appear. You set up a model of the earth's crust and you put in data representing your best understanding of the internal pressures. You know that there unfortunately are large uncertainties in your knowledge, so you just have to choose what seem to be "reasonable" values for various variables. Therefore no single run of your simulation will have strong predictive power, but that's all right. You run it and you get a fault line telling you where the simulated earth shifts. Then you go back and choose other values in the ranges of those variables, and rerun the whole thing. If you do this repeatedly, eventually a pattern will emerge revealing where and how the earth is likely to shift and where it is rock-solid.

This kind of simulation depends on an essential principle of statistics: the idea that when you let variables take on a few sample random values in their ranges, the overall outcome determined by a cluster of such variables will start to emerge after a few trials and soon will give you an accurate model. You don't need to run your simulation millions of times to see valid trends emerging.

This is clearly the kind of assumption that TV networks make when they predict national election results on the basis of early returns from a few select towns in the East. Certainly they don't think that free will is any "freer" in the East than in the West-that whatever the East chooses to do, the West will follow suit. It is just that the cluster of emotional and intellectual pressures on voters is much the same all over--the nation. Obviously, no individual can be taken as representing the whole nation, but a well-selected group of residents of the East Coast can be assumed to be representative of the whole nation in terms of how much they are "pushed" by the various pressures of the election, so that their choices are likely to show general trends of the larger electorate.

Suppose it turned out that New Hampshire's Belknap County and

California's Modoc County had produced, over many national elections, very similar results. Would it follow that one of the two counties had been exerting some sort of causal influence on the other? Would they have had to be in some sort of eerie cosmic resonance mediated by "sympathetic magic" for this to happen? Certainly not. All it takes is for the electorates of the two counties to be similar; then the pressures that determine how people vote will take over and automatically make the results come out similar. It is no more mysterious than the observation that a Belknap County schoolgirl and a Modoc County schoolboy will get the same answer when asked to divide 507 by 13: the laws of arithmetic are the same the world over, and they operate the same in remote minds without any need for "sympathetic magic".

This is all elementary common sense; it should be the kind of thing that any well-educated person should understand clearly. And yet emotionally it cannot help but feel a little peculiar since it flies in the face of free will and regards people's decisions as caused simply by combinations of pressures with unknown values. On the other hand, perhaps that is a better way to look at decisions than to attribute them to "free will", a philosophically murky notion at best.

* * *

This may have seemed like a digression about statistics and the question of individual actions versus group predictability, but as a matter of fact it has plenty to do with the "correct action" to take in the dilemma of my letter. The question we were considering is: To what extent can what a few people do be taken as an indication of what all the people will do? We can sharpen it: To what extent can what one person does be taken as an indication of what all the people will do? The ultimate version of this question, stated in the first person, has a funny twist to it: To what extent does my choice inform me about the choices of the other participants?

You might feel that each person is completely unique and therefore that no one can be relied on as a predictor of how other people will act, especially in an intensely dilemmatic situation. There is more to the story, however. I tried to engineer the situation so that everyone would have the same image of the situation. In the dead center of that image was supposed to be the notion that everyone in the situation was using reasoning alone-including reasoning about the reasoning-to come to an answer.

Now, if reasoning dictates an answer, then everyone should independently come to that answer (just as the Belknap County schoolgirl and the Modoc County schoolboy would independently get 39 as their answer to the division problem). Seeing this fact is itself the critical step in the reasoning toward the correct answer, but unfortunately it eluded nearly everyone to whom I sent the letter. (That is why I came to wish I had included in the letter a paragraph stressing the rationality of the players.) Once you realize

this fact, then it dawns on you that either all rational players will choose D or all rational players will choose C. This is the crux.

Any number of ideal rational thinkers faced with the same situation and undergoing similar throes of reasoning agony will necessarily come up with the identical answer eventually, so long as reasoning alone is the ultimate justification for their conclusion. Otherwise reasoning would be subjective, not objective as arithmetic is. A conclusion reached by reasoning would be a matter of preference, not of necessity. Now some people may believe this of reasoning, but rational thinkers understand that a valid argument must be universally compelling, otherwise it is simply not a valid argument.

If you'll grant this, then you are 90 percent of the way. All you need ask now is, "Since we are all going to submit the same letter, which one would be more logical? That is, which world is better for the individual rational thinker: one with all C's or one with all D's?" The answer is immediate: "I get $57 if we all cooperate, $19 if we all defect. Clearly I prefer $57, hence cooperating is preferred by this particular rational thinker. Since I am typical, cooperating must be preferred by all rational thinkers. So I'll cooperate." Another way of stating it, making it sound weirder, is this: "If I choose C, then everyone will choose C, so I'll get $57. If I choose D, then everyone will choose D, so I'll get $19. I'd rather have $57 than $19, so I'll choose C. Then everyone will, and I'll get $57."

* * *

To many people, this sounds like a belief in voodoo or sympathetic magic, a vision of a universe permeated by tenuous threads of synchronicity, conveying thoughts from mind to mind like pneumatic tubes carrying messages across Paris, and making people resonate to a secret harmony. Nothing could be further from the truth. This solution depends in no way on telepathy or bizarre forms of causality. It's just that the statement "I'll choose C and then everyone will", though entirely correct, is somewhat misleadingly phrased. It involves the word "choice", which is incompatible with the compelling quality of logic. Schoolchildren do not choose what 507 divided by 13 is; they figure it out. Analogously, my letter really did not allow choice; it demanded reasoning. Thus, a better way to phrase the "voodoo" statement would be this: "If reasoning guides me to say C, then, as I am no different from anyone else as far as rational thinking is concerned, it will- guide everyone to say C."

The corresponding foray into the opposite world ("If I choose D, then everyone will choose D") can be understood more clearly by likening it to a musing done by the Belknap County schoolgirl before she divides: "Hmm, I'd guess that 13 into 507 is about 49-maybe 39. I see I'll have to calculate it out. But I know in advance that if I find out that it's 49, then sure as shootin', that Modoc County kid will write down 49 on his paper as well; and if I get 39 as my answer, then so will he." No secret transmissions are involved; all that is needed is the universality and uniformity of arithmetic.

Likewise, the argument "Whatever I do, so will everyone else do" is simply a statement of faith that reasoning is universal, at least among rational thinkers, not an endorsement of any mystical kind of causality.

This analysis shows why you should cooperate even when the opaque envelopes containing the other players' answers are right there on the table in front of you. Faced so concretely with this unalterable set of C's and D's, you might think, "Whatever they have done, I am better off playing D than playing C-for certainly what I now choose can have no retroactive effect on .what they chose. So I defect." Such a thought, however, assumes that the logic that now drives you to playing D has no connection or relation to the logic that earlier drove them to their decisions. But if you accept what was stated in the letter, then you must conclude that the decision you now make will be mirrored by the plays in the envelopes before you. If logic now coerces you to play D, it has already coerced the others to do the same, and for the same reasons; and conversely, if logic coerces you to play C, it has also already coerced the others to do that.

Imagine a pile of envelopes on your desk, all containing other people's answers to the arithmetic problem, "What is 507 divided by 13?" Having hurriedly calculated your answer, you are about to seal a sheet saying "49" inside your envelope, when at the last moment you decide to check it. You discover your error, and change the `4' to a '3'. Do you at that moment envision all the answers inside the other envelopes suddenly pivoting on their heels and switching from "49" to "39"? Of course not! You simply recognize that what is changing is your image of the contents of those envelopes, not the contents themselves. You used to think there were many "49"s. You now think there are many "39"s. However, it doesn't follow that there was a moment in between, at which you thought, "They're all switching from `49' to '39'!" In fact, you'd be crazy to think that.

It's similar with D's and C's. If at first you're inclined to play one way but on careful consideration you switch to the other way, the other players obviously won't retroactively or synchronistically follow you-but if you give them credit for being able to see the logic you've seen, you have to assume that their answers are what yours is. In short, you aren't going to be able to undercut them; you are simply "in cahoots" with them, like it or not! Either all D's, or all C's. Take your pick.

Actually, saying "Take your pick" is 100 percent misleading. It's not as if you could merely "pick", and then other people-even in the past-would magically follow suit! The point is that since you are going to be "choosing" by using what you believe to be compelling logic, if you truly respect your logic's compelling quality, you would have to believe that others would buy it as well, which means that you are certainly not "just picking". In fact, the more convinced you are of what you are playing, the more certain you should be that others will also play (or have already played) the same way, and for the same reasons. This holds whether you play C or D, and it is the real core of the solution. Instead of being a paradox, it's a self-reinforcing solution: a benign circle of logic.

* * *

If this still sounds like retrograde causality to you, consider this little tale, which may help make it all make more sense. Suppose you and Jane are classical music lovers. Over the years, you have discovered that you have incredibly similar tastes in music-a remarkable coincidence! Now one day you find out that two concerts are being given simultaneously in the town where you live. Both of them sound excellent to you, but Concert A simply cannot be missed, whereas Concert B is a strong temptation that you'll have to resist. Still, you're extremely curious about Concert B, because it features Zilenko Buznani, a violinist you've always heard amazing things about.

At first, you're disappointed, but then a flash crosses your mind: "Maybe I can at least get a first-hand report about Zilenko Buznani's playing from Jane. Since she and I hear everything through. virtually the same ears, it would be almost as good as my going if she would go." This is comforting for a moment, until it occurs to you that something is wrong here. For the same reasons as you do, Jane will insist on hearing Concert A. After all, she loves music in the same way as you do-that's precisely why you wish she would tell you about Concert B! The more you feel Jane's taste is the same as yours, the more you wish she would go to the other concert, so that you could know what it was like to have gone to it. But the more her taste is the same is yours, the less she will want to go to it!

The two of you are tied together by a bond of common taste. And if it turns out that you are different enough in taste to disagree about which concert is better, then that will tend to make you lose interest in what she might report, since you no longer can trust her opinion as that of someone who hears music "through your ears". In other words, hoping she'll choose Concert B is pointless, since it undermines your reasons for caring which concert she chooses!

