

# the prisoners' dilemma

Barry Krusch

© 1994, 2010 by Barry Krusch All rights reserved.

LAST UPDATED: January 16, 1995

The latest version of this document may be obtained at [www.krusch.com](http://www.krusch.com).

In 1994 I taught an 8th grade junior high school class, and every Monday we would play a game called “Who Wants It?”

The game worked this way: I would offer to give a dollar away to any person in the class, provided that that person was the only person who raised their hand when I asked “who wants it?” If more than one person raised their hand, I would drop down the offer a quarter each time, until the reward was *nada*.

Was I unnecessarily putting my hard-earned cash at risk? Hardly — in fact, for me the outcome was predetermined. I knew I’d be giving away little, if any, money. I knew this because I knew about the *Prisoners’ Dilemma*, a critically important topic in game theory. Just as I expected, week after week, the kids would compete with each other for the buck, past the quarter right on down to *nada*, and all would end up empty-handed. It never occurred to these poor souls to cooperate; it never occurred to them to get together and have one person raise their hand, and then split the cash. The look of dismay on these kids’ faces as they all watched each other raising their hands was palpable — they knew they were trapped in a cognitive web they did not have the intellectual capacity to escape. It was sad, but, I have to say, quite instructive.

Just a bunch of immature, silly junior high school students? Surely we big, smart adults would not be so foolish as to compete with each other in those situations where mutual cooperation would benefit us all! Uh . . . yeah. As Henry Hazlitt pointed out nearly fifty years ago in his book *A New Constitution Now*,

Men do not act in accordance with their interests; they act in accordance with their illusions. To know what one’s real interests are is an intellectual feat of which few men seem to be capable. ‘If all men acted from enlightened self-interest,’ as Bertrand Russell has put it, ‘the world would be a paradise in comparison with what it is.’<sup>1</sup>

Recognition of this problem is older than fifty years; in fact, it was originally described with some formal rigor in 1832 by William Forster Lloyd, a professor of political economy at Oxford University.<sup>2</sup> As Lloyd noted, cattle-owners have a short-term interest in increasing the size of their herds. Yet when cattle graze on common pasture, an “indefinite increase in the size of the herds sooner or later produces a number of animals that is far beyond the biological ‘carrying capacity’ of the pasture.”<sup>3</sup> This phenomenon is known as the *Tragedy of the Commons*, formulated by Hardin as follows:

---

<sup>1</sup> *A New Constitution Now*, p. 55.

<sup>2</sup> *Filters Against Folly*, p. 90.

<sup>3</sup> *Filters Against Folly*, p. 91.

Imagine yourself as a herdsman . . . when the total population of herd animals has just reached the carrying capacity of the land. Suppose you have a chance to acquire ten more animals. Suppose also that you are in complete possession of the facts—that you understand carrying capacity and the dangers of transgressing it. Should you, or should you not, add ten more animals to your herd?

Since the additional animals are (by hypothesis) ten more than the carrying capacity, all your animals will have a little less food per capita next year than this. So will everyone else's animals. Even so, you expect a net gain from the acquisition, for this reason: the gain is all yours, but the loss (from transgressing the carrying capacity) is shared among all the herdsman. Your share of the loss is only a small fraction of the total. Balancing *your* gain against *your* loss you decide to take on ten more animals. In economics this is called a rational decision. To behave otherwise would be to behave irrationally—in the short run.

Every other herdsman in a commons must, if rational, reach the same decision—not only this year but in every succeeding year. In the long run this kind of behavior produces disaster for all, as overgrazing turns semidesert into desert. Even if you understand completely the disastrous consequence of living by the rules of the commons, you are unable to behave otherwise. The rules pay you to do the wrong thing.

As a good citizen you might refuse to add to your herd, but what makes you think every other herdsman would also be a good citizen? . . . As selfish and rational exploiters prosper at the expense of the public-spirited, envy will cause some of the latter to join the 'rational' decision makers in their ruinous behavior. What might begin as the selfish rationalism of a few, ends in the corruption of the many.<sup>4</sup>

The Fall of Man. As time progresses, poetic observations become prosaic; our subconscious insights make their way from dusk to dawn. Lloyd's 19th-century observation that people do not act in accordance with their long-term interests turned out to be a very telling one, and one that grew concomitantly in importance as society "progressed", and nuclear weapons were invented. With this new importance came new realization.

