Excerpts from Douglas Hofstadter's *Metamagical Themas* (a collection mostly of articles he wrote for *Scientific American*.)

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 nothing 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 matter what the dealer chooses to do*, I'm better off leaving an empty bag. So I'll leave an empty bag."

* * *

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 have 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 bag of what you need while giving up a full bag 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 points 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.

		Dealer: Cooperates	Dealer: Defects
You: Cooperate		(2 , 2)	(-1 , 4)
You: Defect		(4 , -1)	(0 , 0)

(a)

		Your Accomplice: Stay mum	Your Accomplice: Sings
You: Stay mum		(-2 , -2)	(-5 , 0)
You: Sing		(0 , -5)	(-4 , -4)

(b)

		Player B: Cooperates	Player B: Defects
Player A: Cooperate		(3 , 3)	(0 , 5)
Player A: Defect		(5 , 0)	(1 , 1)

(c)

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 commothhts or services, in which both participants have a choice: to cooperate (i.e., *to deliver the goodf 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. The 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.

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 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 mo you feel you *must* defect. Either of you would feel that way, of course, 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-lb. 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-ic. 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:

$$(1)\ T > R > P > S$$

$$(2)\ (\mathrm{T} + \mathrm{S}) / 2 < R$$

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 time. Now if each organism is Continually swimming around and bumping into the others, eventually each one will have met every 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 having gained more from 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 up 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 from an early draft of that book.)

Furthermore, he and evolutionary biologist William D. Hamilton have worked out an published many of the implications of these discoveries for evolutionary theory. Their work has won much acclaim—including the 1981 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 noncooperative 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 'D', 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 TAT* off guard. When it defects, *TIT FOR TAT* retaliates with a single defection, while *JOSS* "innocently" goes back 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 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 be 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 many cases this leads to the same type of breakdown of trust, whereas *TIT FOR TAT,* 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 TATS,* 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 to 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 TAT,* 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 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* reta1iation. 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 63 x 63 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 63 x 63 matrix of program vs. 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 result of the previous one. In particular if you take a program's score in 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—and 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 dynamical 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 rank 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, *HARRING TON* 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, *HARRLINGTON* 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 *TIT 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 player, 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 *RANDOM,* and it finished next to last. The last-place finisher's strategy was responsive, but its behavior was so inscrutable that it *looked* unresponsive. And in a more recent computer tournament conducted by Marek Lugo, and myself in the Computer Science Department at Indiana University three *ALL-D's* came in at the very bottom (out of 53), with a couple of *RANDOM*'s giving them a tough fight for the honor.

One way to explain *TIT FOR TAT'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 TAT.* Even rules that were designed to see

> what they could get away with quickly apologize to *TIT FOR TAT.* Any rule that tries to take advantage of *TIT FOR TAT* will simply 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 very 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 TAT;* in particular, he believes that *TIT FOR TAT* 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 *TIT FOR TAT* 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 TAT* 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 simply a little better than *TIT FOR TAT* 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 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 *ALL-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 TAT* 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 TAT* could be called "reflexive" (in 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 of individuals who rely on reciprocity*.

[…]

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 to get the oil changed. I know little about auto mechanics, so when I come 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?

[…]

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 yo name just the word "cooperate" or the word "defect". if two "cooperate, 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. The payoff 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, 11 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 $54! 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?
>
> […]

* * *

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 Prisoners 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 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 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 undergoing similar throes of reasoning agony will necessarily come up 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 now is, "Since we are all going to submit the same letter, which one 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, cooperating is preferred by this particular rational thinker. Since I am typical, cooperating must be preferred by *all* rational thinkers, So I 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 Cs. 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.

* * *

[…] 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 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 I?" 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 simulator 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 take into account was that cool-headed calculating people should take into account that cool-headed calculating people should take into account that coolheaded 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 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 representing D'ing 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 $1 + 3p - p^2$, which grows monotonically as *p* increases from *0* to *1*. Therefore, the optimum value of *p* 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 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 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)!

[…]

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 '1'. 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 *$l,000,000 / W* 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 enlighten 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 naïvely 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
10^{100} (a googol): 9
$(10^{10})^{100}$ (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 thousand-plus entries of '1' there was one that came from a lucky super-rational 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 shortsighted 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 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).

[…]

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 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 is optimum 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 pena1ized by society in the same way as society "allows" someone to rob a bank and 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 arrarngements 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 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: be 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 lit 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 types of gigantic social problems—Prisoner's Dilemmas, Tragedies of 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 Gödel sentences (and related ones) in mathematical logic. […] Those civilizations to which cooperation appears axiomatic—Type I societies— wind 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 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. […] 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 situation 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.