The analogy is clear, I hope. Choosing D undermines your reasons for doing so. To the extent that all of you really are rational thinkers, you really will think in the same tracks. And my letter was supposed to establish beyond doubt the notion that you are all "in synch"; that is, to ensure that you can depend on the others' thoughts to be rational, which is all you need.

Well, not quite. You need to depend not just on their being rational, but on their depending on everyone else to be rational, and on their depending on everyone to depend on everyone to be rational-and so on. A group of reasoners in this relationship to each other I call superrational. Superrational thinkers, by recursive definition, include in their calculations the fact that they are in a group of superrational thinkers. In this way, they resemble elementary particles that are renormalized.

A renormalized electron's style of interacting with, say, a renormalized photon takes into account that the photon's quantum-mechanical structure includes "virtual electrons" and that the electron's quantum-mechanical structure includes "virtual photons"; moreover it takes into account that all

these virtual particles (themselves renormalized) also interact with one another. An infinite cascade of possibilities ensues but is taken into account in one fell swoop by nature. Similarly, superrationality, or renormalized reasoning, involves seeing all the consequences of the fact that other renormalized reasoners are involved in the same situation-and doing so in a finite swoop rather than succumbing to an infinite regress of reasoning about reasoning about reasoning ...

* * *

`C' is the answer I was hoping to receive from everyone. I was not so optimistic as to believe that literally everyone would arrive at this conclusion, but I expected a majority would-thus my dismay when the early returns strongly favored defecting. As more phone calls came in, I did receive some C's, but for the wrong reasons. Dan Dennett cooperated, saying, "I would rather be the person who bought the Brooklyn Bridge than the person who sold it. Similarly, I'd feel better spending $3 gained by cooperating than $10 gained by defecting."

Charles Brenner, who I'd figured to be a sure-fire D, took me by surprise and C'd. When I asked him why, he candidly replied, "Because I don't want to go on record in an international journal as a defector." Very well. Know, World, that Charles Brenner is a cooperator!

Many people flirted with the idea that everybody would think "about the same", but did not take it seriously enough. Scott Buresh confided to me: "It was not an easy choice. I found myself in an oscillation mode: back and forth. I made an assumption: that everybody went through the same mental processes I went through. Now I personally found myself wanting to cooperate roughly one third of the time. Based on that figure and the assumption that I was typical, I figured about one third of the people would cooperate. So I computed how much I stood to make in a field where six or seven people cooperate. It came out that if I were a D, I'd get about three times as much as if I were a C. So I'd have to defect. Water seeks out its own level, and I sank to the lower righthand corner of the matrix." At this point, I told Scott that so far, a substantial majority had defected. He reacted swiftly: "Those rats-how can they all defect? It makes me so mad! I'm really disappointed in your friends, Doug."

So was I, when the final results were in: Fourteen people had defected and six had cooperated-exactly what the networks would have predicted! Defectors thus received $43 while cooperators got $15. I wonder what Dorothy's saying to Peter about now? I bet she's chuckling and saying, "I told you I'd do better this way, didn't l?" Ah, me ... What can you do with people like that?

A striking aspect of Scott Buresh's answer is that, in effect, he treated his own brain as a simulation of other people's brains and ran the simulation enough to get a sense of what a "typical person" would do. This is very

much in the spirit of my letter. Having assessed what the statistics are likely to be, Scott then did a cool-headed calculation to maximize his profit, based on the assumption of six or seven cooperators. Of course, it came out in favor of defecting. In fact, it would have, no matter what the number of cooperators was! Any such calculation will always come out in favor of defecting. As long as you feel your decision is independent of others' decisions, you should defect. What Scott failed to tike into account was that cool-headed calculating people should take into account that cool-headed calculating people should take into account that cool-headed calculating people should take into account that ...

This sounds awfully hard to take into account in a finite way, but actually it's the easiest thing in the world. All it means is that all these heavy-duty rational thinkers are going to see that they are in a symmetric situation, so that whatever reason dictates to one, it will dictate to all. From that point on, the process is very simple. Which is better for an individual if it is a universal choice: C or D? That's all.

* * *

Actually, it's not quite all, for I've swept one possibility under the rug: maybe throwing a die could be better than making a deterministic choice. Like Chris Morgan, you might think the best thing to do is to choose C with probability p and D with probability 1-p. Chris arbitrarily let p be 1/2, but it could be any number between 0 and 1, where the two extremes represent Ding and C'ing respectively. What value of p would be chosen by superrational players? It is easy to figure out in a two-person Prisoner's Dilemma, where you assume that both players use the same value of p. The expected earnings for each, as a function of p, come out to be I+ 3p -p2, which grows monotonically as p increases from 0 to 1. Therefore, the optimum value ofp is 1, meaning certain cooperation. In the case of more players, the computations get more complex but the answer doesn't change: the expectation is always maximal when p equals 1. Thus this approach confirms the earlier one, which didn't entertain probabilistic strategies. - Rolling a die to determine what you'll do didn't add anything new to the standard Prisoner's Dilemma, but what about the modified-matrix version I gave in the P. S. to my letter? I'll let you figure that one out for yourself. And what about the Platonia Dilemma? There, two things are very clear: (1) if you decide not to send a telegram, your chances of winning are zero; (2) if everyone sends a telegram, your chances of winning are zero. If you believe that what you choose will be the same as what everyone else chooses because you are all superrational, then neither of these alternatives is very appealing. With dice, however, a new option presents itself to roll a die with probability p of coming up "good" and then to send in your name if and only if "good" comes up.

Now imagine twenty people all doing this, and figure out what value of

p maximizes the likelihood of exactly one person getting the go-ahead. It turns out that it is p = 1/20, or more generally, p=1/N where N is the number of participants. In the limit where N approaches infinity, the chance that exactly one person will get the go-ahead is 1/e, which is just-under 37 percent. With twenty superrational players all throwing icosahedral dice, the chance that you will come up the big winner is very close to 1/(20e), which is a little below two percent. That's not at all bad! Certainly it's a lot better than zero percent.

The objection many people raise is: "What if my roll comes up bad? Then why shouldn't I send in my name anyway? After all, if I fail to, I'll have no chance whatsoever of winning. I'm no better off than if I had never rolled my die and had just voluntarily withdrawn!" This objection seems overwhelming at first, but actually it is fallacious, being based on a misrepresentation of the meaning of "making a decision". A genuine decision to abide by the throw of a die means that you really must abide by the throw of the die; if under certain circumstances you ignore the die and do something else, then you never made the decision you claimed to have made. Your decision is revealed by your actions, not by your words before acting!

If you like the idea of rolling a die but fear that your will power may not be up to resisting the temptation to defect, imagine a third "Policansky button": this one says `R' for "Roll", and if you press it, it rolls a die (perhaps simulated) and then instantly and irrevocably either sends your name or does not, depending on which way the die came up. This way you are never allowed to go back on your decision after the die is cast. Pushing that button is making a genuine decision to abide by the roll of a die. It would be easier on any ordinary human to be thus shielded from the temptation, but any superrational player would have no trouble holding back after a bad roll.

* * *

This talk of holding back in the face of strong temptation brings me to the climax of this column: the announcement of a Luring Lottery open to all readers and nonreaders of Scientific American. The prize of this lottery is $1,000,000/N, where N is the number of entries submitted. Just think: If you are the only entrant (and if you submit only one entry), a cool million is yours! Perhaps, though, you doubt this will come about. It does seem a trifle iffy. If you'd like to increase your chances of winning, you are encouraged to send in multiple entries-no limit! Just send in one postcard per entry. If you send in 100 entries, you'll have 100 times the chance of some poor slob who sends in just one. Come to think of it, why should you have to send in multiple entries separately? Just send one postcard with your name and address and a positive integer (telling how many entries you're making) to:

Luring Lottery
c/o Scientific American
415 Madison Avenue
New York, N.Y. 10017

You will be given the same chance of winning as if you had sent in that number of postcards with `1' written on them. Illegible, incoherent, ill-specified, or incomprehensible entries will be disqualified. Only entries received by midnight June 30, 1983 will be considered. Good luck to you (but certainly not to any-other reader of this column)!

Post Scriptum.

The emotions churned up by the Prisoner's Dilemma are among the strongest I have ever encountered, and for good reason. Not only is it a wonderful intellectual puzzle, akin to some of the most famous paradoxes of all time, but also it captures in a powerful and pithy way the essence of a myriad deep and disturbing situations that we are familiar with from life. Some are choices we make every day; others are the kind of agonizing choices that we all occasionally muse about but hope the world will never make us face.

My friend Bob Wolf, a mathematician whose specialty is logic, adamantly advocated choosing D in the case of the letters I sent out. To defend his choice, he began by saying that it was clearly "a paradox with no rational solution", and thus there was no way to know what people would do. Then he said, "Therefore, I will choose D. I do better that way than any other way." I protested strenuously: "How dare you say `therefore' when you've just gotten through describing this situation as a paradox and claiming there is no rational answer? How dare you say logic is forcing an answer down your throat, when the premise of your `logic' is that there is no logical answer?" I never got what I considered a satisfactory answer from Bob, although neither of us could budge the other. However, I did finally get some insight into Bob's vision when he, pushed hard by my probing, invented a situation with a new twist to it, which I call "Wolf's Dilemma".

Imagine that twenty people are selected from your high school graduation class, you among them. You don't know which others have been selected, and you are told they are scattered all over the country. All you know is that they are all connected to a central computer. Each of you is in a little cubicle, seated on a chair and facing one button on an otherwise blank wall. You are given ten minutes to decide whether or not to push your button. At the end of that time, a light will go on for ten seconds, and while it is on, you may

either push or refrain from pushing. All the responses will then go to the central computer, and one minute later, they will result in consequences. Fortunately, the consequences can only be good. If you pushed your button, you will get $100, no strings attached, emerging from a small slot below the button. If nobody pushed their button, then everybody will get $1,000. But if there was even a single button-pusher, the refrainers will get nothing at all.