After World War II, the potential of the superpowers to turn the entire planet into the Sahara led to extensive research into game theory at Defense Department "think tanks". Eventually, a concept known as the *Prisoner's Dilemma* was discovered in 1950 by Melvin Dresher and Merrill Flood of the RAND corporation. Since their original formulation of the problem is "less clear to the uninitiated", Hofstadter (1985) developed a parallel example in one of his two superb articles on the *Prisoners' Dilemma* anthologized in his book *Metamagical Themas*:

---

<sup>4</sup> *Filters Against Folly*, pp. 92-93.

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 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."

The dealer, meanwhile, being in more or less the same boat (though at the other end of it), thinks analogous thoughts and comes to the parallel conclusion that it is best to leave an empty bag. And so both of you, with your impeccable (or impeccable-seeming) logic, leave empty bags, and go away empty-handed.<sup>5</sup>

The original example as formulated by Dresher and Flood is closer to the following<sup>6</sup>:

You and a man named Jack are suspected of having committed an armed robbery, and you are each placed in separate jails, with no means of communication. Some hours later, a District Attorney enters your cell. You are told that there is enough evidence to convict both you and Jack on a lesser charge of illegal possession of firearms, but not enough to convict on the more serious charge of armed robbery. To avoid a lengthy trial, you are given a chance to confess, under the following conditions:

1. If *neither* you nor Jack confess, you will both be convicted of illegal possession, which carries a sentence of **six months**.
2. If *both* you and Jack confess, you will both get the MINIMUM sentence for armed robbery, which is **two years**.
3. If *only one* of you confesses, that person will be considered a state witness,

---

<sup>5</sup> *Metamagical Themas*, pp. 715-716.

<sup>6</sup> The following discussion is based on Watzlwick's formulation of Albert Tucker's description, in *How Real is Real?*, p. 98.

and

**go free**; the other will get the **MAXIMUM** sentence for armed robbery, which is **twenty years**.

To clarify this matrix of possibilities, you decide to construct the following table, which you organize from the *fewest* months you could possibly serve to the *most* months you could possibly serve:

CONFESSION POSSIBILITIES	POSSIBLE MONTHS THAT WILL BE SERVED GIVEN THE POSSIBILITIES	
	ME	JACK
I CONFESS, JACK DOESN'T	0	240
NOBODY CONFESSES	6	6
WE BOTH CONFESS	24	24
I DON'T CONFESS, JACK DOES	240	0

Before you looked at this table, the answer seemed simple; neither you nor Jack should confess, thus limiting the time served by both of you to six months. But then a thought enters your mind: “Jack must figure I don’t want to confess.” You decide to scan the table, and your eyes drop to the last row. Uh oh. “Whoa — if I don’t confess, and he does, he gets off *scot-free* — no months in jail. Hmmm — that’s a powerful incentive for him to confess!” Then you look at the first row of the table. “Uh oh, there’s an even *worse* incentive for him to give it up — my *own* incentive to confess! How can Jack trust *me* when I can get off scot-free by confessing?” Your mind is slowly changing, Dave, I can feel it. You decide to get even more sophisticated by putting this analysis in a new table which contains the average results of the opposing strategies:

	I CONFESS	I DON'T CONFESS
HE CONFESSES	24	240
HE DOESN'T CONFESS	0	6
AVERAGE RESULTS	12	123

Yup, looks pretty clear. The average sentence you can serve with a *confess* strategy is **12** months, while the average sentence from a *don't confess* strategy is **123** months. Clear as day. So, you confess. And Jack, who knows you’re an analytical thinker just like him, chooses likewise. You both end up serving two years, when simple cooperation would have reduced both your sentences by 75 percent. Thus operates the “logic” of individualism embedded in a *Prisoners’ Dilemma* scenario. (By the way, an interesting observation: your guilt or innocence with reference to the charge is *irrelevant*: even if innocent, you must confess! The Salem Witch Trials and plea bargaining come to mind here).

Of course, the problem isn’t in the logic—the problem is in the *premises*. Note the problem of the prisoners:

1. They could not communicate with each other.
2. Even if able to communicate, they could not necessarily trust each other, and
3. They could not trust each other, because
  - a) they came from a society whose culture encouraged the free-flowering of egocentrism, and
  - b) there was no arm of enforcement preventing those individual acts which betrayed the long-term interests of all.