Bob asked me what I would do. Unhesitatingly, I said, "Of course I would not push the button. It's obvious!" To my amazement, though, Bob said he'd push the button with no qualms. I said, "What if you knew your co-players were all logicians?" He said that would make no difference to him. Whereas I gave credit to everybody for being able to see that it was to everyone's advantage to refrain, Bob did not. Or at least he expected that there is enough "flakiness" in people that he would prefer not to rely on the rationality of nineteen other people. But of course in assuming the flakiness of others, he would be his own best example-ruining everyone else's chances of getting $1,000.

What bothered me about Wolf's Dilemma was what I have come to call reverberant doubt. Suppose you are wondering what to do. At first it's obvious that everybody should avoid pushing their button. But you do realize that among twenty people, there might be one who is slightly hesitant and who might waver a bit. This fact is enough to worry you a tiny bit, and thus to make you waver, ever so slightly. But suddenly you realize that if you are wavering, even just a tiny bit, then most likely everyone is wavering a tiny bit. And that's considerably worse than what you'd thought at first-namely, that just one person might be wavering. Uh-oh! Now that you can imagine that everybody is at least contemplating pushing their button, the situation seems a lot more serious. In fact, now it seems quite probable that at least one person will push their button. But if that's the case, then pushing your own button seems the only sensible thing to do. As you catch yourself thinking this thought, you realize it must be the same as everyone else's thought. At this point, it becomes plausible that the majority of participants -possibly even all-will push their button! This clinches it for you, and so you decide to push yours.

Isn't this an amazing and disturbing slide from certain restraint to certain pushing? It is a cascade, a stampede, in which the tiniest flicker of a doubt has become amplified into the gravest avalanche of doubt. That's what I mean by "reverberant doubt". And one of the annoying things about it is that the brighter you are, the more quickly and clearly you see what there is to fear. A bunch of amiable slowpokes might well be more likely to unanimously refrain and get the big payoff than a bunch of razor-sharp logicians who all think perversely recursively reverberantly. It's that "smartness" to see that initial flicker of a doubt that triggers the whole avalanche and sends rationality a-tumblin' into-the abyss. So, dear reader . . . if you push that button in front of you, do you thereby lose $900 or do you thereby gain $100?

* * *

Wolf's Dilemma is not the same as the Prisoner's Dilemma. In the Prisoner's Dilemma, pressure towards defection springs from *hope for asymmetry* (i.e., hope that the other player might be dumber than you and thus make the opposite choice) whereas in Wolf's Dilemma, pressure towards button-pushing springs from fear of asymmetry (i.e., fear that the other player might be dumber than you and thus make the opposite choice). This difference shows up clearly in the games' payoff matrices for the two-person case (compare Figure 30-lb with Figure 29-1c). In the Prisoner's Dilemma, the temptation T is greater than the reward R (5 > 3), whereas in Wolf's Dilemma, R is greater than T (1,000 > 100).

Bob Wolf's choice in his own dilemma revealed to me something about his basic assessment of people and their reliability (or lack thereof). Since his adamant decision to be a button-pusher even in this case stunned me, I decided to explore that cynicism a bit more, and came up with this modified Wolf's Dilemma.

Imagine, as before, that twenty people have been selected from your high school graduation class, and are escorted to small cubicles with one button on the wall. This time, however, each of you is strapped into a chair, and a device containing a revolver is attached to your head. Like it or not, you are now going to play Russian roulette, the odds of your death to be determined by your choice. For anybody who pushes their button, the odds of survival will be set at 90 percent-only one chance in ten of dying. Not too bad, but given that there are twenty of you, it means that almost certainly one or two of you will die, possibly more. And what happens to the refrainers? It all depends on how many of them there are. Let's say there are N refrainers. For each one of them, their chance of being shot will be one in N2. For instance, if five people don't push, each of them will have only a 1/25 chance of dying. If ten people refrain, they will each get a 99 percent chance of survival. The bad cases are, of course, when nearly everybody pushes their button ("playing it safe", so to speak), leaving the refrainers in a tiny minority of three, two, or even one. If you're the sole refrainer, it's curtains for you-one chance in one of your death. Bye-bye! For two refrainers, it's one chance in four for each one. That means there's nearly a 50 percent chance that at least one of the two will perish.

Clearly the crossover line is between three and four refrainers. If you have a reasonable degree of confidence that at least three other people will hold back, you should definitely do so yourself. The only problem is, they're all making their decisions on the basis of trying to guess how many people will refrain, too! It's terribly circular, and you hardly know where to start. Many people, sensing this, just give up, and decide to push their button. (Actually, of course, how do I know? I've never seen people in such a situation-but it seems that way from evidence of real-life situations resembling this, and of course from how people respond to a mere description of this situation,

where they aren't really faced with any dire consequences at all. Still, I tend to believe them, by and large.) Calling such a decision "playing it safe" is quite ironic, because if only everybody "played it dangerous", they'd have a chance of only one in 400 of dying! So I ask you: Which way is safe, and which way dangerous? It seems to me that this Wolf Trap epitomizes the phrase "We have nothing to fear but fear itself."

Variations on Wolf's Dilemma include some even more frightening and unstable scenarios. For instance, suppose the conditions are that each button-pusher has a 50 percent chance of survival, but if there is unanimous refraining from pushing the button, everyone's life will be spared-and as before, if anyone pushes their button, all refrainers will die. You can play around with the number of participants, the survival chance, and so on. Each such variation reveals a new facet of grimness. These visions are truly horrific, yet all are just allegorical renditions of ordinary life's decisions, day in, day out.

* * *

I had originally intended to close the column with the following paragraph, but was dissuaded from it by friends and editors:

> I am sorry to say that I am simply inundated with letters from well-meaning readers, and I have discovered, to my regret, that I can barely find time to read all those letters, let alone answer them. I have been racking my brains for months trying to come up with some strategy for dealing with all this correspondence, but frankly I have not found a good solution yet. Therefore, I thought I would appeal to the collective genius of you-all out there. If you can think of some way for me to ease the burden of my correspondence, please send your idea to me. I shall be most grateful.

31

Irrationality Is the Square Root of All Evil

September, 1983

THE Luring Lottery, proposed in my June column, created quite a stir. Let me remind you that it was open to anyone; all you had to do was submit a postcard with a clearly specified positive integer on it telling how many entries you wished to make. This integer was to be, in effect, your "weight" in the final drawing, so that if you wrote "100", your name would be 100 times more likely to be drawn than that of someone who wrote 'I'. The only catch was that the cash value of the prize was inversely proportional to the sum of all the weights received by June 30. Specifically, the prize to be awarded was $1,000,0001W, where W is the sum of all the weights sent in.

The Luring Lottery was set up as an exercise in cooperation versus defection. The basic question for each potential entrant was: "Should I restrain myself and submit a small number of entries, or should I `go for it' and submit a large number? That is, should I cooperate, or should I defect?" Whereas in previous examples of cooperation versus defection there was a clear-cut dividing line between cooperators and defectors, here it seems there is a continuum of possible answers, hence of "degree of cooperation". Clearly one can be an extreme cooperator and voluntarily submit nothing, thus in effect cutting off one's nose to spite one's face. Equally clearly, one can be an extreme defector and submit a giant number of entries, hoping to swamp everyone else out but destroying the prize in so doing. However, there remains a lot of middle ground between these two extremes. What about someone who submits two entries, or one? What about someone who throws a six-sided die to decide whether or not to send in a single entry? Or a million-sided die?

Before I go further, it would be good for me to present my generalized and nonmathematical sense of these terms "cooperation" and "defection". As a child, you undoubtedly often encountered adults who admonished you

for walking on the grass or for making noise, saying "Tut, tut, tut just think if everyone did that!" This is the quintessential argument used against the defector, and serves to define the concept:

> A defection is an action such that, if everyone did it, things would clearly be worse (for everyone) than if everyone refrained from doing it, and yet which tempts everyone, since if only one individual (or a sufficiently small number) did it while others refrained, life would be sweeter for that individual (or select group).

Cooperation, of course, is the other side of the coin: the act of resisting temptation. However, it need not be the case that cooperation is passive while defection is active; often it is the exact opposite: The cooperative option may be to participate industriously in some activity, while defection is to lay back and accept the sweet things that result for everybody from the cooperators' hard work. Typical examples of defection are:

- loudly wafting your music through the entire neighborhood on a fine summer's day;
- not worrying about speeding through a four-way stop sign, figuring that the people going in the crosswise direction will stop anyway;
- not being concerned about driving a car everywhere, figuring that there's no point in making a sacrifice when other people will just continue to guzzle gas anyway;
- not worrying about conserving water in a drought, figuring "Everyone else will";
- not voting in a crucial election and excusing yourself by saying "One vote can't make any difference";
- not worrying about having ten children in a period of population explosion, leaving it to other people to curb their reproduction;
- not devoting any time or energy to pressing global issues such as the arms race, famine, pollution, diminishing resources, and so on, saying "Oh, of course I'm very concerned-but there's nothing one person can do."

When there are large numbers of people involved, people don't realize that their own seemingly highly idiosyncratic decisions are likely to be quite typical and are likely to be recreated many times over, on a grand scale; thus, what each couple feels to be their own isolated and private decision (conscious or unconscious) about how many children to have turns into a population explosion. Similarly, "individual" decisions about the futility of working actively toward the good of humanity amount to a giant trend of apathy, and this multiplied apathy translates into insanity at the group level. In a word, *apathy at the individual level translates into insanity at the mass level.*

* * *

Garrett Hardin, an evolutionary biologist, Wrote a famous article about this type of phenomenon, called "The Tragedy of the Commons". His view was that there are two types of rationality: one (I'll call it the "local" type) that strives for the good of the individual, the other (the "global" type) that strives for the good of the group; and that these two types of rationality are in an inevitable and eternal conflict. I would agree with his assessment, provided the individuals are unaware of their joint plight but are simply blindly carrying out their actions as if in isolation.

However, if they are fully aware of their joint situation, and yet in the face of it they blithely continue to act as if their situation were not a communal one, then I maintain that they are acting totally irrationally. In other words, with an enlightened citizenry, "local" rationality is not rational, period. It is damaging not just to the group, but to the individual. For example, people who defected in the One-Shot Prisoner's Dilemma situation I described in June did worse than if all had cooperated.