But, as one might expect, the Prisoner's Dilemma is not merely confined to hapless suspects in the lock-up. Locks, chess clocks, cops, umpires, steroids, shoplifting scanners, two-way mirrors, video surveillance, parking meters, QWERTY keyboards, radar detectors and toll booths are contemporary monuments to the existence of this phenomenon. Consider these notorious examples from real-life:

- A club has a fire, and all rush for the exits, preventing the exit of anyone; as a result, all perish.
- At a concert, A stands on his toes to better see the performer. The person behind A must now stand, and the effect ripples throughout the auditorium. Soon all are standing, and no one has a better view than they would have had in a sitting position, except that now they must stand versus sit.
- Big kids "pick on" little kids, and society allows it ("you have to learn how to defend yourself"). Thus, a kid must "act tough" to develop a "rep," and avoid being singled out. Soon kids become teens become adults, and fists become knives become guns, and gangs become organized gangs become organized crime. Eventually the "big kids" are confronted by even bigger and badder kids.
- The phenomenon of military escalation engendered by a genuine need to defend:

"If one nation maintains constantly a disciplined army, ready for the service of ambition or revenge, it obliges the most pacific nations who may be within the reach of its enterprises to take corresponding precautions."<sup>7</sup>
- Athlete A uses steroids, which gives him a competitive advantage. Other

---

<sup>7</sup> *Federalist 41*, Madison.

- athletes are forced to use steroids to retain parity. As a result, no athlete is given a competitive advantage, but all are subjected to the hazards of steroids.<sup>8</sup>
- Model A gets breast implants, which help her get jobs. Other models are forced to get breast implants to retain parity. As a result, no model is given a competitive advantage, but all are subjected to the hazards of breast implants.
  - In a college dorm, A blares his stereo. To hear his stereo, B must blare his. Soon the whole floor is awash in a cacophony of noise as a “stereo war” develops.
  - A says, “never take the first offer.” Soon B says it, and eventually C. Gradually, society becomes filled with people who refuse to take the first offer, forcing those who otherwise wished to make honest first offers to become disingenuous, leading to a society with radically diminished trust.
  - A depositor hears that a bank is in trouble, and goes to pull out his savings. Others are forced to pull out their savings as a “run” begins on the bank’s savings, and the bank collapses.
  - State X offers a lottery, and soon the citizens of State Y are sending money out of the state to state X. State X, which did not wish to have a lottery because it felt that lotteries had a corrupting effect on the citizenry, is forced, out of financial necessity, to have a lottery.
  - Mr. A pirates software. So does Mr. B. So does Mr. C. As a consequence, the software that would have been otherwise written for the benefit of Messrs. A, B, and C is not written, because there is a radically diminished market for software.
  - Mr. Z writes a “virus,” which destroys the operating system of people’s computers. Mr. Z thus provides a role model for others, who also write viruses, one of which attacks Mr. Z’s computer and destroys his data.
  - Car companies institute a policy of “planned obsolescence” due to their control of the automobile market. Foreign automobile manufacturers are given an opportunity to enter the market, and the domestic car companies lose market share, and American workers lose jobs.
  - College student X (and I do mean “X”) rips out the pages from a book a professor has put on reserve, and other students follow suit; that’s why X now can’t get the article he needs.
  - In a public debate where ready access to the facts is not available, X misstates

---

<sup>8</sup> NYT, 7/28/91, p. S-1.

the truth, and the audience sees X's viewpoint as more legitimate, forcing Y to misstate the truth to retain parity. The presence of counterfeit information delegitimizes genuine information.

- Congressman X votes to keep a military base open to improve his chances for election, even though it will hurt the country by adding to the deficit. Voters vote for a representative who “brought jobs” to the district, even though the policy will have a devastating effect on the national economy.

In *Metamagical Themas*, Hofstadter gives additional examples of the working-out of the *Prisoners' Dilemma* in everyday life, in order of seriousness:

- loudly wafting your music through the entire neighborhood on a fine summer's day;
- 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 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.’<sup>9</sup>

While this last example is, in the final analysis, the most serious, its effects are harder to see. The driving example is probably the one most likely to be confronted on a daily basis, and the one which is the most visible. It's probably no surprise, in a country which has such a love affair with the automobile, that the Prisoner's Dilemma situations which most frequently confront us are found in *traffic*. Among these are the following:

- drivers who don't wear seat belts, thus driving up the cost of everyone else's insurance;
- people who avoid driving small cars because big cars are safer, putting the people in small cars at greater risk, which forces them, in turn, to buy large cars;
- the person who drives “gas-guzzlers” because he “likes” big cars, giving others permission to do what they like as well, increasing America's dependence on foreign oil, which increases the price of gasoline as well as

---

<sup>9</sup> *Metamagical Themas*, p. 757.

- increasing the risk of war to secure scarce petroleum resources;
- the phenomenon of *rush-hour traffic*, the congestion of streets, roads, and highways at designated times, where every person's individual desire to get to work *promptly* makes everybody *late*;
  - the pollution which results from the collective actions of individual drivers who believe that their one act of driving "doesn't matter," resulting in air no one wants to breathe;
  - those drivers who go to the front of long lines at freeway exits and "butt in," forcing other drivers to retaliate or "take it";
  - and finally, *gridlock*, the total cessation of traffic flow that results from clogged intersections. Gridlock is preceded by a phenomenon known to traffic engineers as *spillback*, which results when drivers move into an intersection with no place to go, and thus block the other drivers from passing through. Each driver who "spills back" hates being blocked himself, but blocks others to save a second's worth of traveling time.

As the above situations clearly indicate, the rudeness of driver A means that driver B has to be equally rude to secure his or her rights, and the rudeness begins to *escalate*. Localities without adequate sanctions against these acts are transformed into *dog-eat-dog* cultures. Says Hofstadter:

I have been struck by the relative savagery of the driving environment in the Boston area. I know of no other city in which people are so willing to take the law into their own hands, and to create complete anarchy. There seems to be less respect for such things as red lights, stop signs, lines in the street, speed limits, other people's cars, and so forth, than in any other city, state, or country that I have ever driven in. This incessant "me-first" attitude seems to be a vicious, self-reinforcing circle. Since there *are* so many people who do whatever they want, nobody can afford to be polite and let other people in ahead of them (say), for then they will be taken advantage of repeatedly and will wind up losing totally.<sup>10</sup>

Boston is not alone. In some cities, such as Los Angeles, frustrated drivers have been known to arm themselves, and shoot others.

Well, junior high school students and normal adults are *Prisoners' Dilemma* prey, but surely the best among us can rise above the fray. Right? Wrong. Out there in the real world is a gloomy illustration of the depths of irrationality to which our best-educated individuals can sink. In his June, 1983 column in *Scientific American*, Hofstadter announced a lottery not unlike the "Who Wants

---

<sup>10</sup> *Metamagical Themas*, pp. 732-733.

It?" game. In that lottery, the prize to be awarded was \$1,000,000 divided by the number of entries received; so, if 1,000,000 entries were received, and your name was picked, you would win \$1.00. Obviously, the rational behavior<sup>11</sup> for the purchasers of that issue of *Scientific American* would have been to designate one subscriber to enter one time, with the others holding out. After winning, that subscriber would then divide up the money among all who cooperated. Thus, if there were 100,000 cooperating subscribers, each would have won \$10.00. Since the readers of *Scientific American*, of all people, could be expected to be more rational than the rest of us, they would be presumably be the most likely to discover the most rational solution. *Presumably*. But, as Hofstadter reported,

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 . . . 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 . . . meanwhile, all this monumental effort [was] to the detriment of *everyone*.<sup>12</sup>

Perhaps now we can see why the problem was formulated in terms of prisoners; in society, individuals are the *prisoners* of their own (collective) beliefs. Individual Heaven is Communal Hell; our problem is that we are surrounded by people just like us!

Is the problem hopeless? No. *Prisoners' Dilemma* theory provides at least three ways out:

1. Enable *communication* to enable the formation of *culture*, to
2. Establish *social codes* which combat *Prisoners' Dilemma* effects, while simultaneously
3. Creating a cultural organization, formal or informal, that *enforces* these social codes. (Best type is "bottom-up" as opposed to "top-down").

How fortunate the prisoners would have been if they had had at least the first two options open to them all their lives. Consequently, both would have cooperated, to their mutual advantage. *The ultimate act of individualism is to*

---

<sup>11</sup> Assuming that the acquisition of wealth is always rational.

<sup>12</sup> *Metamagical Themas*, p. 760 (paragraphs combined).

*implement a culture that operates in the background to curb irrational exercises of individuality.* The cure for the Prisoner's Dilemma disease has been known for centuries — in the old days, it was called *ethics*, as expressed in the *community*. In the old days, the expression was “the whole society raises the child.” Morality was enforced in a small way on a daily basis; and by “nipping immorality in the bud”, we avoided the larger acts of immorality which grew out of the small acts.

In the modern era, however, we have gotten away from the sense of *community*. Who talks to their neighbor anymore? And talks to *strangers*? — well, let's just say we do less of it now than we used to.

Alas, the cost of modern society is a cost that, like all costs, must be paid. And the cost, as technology rises, is becoming increasingly harder to pay. We need only look around at society, where the costs unpaid are all around us.

It may be that we may have to re-evaluate just where we're headed if we keep on going down the same path, and may have to explore new paradigms that will help us solve this *Prisoners' Dilemma*.