This was the central point of my June column, in which I wrote about renormalized rationality, or superrationality. Once you know you are a typical member of a class of individuals, you must act as if your own individual actions were to be multiplied manyfold, because they inevitably will be. In effect, to sample yourself is to sample the field, and if you fail to do what you wish the rest would do, you will be very disappointed by the rest as well. Thus it pays a lot to reflect carefully about one's situation in the world before defecting, that is, jumping to do the naively selfish act. You had better be prepared for a lot of other people copping out as well, and offering the same flimsy excuse.

People strongly resist seeing themselves as parts of statistical phenomena, and understandably so, because it seems to undermine their sense of free will and individuality. Yet how true it is that each of our "unique" thoughts is mirrored a million times over in the minds of strangers! Nowhere was this better illustrated than in the response to the Luring Lottery. It is hard to know precisely what constitutes the "field", in this case. It was declared universally open, to readers and nonreaders alike. However, we would be safe in assuming that few nonreaders ever became aware of it, so let's start with the circulation of Scientific American, which is about a million. Most of them, however, probably did no more than glance over my June column, if that; and of the ones who did more than that (let's say 100,000), still only a fraction-maybe one in ten-read it carefully from start to finish. I would thus estimate that there were perhaps 10,000 people motivated enough to read it carefully and to ponder the issues seriously. In any case, I'll take this figure as the population of the "field".

In my June column, I spelled out plainly, for all to see, the superrational argument that applies to the Platonia Dilemma, for rolling an N-sided die and entering only if it came up on the proper side. Here, a similar argument goes through. In the Platonia Dilemma, where more than one entry is fatal to all, the ideal die turned out to have N faces, where N is the number of

players-hence, with 10,000 players, a 10,000-sided die. In the Luring Lottery, the consequences aren't so drastic if more than one entry is submitted. Thus, the ideal number of faces on the die turns out to be about 2/3 as many-in the case of 10,000 players, a 6,667-sided die would do admirably. Giving the die fewer than 10,000 sides of course slightly increases each player's chance of sending in one entry. This is to make it quite likely that at least one entry will arrive!

With 6,667 faces on the die, each superrational player's chance of winning is not quite 1 in 10,000, but more like 1 in 13,000; this is because there is about a 22 percent chance that no one's die will land right, so no one will send in any entry at all, and no one will win. But if you give the die still fewer faces-say 3,000-the expected size of the pot gets considerably smaller, since the expected number of entrants grows. And if you give it more faces -say 20,000-then you run a considerable risk of having no entries at all. So there's a trade-off whose ideal solution can be calculated without too much trouble, and 6,667 faces turns out to be about optimal. With that many faces, the expected value of the pot is maximal: nearly $520,000-not to be sneered at.

Now this means that had everyone followed my example in the June column, I would probably have received a total of one or two postcards with ' 1' written on them, and one of those lucky people would have gotten a huge sum of money! But do you think that is what happened? Of course not! Instead, I was inundated with postcards and letters from all over the world -over 2,000 of them. What was the breakdown of entries? I have exhibited part of it in a table, below:

1: 1,133
2: 31
3: 16
4: 8
5: 16
6: 0
7: 9
8: 1
9: 1
10: 49
100: 61
1,000: 46
1,000,000: 33
1,000,000,000: 11
602,300,000,000,000,000,000,000 (Avogadro's number): 1
10100 (a googol): 9
1010100 (a googolplex): 14

Curiously, many if not most of the people who submitted just one entry patted themselves on the back for being "cooperators". Hogwash! The real cooperators were those among the 10,000 or so avid readers who calculated the proper number of faces of the die, used a random-number table or something equivalent, and then-most likely-rolled themselves out. A few people wrote to tell me they had rolled themselves out in this way. I appreciated hearing from them. It is conceivable, just barely, that among the thousand-plus entries of `1' there was one that came from a lucky superrational cooperator-but I doubt it. The people who simply withdrew without throwing a die I would characterize as well-meaning but a bit lazy, not true cooperators-something like people who simply contribute money to a political cause but then don't want to be bothered any longer about it. It's the lazy way of claiming cooperation.

By the way, I haven't by any means finished with my score chart. However, it is a bit disheartening to try to relate what happened. Basically, it is this. Dozens and dozens of readers strained their hardest to come up with inconceivably large numbers. Some filled their whole postcard with tiny `9's, others filled their card with rows of exclamation points, thus creating iterated factorials of gigantic sizes, and so on. A handful of people carried this game much further, recognizing that the optimal solution avoids all pattern (to see why, read Gregory Chaitin's article "Randomness and Mathematical Proof"), and consists simply of a "dense pack" of definitions built on definitions, followed by one final line in which the "fanciest" of the definitions is applied to a relatively small number such as 2, or better yet, 9.

I received, as I say, a few such entries. Some of them exploited such powerful concepts of mathematical logic and set theory that to evaluate which one was the largest, became a very serious problem, and in fact it is not even clear that I, or for that matter anyone else, would be able to determine which is the largest integer submitted. I was strongly reminded of the lunacy and pointlessness of the current arms race, in which two sides vie against each other to produce arsenals so huge that not even teams of experts can meaningfully say which one is larger-and meanwhile, all this monumental effort is to the detriment of everyone.

* * *

Did I find this amusing? Somewhat, of course. But at the same time, I found it disturbing and disappointing. Not that I hadn't expected it. Indeed, it was precisely what I had expected, and it was one reason I was so sure the Luring Lottery would be no risk for the magazine.

This short-sighted race for "first place" reveals the way in which people in a huge crowd erroneously consider their own fancies to be totally unique. I suspect that nearly everyone who submitted a number above 1,000,000 actually believed they were going to be the only one to do so. Many of those who submitted numbers such as a googolplex, or a `9' followed by

thousands of factorial signs, explicitly indicated that they were pretty sure that they were going to "win". And then those people who pulled out all the stops and sent in definitions that would boggle most mathematicians were very sure they were going to win. As it turns out, I don't know who won, and it doesn't matter, since the prize is zero to such a good approximation that even God wouldn't know the difference.

Well, what conclusion do I draw from all this? None too serious, but I do hope that it will give my readers pause for thought next time they face a "cooperate-or-defect" decision, which will likely happen within minutes for each of you, since we face such decisions many times each day. Some of them are small, but some will have monumental repercussions. The globe's future is in your hands-and yes, I mean you (as well as every other reader of this column).

* * *

And with this perhaps sobering conclusion, I would like to draw my term as a columnist for Scientific American to a close. It has been a valuable and beneficial opportunity for me. I have enjoyed having a platform from which to express my ideas and concerns, I have-at least sometimes-enjoyed receiving the huge shipments of mail forwarded to me from New York several times a month, and I have certainly been happy to make new friends through this channel. I won't miss the monthly deadline, but I will undoubtedly come across ideas, from time to time, that would have made perfect "Metamagical Themas". I will be keeping them in mind, and maybe at some future time will write a similar set of essays.

But for now, it is time for me to move on to other territory: I look forward to a return to my professional work, and to a more private life. Good-bye, and best wishes to you and to all other readers of this magazine, this issue, this copy, this piece, this page, this column, this paragraph, this sentence, and, last but not least, this "this".

Post Scriptum.

What do you do when in a crushingly cold winter, you hear over the radio that there is a severe natural gas shortage in your part of the country, and everyone is requested to turn their thermostat down to 60 degrees? There's no way anyone will know if you've complied or not. Why shouldn't you toast in your house and let all the rest of the people cut down their consumption? After all, what you do surely can't affect what anyone else does.

This is a typical "tragedy of the commons" situation. A common resource has reached the point of saturation or exhaustion, and the questions for each individual now are: "How shall I behave? Am I typical? How does a

lone person's action affect the big picture?" Garrett Hardin's article "The Tragedy of the Commons" frames the scene in terms of grazing land shared by a number of herders. Each one is tempted to increase their own number of animals even when the land is being used beyond its optimum capacity, because the individual gain outweighs the individual loss, even though in the long run, that decision, multiplied throughout the population of herders, will destroy the land totally.

The real reason behind Hardin's article was to talk about the population explosion and to stress the need for rational global planning-in fact, for coercive techniques similar to parking tickets and jail sentences. His idea is that families should be allowed to have many children (and thus to use a large share of the common resources) but that they should be penalized by society in the same way as society "allows" someone to rob a bank and then applies sanctions to those who have made that choice. In an era when resources are running out in a way humanity has never had to face heretofore, new kinds of social arrangements and expectations must be imposed, Hardin feels, by society as a whole. He is a dire pessimist about any kind of superrational cooperation, emphasizing that cooperators in the birth-control game will breed themselves right out of the population. A perfect illustration of why this is so is the man I heard about recently: he secretly had ten wives and by them had sired something like 35 children by the time he was 30. With genes of that sort proliferating wildly, there is little hope for the more modest breeders among us to gain the upper hand. Hardin puts it bluntly: "Conscience is self-eliminating." He goes even further and says:

> The argument has here been stated in the context of the population problem, but it applies equally well to any instance in which society appeals to an individual exploiting a commons to restrain himself for the general good-by means of his conscience. To make such an appeal is to set up a selective system that works toward the elimination of conscience from the race.

An even more pessimistic vision of the future is proffered us by -one Walter Bradford Ellis, a hypothetical speaker representing the views of his inventor, Louis Pascal, in a hypothetical speech:

> The United States-indeed the whole earth-is fast running out of the resources it depends on for its existence. Well before the last of the world's supplies of oil and natural gas are exhausted early in the next century, shortages of these and other substances will have brought about the collapse of our whole economy and, indeed, of our whole technology. And without the wonders of modern technology, America will be left a grossly overpopulated, utterly impoverished, helpless, dying land. Thus I foresee a whole world full of wretched, starving people with no hope of escape, for the only countries which could have aided them will soon be no better off than the rest. And thus unless we are saved from this future by the blessing of a nuclear war or a truly lethal

> pestilence, I see stretching off into eternity a world of indescribable suffering and hopelessness. It is a vision of truly unspeakable horror mitigated only by the fact that try as I might I could not possibly concoct a creature more deserving of such a fate.

Whew! The circularity of the final thought reminds me of an idea I once had: that it will be just as well if humanity destroys itself in a nuclear holocaust, because civilizations that destroy themselves are barbaric and stupid, and who would want to have one of them around, polluting the universe?

Pascal's thoughts, expressed in his article "Human Tragedy and Natural Selection" and in his rejoinder to an article by two critics called "The Loving Parent Meets the Selfish Gene" (which is where Ellis' speech is printed), are strikingly reminiscent of the thoughts of his earlier namesake Blaise, who in an unexpected use of his own calculus of probabilities managed to convince himself that the best possible way to spend his life was in devotion to a God who he wasn't sure (and couldn't be sure) existed. In fact, Pascal felt, even if the chances of God's existence were one in ,a million, faith in that God would pay off in the end, because the potential rewards (or punishments) if Heaven and Hell exist are infinite, and all earthly rewards and punishments, no matter how great, are still finite. The favored behavior is to be a believer, Pascal "calculated"-regardless of what you do believe. Thus Blaise Pascal devoted his brilliant mind to theology.

Louis Pascal, following in his forebear's mindsteps, has opted to devote his life to the world's population problem. And he can produce mathematical arguments to show why you should, too. To my mind, there is no question that such arguments have considerable force. There are always points to nitpick over, but in essence, thinkers like Hardin and Pascal and Anne and Paul Ehrlich and many others have recognized and internalized the novelty of the human situation at this moment in history: the moment when humanity has to grapple with dwindling resources and overwhelmingly huge weapons systems. Not many people are willing to wrestle with this beast, and consequently the burden falls all the more heavily on those few who are.

* * *

It has disturbed me how vehemently and staunchly my clear-headed friends have been able to defend their decisions to defect. They seem to be able to digest my argument about superrationality, to mull it over, to begrudge some curious kind of validity to it, but ultimately to feel on a gut level that it is wrong, and to reject it. This has led me to consider the notion that my faith in the superrational argument might be similar to a self-fulfilling prophecy or self-supporting claim, something like being absolutely convinced beyond a shadow of a doubt that the Henkin sentence "This sentence is true" actually must be true-when, of course, it is equally defensible to believe it to be false. The sentence is undecidable; its truth

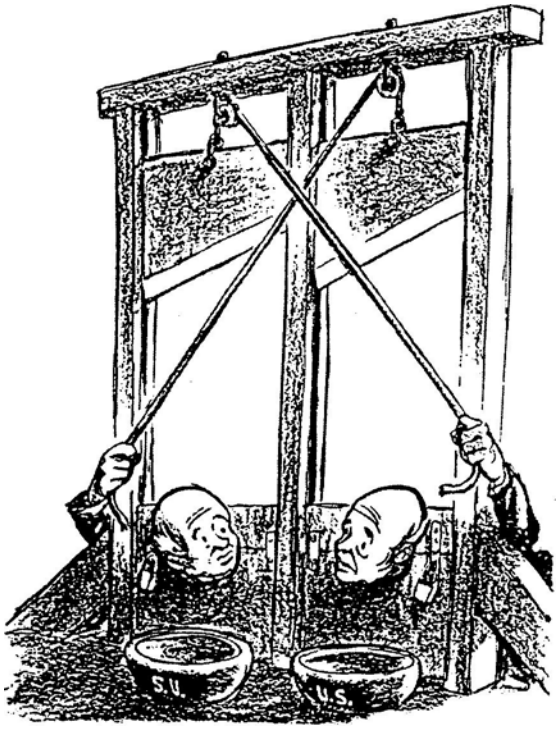
value is stable, whichever way you wish it to go (in this way, it is the diametric opposite of the Epimenides sentence "This sentence is false", whose truth value flips faster than the tip of a happy pup's tail). One difference, though, between the Prisoner's Dilemma and oddball self-referential sentences is that whereas your beliefs about such sentences' truth values usually have inconsequential consequences, with the Prisoner's Dilemma, it's quite another matter.

I sometimes wonder whether there haven't been many civilizations Out There, in our galaxy and beyond, that have already dealt with just these types of gigantic social problems-Prisoner's Dilemmas, Tragedies of the Commons, and so forth. Most likely some would have survived, some would have perished. And it occurs to me that perhaps the ultimate difference in those societies may have been the survival of the meme that, in effect, asserts the logical, rational validity of cooperation in a one-shot Prisoner's Dilemma. In a way, this would be the opposite thesis to Hardin's. It would say that lack of conscience is self-eliminating-provided you wait long enough that natural selection can act at the level of entire societies.

Perhaps on some planets, Type I societies have evolved, while on others, Type II societies have evolved. By definition, members of Type I societies believe in the rationality of lone, uncoerced, one-shot cooperation (when faced with members of Type I societies), whereas members of Type II societies reject the rationality of lone, uncoerced, one-shot cooperation, irrespective of who they are facing. (Notice the tricky circularity of the definition of Type I societies. Yet it is not a vacuous definition!) Both types of society find their respective answer to be obvious-they just happen to find opposite answers. Who knows-we might even happen to have some Type I societies here on earth. I cannot help but wonder how things would turn out if my little one-shot Prisoner's Dilemma experiment were carried out in Japan instead of the U.S. In any case, the vital question is: Which type of society survives, in the long run?

It could be that the one-shot Prisoner's Dilemma situations that I have described are undecidable propositions within the logic that we humans have developed so far, and that new axioms can be added, like the parallel postulate in geometry, or Godel sentences (and related ones) in mathematical logic. (Take a look at Figure 31-1, and see what kind of logic will extract those two poor devils from their one-shot dilemma.) Those civilizations to which cooperation appears axiomatic-Type I societieswind up surviving, I would venture to guess, whereas those to which defection appears axiomatic-Type II societies-wind up perishing. This suggestion may seem all wet to you, but watch those superpowers building those bombs, more and more of them every day, helplessly trapped in a rising spiral, and think about it. Evolution is a merciless pruner of ill logic.

Most philosophers and logicians are convinced that truths of logic are "analytic" and a priori; they do not like to think that such basic ideas are grounded in mundane, arbitrary things like survival. They might admit that



"The problem is how to turn loose without letting go."

FIGURE 31-1. *One powerful metaphor for the absurdity we have collectively dug ourselves into. The symmetry of the situation is acutely portrayed in this cartoon drawn by Bill Mauldin in 1960. Note that if either person releases his rope, thus chopping of his counterpart's head, that person's hand will go limp, thus releasing his rope and causing the other blade to fall and chop of the head of the instigator. That idea is a centerpiece of our current nuclear deterrence strategy: Even if we are wiped of the globe, our trUSty missiles will still wreak divine revenge on the evil empire of Satanic Uglies who dared do harm to US.*

natural selection tends to favor good logic-but they would certainly hate the suggestion that natural selection defines good logic! Yet truth and survival value are all tangled together, and civilizations that survive certainly have glimpsed higher truths than those that perish. When you argue with someone whose ideas you are sure are wrong but who dances an infuriatingly inconsistent yet self-consistent verbal dance in front of you, your one solace is that something in life may yet change this person's mind, even though your own best logic is helpless to do so. Ultimately, beliefs have to be grounded in experience, whether that experience is the organism's or its ancestors' or its peer group's. (That s what Chapter 5, particularly its

P.S., was all about.) My feeling is that the concept of superrationality is one whose truth will come to dominate among intelligent beings in the universe simply because its adherents will survive certain kinds of situations where its opponents will perish. Let's wait a few spins of the galaxy and see. After all, healthy logic is whatever remains after evolution's merciless pruning.

*** * ***

I was describing the Copycat project (Chapter 24) to physicist Victor Weisskopf, and I gave him our canonical example: "If abc goes to abd, what does xyz go to?" After we had discussed various possible answers and settled on wyz as the most compelling for reasons of symmetry, he surprised me by saying this: "You know, the root of the world's deepest problems is the tragic inability on the part of the world's leaders to see such basic symmetries. For instance, that the U.S. is to the S.U. what the S.U. is to the U.S.-that is too much for them to accept." Oh, but how could Weisskopf be so silly? After all, we're not trying to export communism to the entire world!

Logician Raymond Smullyan, who first heard about the Prisoner's Dilemma from me and who was absolutely delighted by it, also surprised me, but in a different way: He vehemently insisted on the correctness of defection in a one-shot situation no matter who might be on the other side, including his twin or his clone! (He did waver about his mirror image.) But just as I was giving up on him as a lost cause, he conceded this much to me: "I suspect, Doug, that this problem is a lot knottier than you or I suspect." Indeed, I suspect so, Raymond.

32

The Tale of Happiton

June, 1983

HAPPITON was a happy little town. It had 20,000 inhabitants, give or take 7, and they were productive citizens who mowed their lawns quite regularly. Folks in Happiton were pretty healthy. They had a life expectancy of 75 years or so, and lots of them lived to ripe old ages. Down at the town square, there was a nice big courthouse with all sorts of relics from WW II and monuments to various heroes and whatnot. People were proud, and had the right to be proud, of Happiton.

On the top of the courthouse, there was a big bell that boomed every hour on the hour, and you could hear it far and wide-even as far out as Shady Oaks Drive, way out nearly in the countryside.

One day at noon, a few people standing near the courthouse noticed that right after the noon bell rang, there was a funny little sound coming from up in the belfry. And for the next few days, folks noticed that this scratching sound was occurring after every hour. So on Wednesday, Curt Dempster climbed up into the belfry and took a look. To his surprise, he found a crazy kind of contraption rigged up to the bell. There was this mechanical hand, sort of a robot arm, and next to it were five weird-looking dice that it could throw into a little pan. They all had twenty sides on them, but instead of being numbered 1 through 20, they were just numbered 0 through 9, but with each digit appearing on two opposite sides. There was also a TV camera that pointed at the pan and it seemed to be attached to a microcomputer or something. That's all Curt could figure out. But then he noticed that on top of the computer, there was a neat little envelope marked "To the friendly folks of Happiton". Curt decided that he'd take it downstairs and open it in the presence of his friend the mayor, Janice Fleener. He found Janice easily enough, told her about what he'd found, and then they opened the envelope. How neatly it was written! It said this:

Grotto 19, Hades
June 20, 1983

Dear folks of Happiton,

I've got some bad news and some good news for you. The bad first. You know your bell that rings every hour on the hour? Well, I've set it up so that each time it rings, there is exactly one chance in a hundred thousand-that is, 1/100,000-that a Very Bad Thing will occur. The way I determine if that Bad Thing will occur is, I have this robot arm fling its five dice and see if they all land with `7' on top. Most of the time, they won't. But if they do-and the odds are exactly 1 in 100,000-then great clouds of an unimaginably revoltingsmelling yellow-green gas called "Retchgoo" will come oozing up from a dense network of underground pipes that I've recently installed underneath Happiton, and everyone will die an awful, writhing, agonizing death. Well, that's the bad news.

Now the good news! You all can prevent the Bad Thing from happening, if you send me a bunch of postcards. You see, I happen to like postcards a whole lot (especially postcards of Happiton), but to tell the truth, it doesn't really much matter what they're of. I just love postcards! Thing is, they have to be written personally-not typed, and especially not computer-printed or anything phony like that. The more cards, the better. So how about sending me some postcards-batches, bunches, boxes of them?

Here's the deal. I reckon a typical postcard takes you about 4 minutes to write. Now suppose just one person in all of Happiton spends 4 minutes one day writing me, so the next day, I get one postcard. Well, then, I'll do you all a favor: I'll slow the courthouse clock down a bit, for a day. (I realize this is an inconvenience, since a lot of you tell time by the clock, but believe me, it's a lot more inconvenient to die an agonizing, writhing death from the evil-smelling, yellow-green Retchgoo.) As I was saying, I'll slow the clock down for one day, and by how much? By a factor of 1.0000 1. Okay, I know that doesn't sound too exciting, but just think if all 20,000 of you send me a card! For each card I get that day, I'll toss in a slow-up factor of 1.00001, the next day. That means that by sending me 20,000 postcards a day, you all, working together, can get the clock to slow down by a factor of 1.00001 to the 20, 000th power, which is just a shade over 1.2, meaning it will ring every 72 minutes.

All right, I hear you saying, "72 minutes is just barely over an hour!" So I offer you more! Say that one day I get 160,000 postcards (heavenly!). Well then, the very next day I'll show my gratitude by slowing your clock down, all day long, midnight to midnight, by 1.00001 to the 160,000th power, and that ain't chickenfeed. In fact, it's about 5, and that means the clock will ring only every 5 hours, meaning those sinister dice will only get rolled about 5 times (instead of the usual 24). Obviously, it's better for both of us that way. You have to bear in mind that I don't have any personal interest in seeing that awful Retchgoo come rushing and gushing up out of those pipes and causing every last one of you to perish in grotesque, mouth-foaming, twitching convulsions. All I care about is getting postcards! And to send me 160,000 a day wouldn't cost you folks that much effort, being that it's just 8 postcards a day just about a half hour a day for each of you, the way I reckon it.

So my deal is pretty simple. On any given day, I'll make the clock go off once every X hours, where X is given by this simple formula:

$$X = 1.00001^{N}$$

Here, N is the number of postcards I received the previous day. If N is 20,000, then X will be 1.2, so the bell would ring 20 times per day, instead of 24. If N is 160,000, then X jumps way up to about 5, so the clock would slow way down just under 5 rings per day. If I get no postcards, then the clock will ring once an hour, just as it does now. The formula reflects that, since if N is 0, X will be 1. You can work out other figures yourself. Just think how much safer and securer you'd all feel knowing that your courthouse clock was ticking away so slowly!

I'm looking forward with great enthusiasm to hearing from you all.

Sincerely yours,
Demon #3127

The letter was signed with beautiful medieval-looking flourishes, in an unusual shade of deep red ... ink?

"Bunch of hogwash!" spluttered Curt. "Let's go up there and chuck the whole mess down onto the street and see how far it bounces." While he was saying this, Janice noticed that there was a smaller note clipped onto the back of the last sheet, and turned it over to read it. It said this:

P. S. -It's really not advisable to try to dismantle my little set-up up there in the belfry: I've got a hair trigger linked to the gas pipes, and if anyone tries to dismantle it, pssssst! Sorry.

Janice Fleener and Curt Dempster could hardly believe their eyes. What gall! They got straight on the -phone to the Police Department, and talked to Officer Curran. He sounded poppin' mad when they told him what they'd found, and said he'd do something about it right quick. So he hightailed it over to the courthouse and ran up those stairs two at a time, and when he reached the top, a-huffin' and a-puffin', he swung open the belfry door and took a look. To tell the truth, he was a bit ginger in his inspection, because one thing Officer Curran had learned in his many years of police experience is that an ounce of prevention is worth a pound of cure. So he cautiously looked over the strange contraption, and then he turned around and quite carefully shut the door behind him and went down. He called up the town sewer department and asked them if they could check out whether there was anything funny going on with the pipes underground.

Well, the long and the short of it is that they verified everything in the Demon's letter, and by the time they had done so, the clock had struck five more times and those five dice had rolled five more times. Janice Fleener had in fact had her thirteen-year-old daughter Samantha go up and sit in a

wicker chair right next to the microcomputer and watch the robot arm throw those dice. According to Samantha, an occasional 7 had turned up now and then, but never had two 7's shown up together, let alone 7's on all five of the weird-looking dice!

* * *

The next day, the Happiton Eagle-Telephone came out with a front-page story telling all about the peculiar goings-on. This caused quite a commotion. People everywhere were talking about it, from Lidden's Burger Stop to Bixbee's Druggery. It was truly the talk of the town.

When Doc Hazelthorn, the best pediatrician this side of the Cornyawl River, walked into Ernie's Barbershop, corner of Cherry and Second, the atmosphere was more somber than usual. "Whatcha gonna do, Doc?" said big Ernie, the jovial barber, as he was clipping the few remaining hairs on old Doc's pate. Doc (who was also head of the Happiton City Council) said the news had come as quite a shock to him and his family. Red Dulkins, sitting in the next chair over from Doc, said he felt the same way. And then the two gentlemen waiting to get their hair cut both added their words of agreement. Ernie, summing it up, said the whole town seemed quite upset. As Ernie removed the white smock from Doc's lap and shook the hairs off it, Doc said that he had just decided to bring the matter up first thing at the next City Council meeting, Tuesday evening. "Sounds like a good idea, Doc!" said Ernie. Then Doc told Ernie he couldn't make the usual golf date this weekend, because some friends of his had invited him to go fishing out at Lazy Lake, and Doc just couldn't resist.

Two days after the Demon's note, the Eagle-Telephone ran a feature article in which many residents of Happiton, some prominent, some not so prominent, voiced their opinions. For instance, eleven-year-old Wally Thurston said he'd gone out and bought up the whole supply of picture postcards at the 88-Cent Store, $14.22 worth of postcards, and he'd already started writing a few. Andrea McKenzie, sophomore at Happiton High, said she was really worried and had had nightmares about the gas, but her parents told her not to worry, things had a way of working out. Andrea said maybe her parents weren't taking it so seriously because they were a generation older and didn't have as long to look forward to anyway. She said she was spending an hour each day writing postcards. That came to 15 or 16 cards each day. Hank Hoople, a janitor at Happiton High, sounded rather glum: "It's all fate. If the bullet has your name on it, it's going to happen, whether you like it or not." Many other citizens voiced concern and even alarm about the recent developments.

But some voiced rather different feelings. Ned Furdy, who as far as anyone could tell didn't do much other than hang around Simpson's bar all day (and most of the night) and buttonhole anyone he could, said, "Yeah, it's a problem, all right, but I don't know nothin' about gas and statistics and such.

It should all be left to the mayor and the Town Council, to take care of. They know what they're doin'. Meanwhile, eat, drink, and be merry!" And Lulu Smyth, 77-year-old. proprietor of Lulu's Thread 'N Needles Shop, said "I think it's all a ruckus in a teapot, in my opinion. Far as I'm concerned, I'm gonna keep on sellin' thread 'n needles, and playin' gin rummy every third Wednesday."

* * *

When Doc Hazelthorn came back from his fishing weekend at Lazy Lake, he had some surprising news to report. "Seems there's a demon left a similar set-up in the church steeple down in Dwaynesville", he said. (Dwaynesville was the next town down the road, and the arch-rival of Happiton High in football.) "The Dwaynesville demon isn't threatening them with gas, but with radioactive water. Takes a little longer to die, but it's just as bad. And I hear tell there's a demon with a subterranean volcano up at New Athens." (New Athens was the larger town twenty miles up the Cornyawl from Dwaynesville, and the regional center of commerce.)

A lot of people were clearly quite alarmed by all this, and there was plenty of arguing on the streets about how it had all happened without anyone knowing. One thing that was pretty universally agreed on was that a commission should be set up as soon as possible, charged from here on out with keeping close tabs on all subterranean activity within the city limits, so that this sort of outrage could never happen again. It appeared probable that Curt Dempster, who was the moving force behind this idea, would be appointed its first head.

Ed Thurston (Wally's father) proposed to the Jaycees (of which he was a member in good standing) that they donate $1,000 to support a postcard-writing campaign by town kids. But Enoch Swale, owner of Swale's Pharmacy and the Sleepgood Motel, protested. He had never liked Ed much, and said Ed was proposing it simply because his son would gain status that way. (It was true that Wally had recruited a few kids and that they spent an hour each afternoon after school writing cards. There had been a small article in the paper about it once.) After considerable debate, Ed's motion was narrowly defeated. Enoch had a lot of friends on the City Council.

Nellie Doobar, the math teacher at High, was about the only one who checked out the Demon's math. "Seems right to me", she said to the reporter who called her about it. But this set her to thinking about a few things. In an hour or two, she called back the paper and said, "I figured something out. Right now, the clock is still ringing very close to once every hour. Now there are about 720 hours per month, and so that means there are 720 chances each month for the gas to get out. Since each chance is 1 in 100,000, it turns out that each month, there's a bit less than a 1-in-100 chance that Happiton will get gassed. At that rate, there's about 11 chances in 12 that Happiton will make it through each year. That may sound pretty

good, but the chances we'll make it through any 8-year period are almost exactly 50-50, exactly the same as tossing a coin. So we can't really count on very many years ..."

This made big headlines in the next afternoon's Eagle-Telephone-in fact, even bigger than the plans for the County Fair! Some folks started calling up Mrs. Doobar anonymously and telling her she'd better watch out what she was saying if she didn't want to wind up with a puffy face or a fat lip. Seems like they couldn't quite keep it straight that Mrs. Doobar wasn't the one who'd set the thing up in the first place.

After a few days, though, the nasty calls died down pretty much. Then Mrs. Doobar called up the paper again and told the reporter, "I've been calculating a bit more here, and I've come up with the following, and they're facts every last one of them. If all 20,000 of us were to spend half an hour a day writing postcards to the Demon, that would amount to 160,000 postcards a day, and just as the Demon said, the bell would ring pretty near every five hours instead of every hour, and that would mean that the chances of us getting wiped out each month would go down considerable. In fact, there would only be about 1 chance in 700 that we'd go down the tubes in any given month, and only about a chance in 60 that we'd get zapped each year. Now I'd say that's a darn sight better than 1 chance in 12 per year, which is what it is if we don't write any postcards (as is more or less the case now, except for Wally Thurston and Andrea McKenzie and a few other kids I heard of). And for every 8-year period, we'd only be running a 13 percent risk instead of a 50 percent risk."

"That sounds pretty good", said the reporter cheerfully.

"Well," replied Mrs. Doobar, "it's not too bad, but we can get a whole lot better by doublin' the number of postcards."

"How's that, Mrs. Doobar?" asked the reporter. "Wouldn't it just get twice as good?"

"No, you see, it's an exponential curve," said Mrs. Doobar, "which means that if you *double* N, you *square* X. "

"That's Greek to me", quipped the reporter.

"N is the number of postcards and X is the time between rings", she replied quite patiently. "If we all write a half hour a day, X is 5 hours. But that means that if we all write a whole hour a day, like Andrea McKenzie in my algebra class, X jumps up to 25 hours, meaning that the clock would ring only about once a day, and obviously, that would reduce the danger a lot. Chances are, hundreds of years would pass before five 7's would turn up together on those infernal dice. Seems to me that under those circumstances, we could pretty much live our lives without worrying about the gas at all. And that's for writing about an hour a day, each one of us."

The reporter wanted some more figures detailing how much different amounts of postcard-writing by the populace would pay off, so Mrs. Doobar obliged by going back and doing some more figuring. She figured out that if 10,000 people-half the population of Happiton-did 2 hours a day for

the year, they could get the same result-one ring ever)' 25 hours. If only 5,000 people spent 2 hours a day, or if 10,000 people spent one hour a day, then it would go back to one ring every 5 hours (still a lot safer than one every hour). Or, still another way of looking at it, if just 1250 of them worked full-time (8 hours a day), they could achieve the same thing.

"What about if we all pitch in and do 4 minutes a day, Mrs. Doobar?" asked the reporter.

"Fact is, 'twouldn't be worth a damn thing! (Pardon my French.)" she replied. "N is 20,000 that way, and even though that sounds pretty big, X works out to be just 1.2, meaning one ring every 1.2 hours, or 72 minutes. That way, we still have about a chance of 1 in 166 every month of getting wiped out, and 1 in 14 every year of getting it. Now that's real scary, in my book. Writing cards only starts making a noticeable difference at about 15 minutes a day per person."

* * *

By this time, several weeks had passed, and summer was getting into full swing. The County Fair was buzzing with activity, and each evening after folks came home, they could see loads of fireflies flickering around the trees in their yards. Evenings were peaceful and relaxed. Doc'Hazelthorn was playing golf every weekend, and his scores were getting down into the low 90's. He was feeling pretty good. Once in a while he remembered the Demon, especially when he walked downtown and passed the courthouse tower, and every so often he would shudder. But he wasn't sure what he and the City Council could do about it.

The Demon and the gas still made for interesting talk, but were`no longer such big news. Mrs. Doobar's latest revelations made the paper, but. were. relegated this time to the second section, two pages before the comics, right next to the daily horoscope column. Andrea McKenzie read the article avidly, and showed it to a lot of her school friends, but to her surprise, it didn't seem to stir up much interest in them. At first, her best friend, Kathi Hamilton, a very bright girl who had plans to go to State and major in history, enthusiastically joined Andrea and wrote quite a few cards each day. But after a few days, Kathi's enthusiasm began to wane.

"What's the point, Andrea?" Kathi asked. "A handful of postcards from me isn't going to make. the slightest bit of difference. Didn't you read Mrs. Doobar's article? There have got to be 160,000 a day to make a big difference."

"That's just the point, Kath!" replied Andrea exasperatedly. "If you and everyone else will just do your part, we'll reach that number-but you can't cop out!" Kathi didn't see the logic, and spent most of her time doing her homework for the summer school course in World History she was taking. After all, how could she get into State if she flunked World History?

Andrea just couldn't figure out how come Kathi, of all people, so

interested in history and the flow of time and world-events, could not see her own life being touched by such factors, so she asked Kathi, "How do you know there will be any you-left to go to State, if you don't write postcards? Each year, there's a 1-in-12 chance of you and me and all of us being wiped out! Don't you even want to work against that? If people would just care, they could change things! An hour a day! Half an hour a day! Fifteen minutes a day!"

"Oh, come on, Andrea!" said Kathi annoyedly, "Be realistic."

"Darn it all, I'm the one who's being realistic", said Andrea. "If you don't help out, you're adding to the burden of someone else."

"For Pete's sake, Andrea", Kathi protested angrily, "I'm not adding to anyone else's burden. Everyone can help out as much as they want, and no one's obliged to do anything at all. Sure, I'd like it if everyone were helping, but you can see for yourself, practically nobody is. So I'm not going to waste my time. I need to pass World History."

And sure enough, Andrea had to do no more than listen each hour, right on the hour, to hear that bell ring to realize that nobody was doing much. It once had sounded so pleasant and reassuring, and now it sounded creepy and ominous to her, just like the fireflies and the barbecues. Those fireflies and barbecues really bugged Andrea, because they seemed so normal, so much like any other summer-only this summer was not like any other summer. Yet nobody seemed to realize that. Or, ' rather, there was an undercurrent that things were not quite as they should be, but nothing was being done ...

. One Saturday, Mr. Hobbs, the electrician, came around to fix a broken refrigerator at the McKenzies' house. Andrea talked to him about writing postcards to the Demon. Mr. Hobbs said to her, "No time, no time! Too busy fixin' air conditioners! In this heat wave, they been breakin' down all over town. I-work a 10-hour day as it is, and now it's up to 11, 12 hours a day, includin' weekends. I got no time for postcards, Andrea." And Andrea .saw that for Mr. Hobbs, it was true. He had a big family and his children went to parochial school, and he had to pay for them all, and ...

Andrea's older sister's boyfriend, Wayne, was a star halfback at Happiton High. One evening he was over and teased Andrea about her postcards. She asked him, "Why don't you write any, Wayne?".

"I'm out lifeguardin' every day, and the rest of the time I got scrimmages - for the fall season."

"But you could take some time out just 15 minutes a day-and write a few postcards!" she argued. He just laughed and looked a little fidgety. "I don't know, Andrea",-he said. "Anyway, me 'n Ellen have got better things to do-huh, Ellen?" Ellen giggled and blushed a little. Then they ran out of the house and jumped into Wayne's sports car to go bowling at the Happi-Bowl.

*** * ***

Andrea was puzzled by all her friends' attitudes. She couldn't understand why everyone had started out so concerned but then their concern had fizzled,, as if the problem had gone away. One day when she was walking home from school, she saw old Granny Sparks out watering her garden. Granny, as everyone called her, lived kitty-corner from the McKenzies and was always chatty, so Andrea stopped and asked Granny Sparks what she thought of all this. "Pshaw! Fiddlesticks!" said Granny indignantly. "Now Andrea, don't you go around believin' all that malarkey they print in the newspapers! Things are the same here as they always been. I oughta know -I've been livin' here nigh on 85 years!

Indeed, that was what bothered Andrea. Everything seemed so annoyingly normal. The teenagers with their cruising cars and loud motorcycles. The usual boring horror movies at the Key Theater down on the square across from the courthouse. The band in the park. The parades. And especially, the damn fireflies! Practically nobody seemed moved or affected by what to her seemed the most overwhelming news she'd ever heard. The only other truly sane person she could think of was little Wally Thurston, that elevenyear-old from across town. What a ridiculous irony, that an eleven-year-old was saner than all the adults!

Long about August 1, there was an editorial in the paper that gave Andrea a real lift. It came from out of the blue. It was written by the paper's chief editor, "Buttons" Brown. He was an old-time journalist from St. Jo, Missouri. His editorial was real short. It went like this:

The Disobedi-Ant

> The story of the Disobedi-Ant is very short. It refused to believe that its powerful impulses to play instead of work were anything but unique expressions of its very unique self, and it went its merry way, singing, "What I choose to do has nothing to do with what any-ant else chooses to do! What could be more self-evident?"
>
> Coincidentally enough, so went the reasoning of all its colony-mates. In fact, , the same refrain was independently invented by every last ant in the colony, and each ant thought it original. It echoed throughout the colony, even with the same melody.
>
> The colony perished.

Andrea thought this was a terrific allegory, and showed it to all her friends. They mostly liked it, but to her surprise, not one of them started writing postcards.

All in all, folks were pretty much back to daily life. After all, nothing much - seemed really to have changed. The weather had turned real hot, and folks congregated around the various swimming pools in town. There were lots of barbecues in the evenings, and, every once in a while somebody'd make a joke or two about the Demon and the postcards. Folks would chuckle and

then change the topic. Mostly, people spent their time doing what they'd always done, and enjoying the blue skies. And mowing their lawns regularly, since they wanted the town to look nice.

Post Scriptum

The atomic bomb has changed everything
except our way of thinking. And so we
drift helplessly towards unparalleled disaster.
-Albert Einstein

People of every era always feel that their era has the severest problems that people have ever faced. At first this sounds silly. How can every era be the toughest? But it's not silly. Things can be getting constantly more dangerous and frightful, and that would mean that each new generation truly is facing unprecedentedly serious problems. As for us, we have the problem of extinction on our hands.

Someone once said that our current situation vis-a-vis the Soviet Union is like two people standing knee-deep in a room filled with gasoline. Both hold open matchbooks in their hands. One person is jeering at the other:.. "Ha ha ha! My matchbook is full, and yours is only *half* full! Ha ha ha!"

The reality of our situation is about that simple. The vast majority of people, however, refuse to let this reality seep into their systems and change their day-to-day behaviors. And thus the validity of Einstein's gloomy utterance.

* * *

I remember many years ago reading an estimate that the famous geneticist George Wald had made about nuclear war. He said he figured there was a two percent chance per year of a nuclear war taking place. This amounts to throwing one 50-sided die (or a couple of seven-sided dice) once a year, and hoping that it doesn't come up on the bad side. How Wald arrived at his figure of two percent per year, I don't know. But it was vivid. The figure has stuck with me for a couple of decades. I tend to think that the chances are greater nowadays than they were back then: maybe about five percent per year. But who can say?

The *Bulletin of the Atomic Scientists* features a clock on its cover. This clock doesn't tick, it just hovers. It hovers near midnight, sometimes getting closer, sometimes receding a bit. Right now, it's at three minutes to midnight. Back at the signing of SALT I, it was at twelve minutes before midnight. The closest it ever came was two minutes before midnight, and I think that was at the time of the Cuban missile crisis.

The purpose of the clock is to symbolize the current danger of a nuclear

holocaust. It's a little like those "Danger of Eire Today" signs that Smokey the Bear holds up for you as you enter a national forest in the summer. It is a subjective estimate, made by the magazine's board of directors. Now what is the meaning of "danger", if not probability of disaster per unit time? Surely, the more dangerous a place or situation, the faster you want to get out of it, for, just that reason. Therefore, it seemed to me that the Bulletin's number of minutes before midnight, B, was really a coded way of expressing a Wald number, W-a probability of nuclear war per year. And so I decided to make a subjective table, matching up the values of B that I knew about with my own best estimates of W. After a bit of experimentation, I came up with the following table:

Bulletin Clock (minutes before midnight)	Wald's percentage (probability per year)
1 min	20 percent
2 mins	10 percent
3 rains	7 percent
4 mins	5 percent
5 mins	4 percent
7 mins	3 percent
10 mins :	2 percent
12 mins	1.5 percent
20 mins	1 percent

A fairly accurate summary of this subjective correspondence is given by the following simple equation:

$$W = 20/B$$

This estimates for you the holocaust danger per orbit of the earth, as a function of the current setting of the Bulletin's clock.

W and B may not be estimable in any truly scientific way, but there is a definite reality behind them, even if not. so simple as that of N and X in Happiton. Obviously it is not a "random'." dicelike process that will determine whether nuclear war erupts in any given year. Nonetheless, it makes good sense to think of it in terms of a probability per year, since what actually does determine history is a lot of things that are in effect random, from the point of view of any less-than-omniscient being. What other people (or countries) do is unpredictable and uncontrollable: it might as well be random.

If tensions get unbearably high in the Middle East or in Central America, that is not something that we could have predicted or forestalled. If some terrorist group manufactures and uses or threatens to use -a nuclear bomb, that is essentially a "random" event. If overpopulation in Asia or starvation

In Africa or crop failures in the Soviet Union or oil gluts or shortages create huge tensions between nations, that is like a random variable, like a throw of dice. Who could have predicted the crazy flareup between Britain and Argentina over the silly Falkland Islands? Who knows where the next hot spot will turn out to be? The global temperature can change as swiftly and capriciously as a bright summer day can turn sultry and menacing-even in Happiton.

* * *

It is the vivid imagery behind the Wald number and the Bulletin clock that first got me thinking in terms of the Happiton metaphor. The story was pretty easy to write, once the metaphor had been concocted. I had to work out the mathematics as I went along, but otherwise it flowed easily. It was crucial to me that the numbers in. the allegory seem realistic. The most important numbers were: (1) the chance of devastation per year, which came out about right, as I see it; and (2) the amount of time per day that I think would begin to make a significant difference if devoted by a typical person to some sort of activity geared toward the right ends. In Happiton, that threshold turned out to be about fifteen minutes per day per person. Fifteen minutes a day is just about the amount of time that I think would begin to make a real difference in the real world, but there are two ways that one might draw a distinction between the situation in Happiton and the actual case.

Firstly, some people say that the situation in Happiton is much simpler than that of global competition and potential nuclear war. In Happiton, it's obvious that writing postcards will do some good, whereas it's not so obvious (they claim) what kind of action will do any good in the real world. Working hard for a freeze or for a reduction of US-SU tensions might even be harmful, they claim! The situation is so complex that nothing corresponds to the simplistic and sure-fire recipe of writing postcards.

Ah, but there is a big fallacy here. Writing postcards in Happiton is not sure-fire. The gas could still come oozing up at any time. All that changes is the odds. Now in the real world, we must follow our own best estimates, in the absence of perfect information, as to what actions are likely to be positive andd what ones to be negative. You can only follow your nose. You can never be sure that any action, no matter how well intended, is going to improve. the situation. That's just the way life is.

I happen to believe that the odds of a holocaust will be reduced (perhaps by a factor of 1.0000001-) by writing to my representatives and senators fairly regularly, by attending local freeze meetings, by contributing to various organizations, by giving lectures here and there on the topic, and by writing articles like this. How can I know that it will do any good? I can't, of course. And it's no different in Happiton. The best of intentions can backfire for totally unforeseeable reasons. It might turn out that little Wally Thurston, by moving his pencil in a certain graceful curlicue motion one

afternoon while writing his 1,000th postcard to the Demon, stirs up certain air molecules which, by bouncing and jouncing against other ones helter-skelter, wind up giving that tiny last push to the caroming icosahedral dice atop the belfry, and bang! They all come up '7'! Wally, oh Wally, why such folly? Why did you ever write those postcards?

Those who would caution people that it might be counter-productive to work against the arms race-unless they believe one should work for the arms race-are in effect counseling paralysis. But would they do so in other areas of life? You never know if that car trip to the grocery store won't be the last thing you do in your life. All life is a gamble.

The second distinction between Happiton and reality is this. In Happiton, for fifteen minutes a day to make a noticeable dent, it would have had to be donated by all 20,000 citizens, adults and children. Obviously I do not think that is realistic in our country. The fifteen minutes a day per person that I would like to see spent by real people in this country is limited to adults (or at least people of high-school age), and I don't even include most adults in this. I cannot realistically hope that everyone will be motivated to become politically active. Perhaps a highly active minority of five percent would be enough. It is amazing how visible and influential an articulate and vocal minority of,that size can be! So, being realistic, I limit 'my desires to an average of fifteen minutes of activity per day for five percent of the adult American population. I sincerely believe that with about this much work, a kind of turning point would be reached-and that at 30 minutes or 60 minutes per day (exactly as in Happiton), truly significant changes in the national mood (and hence in the global danger level) could be effected.

* * *

I think I have explained what Happiton was written for. Trigger activity it may not. I'm growing a little more realistic, and I don't expect much of anything. But I would like to understand human nature. better, to understand what it is that makes us so much like stupid gnats dully buzzing above a freeway, unable to see the onrushing truck, 100 yards down the road, against whose windshield we are about to be smashed.

One last thought: Although to me it seems that nuclear war is the gravest threat before us, I would grant that to other people it might appear otherwise. I don't care so much what kinds of efforts people invest their time in, as long as they do something. The exact thing that corresponds to the threat to Happiton doesn't much matter. It could be nuclear weapons, chemical or biological weapons, the population explosion, the U.S.'s ever-deepening involvement in Central America, or even something more contained, like the environmental devastation inside the U.S. What it seems to me is needed is a healthy dose of indignation: a spark, a flame, a fire inside. Until that happens, that courthouse clock'll be tickin' away, once every hour, on the hour, until ...

Post Post Scriptum.

Two magazines are devoted to the prevention of nuclear war. They are: the Bulletin of the Atomic Scientists and Nuclear Times. The Bulletin, founded in 1945, aims to forestall nuclear holocaust by promoting awareness and understanding of the issues involved. It describes itself as "a magazine of science and world affairs". Its address is: 5801 South Kenwood Avenue, Chicago, Illinois 60637.

Nuclear Times is a more recent arrival, and calls itself "the news magazine of the antinuclear weapons movement". Its articles are shorter and lighter than those of the Bulletin, but it keeps you up to date on what's happening all over the country and the world. Its address is: Room 512, 298 Fifth Avenue, New York, New York 10001.

The following organizations are effective and important forces in the attempt to slow down the arms race and to reduce global tensions. Most of them put out excellent literature, which is available in, large quantities at low prices (sometimes free) for distribution. Needless to say, they can always use more members and more funding. Many have local chapters.

The Council for a Livable World
11 Beacon Street Boston,
Massachusetts 02108

SANE
711 G Street, S.E. Washington,
D.C. 20003

Nuclear Weapons Freeze Campaign
4144 Lindell Boulevard, Suite 404
St. Louis, Missouri 63108

The Center for Defense Information
303 Capital Gallery West
600 Maryland Avenue, S.W.
Washington, D.C. 20024

Physicians for Social Responsibility,
639 Massachusetts Avenue
Cambridge, Massachusetts 02139

International Physicians for the Prevention of Nuclear War
225 Longwood Avenue, Room 200
Boston, Massachusetts 02115

Union of Concerned Scientists
1384 Massachusetts Avenue
Cambridge, Massachusetts 02238