

# The Rules of the Game (Carl Sagan)

Everything morally right derives from one of four sources: it concerns either full perception or intelligent development of what is true; or the preservation of organized society, where every man is rendered his due and all obligations are faithfully discharged; or the greatness and strength of a noble, invincible spirit; or order and moderation in everything said and done, whereby is temperance and self-control.

C I C E R O,  
De Officiis, I, 5 (45-44 B.C.)

I remember the end of a long ago perfect day in 1939—a day that powerfully influenced my thinking, a day when my parents introduced me to the wonders of the New York World's Fair. It was late, well past my bedtime. Safely perched on my father's shoulders, holding onto his ears, my mother reassuringly at my side, I turned to see the great Trylon and Perisphere, the architectural icons of the fair, illuminated in shimmering blue pastels. We were abandoning the future, the "World of Tomorrow," for the BMT subway train. As we paused to rearrange a tray around his neck. He was selling pencils. My father reached into the crumpled brown paper bag that held the remains of our lunches, withdrew an apple, and handed it to the pencil man. I let out a loud wail. I disliked apples then, and had refused this on both at lunch and at dinner. But I had, nevertheless, a proprietary interest in it. It was my apple, and my father had just given it away to a funny-looking stranger — who, to compound my anguish, was now glaring unsympathetically in my direction.

Although my father was a person of nearly limitless patience and tenderness, I could see he was disappointed in me. He swept me up and hugged me tight to him.

"He's a poor stiff, out of work," he said to me, too quietly for the man to hear. "He hasn't eaten all day. We have enough. We can give him an apple."

I reconsidered, stifled my sobs, took another wishful glance at the World of Tomorrow, and gratefully fell asleep in his arms.

\*\*\*

Moral codes that seek to regulate human behavior have been with us not only since the dawn of civilization but also among our pre-civilized, and highly social, hunter-gatherer ancestors. And even earlier. Different societies have different codes. Many cultures say one thing and do another. In a few fortunate societies, an inspired lawgiver lays down a set of rules to live by (and more often than not claims to have been instructed by a god — without which few would follow the prescriptions). For example, the codes of Ashoka (India), Hammurabi (Babylon), Lycurgus (Sparta) and Solon (Athens), which once held sway over mighty civilizations, are today largely defunct. Perhaps they misjudged human nature and asked too much of us. Perhaps experience from one epoch or culture is not wholly applicable to another.

Surprisingly, there are today efforts - tentative but emerging - to approach the matter scientifically; i.e., experimentally.

In our everyday lives, as in the-momentous affairs of nations, we must decide: What does it mean to do the right thing? Should we help a needy stranger? How do we deal with an enemy? Should we ever take advantage of someone who treats us kindly? If hurt by a friend, or helped by an enemy, should we reciprocate in kind; or does the totality of past behavior outweigh any recent departures from the norm.

Examples: Your sister-in-law ignores your snub and invites you over for Christmas dinner. Should you accept? Shattering a four-year-long worldwide voluntary moratorium, China resumes nuclear weapons testing; should we? How much should we give to charity? Serbian soldiers systematically rape Bosnian women; should Bosnian soldiers systematically rape Serbian women? After centuries of oppression, the National Party leader F. W. de Klerk makes overtures to the African National Congress; should Nelson Mandela and the ANC have reciprocated? A coworker makes you look bad in front of the boss; should you try to get even? Should we cheat on our income tax returns? If we can get away with it? If an oil company supports a symphony orchestra or sponsors a refined TV drama, ought we to ignore its pollution of the environment? Should you cheat at cards? On a larger scale: Should we kill killers?

In making such decisions, we're concerned not only with doing right but also with what works—what makes us and the rest of society happier and more secure. There's a tension between what we call ethical and what we call pragmatic. If, even in the long run, ethical behavior were self-defeating, eventually we would not call it ethical, but foolish. (We might even claim to respect it but ignore it in practice.) Bearing in mind the variety and complexity of human behavior, are there any simple rules—whether we call them ethical or pragmatic—that actually work?

How do we decide what to do? Our responses are partly determined by our perceived self-interest. We reciprocate in kind or act contrary because we hope it will accomplish what we want. Nations assemble or blow up nuclear weapons so other countries won't trifle with them. We return good for evil because we know that we can thereby sometimes touch people's sense of justice, or shame them into being nice. But sometimes we're not motivated selfishly. Some people seem just naturally kind. We may accept aggravation from aged parents or from children, because we love them and want them to be happy, even if it's at some cost to us. Sometimes we're tough with our children and cause them a little unhappiness, because we want to mold their characters and believe that the long-term results will bring them more happiness than the short-term pain.

Cases are different. Peoples and nations are different. Knowing how to negotiate this labyrinth is part of wisdom. But bearing in mind the variety and complexity of human behavior, are there some simple rules, whether we call them ethical or pragmatic, that actually work? Or maybe we should avoid trying to think it through and just do what feels right. But even then how do we determine what "feels right"?

\*\*\*

The most admired standard of behavior, in the West, at least, is the Golden Rule, attributed to Jesus of Nazareth. Everyone knows its formulation in the first-century Gospel of St. Matthew: **Do unto others as you would have them do unto you.** Almost no one follows it. When the Chinese philosopher Kung-Tzu (known as Confucius in the

West) was asked in the fifth century B.C. his opinion of the Golden Rule, of repaying evil with kindness, he replied, "Then with what will you repay kindness?" Shall the poor woman who envies her neighbor's wealth give what little she has to the rich? Shall the masochist inflict pain on his neighbor? The Golden Rule takes no account of human differences. Are we really capable, after our cheek has been slapped, of turning the other cheek so it can be slapped? With a heartless adversary, isn't this just a guarantee of more suffering?

The Silver Rule is different: **Do not do unto others what you would not have them do unto you.** It also can be found worldwide, including, a generation before Jesus, in the writings of Rabbi Hillel. The most inspiring twentieth-century exemplars of the Silver Rule are Mohandas Gandhi and Dr. Martin Luther King Jr. They counseled oppressed peoples not to repay violence with violence, but not to be compliant and obedient either. Nonviolent civil disobedience was what they advocated—putting your body on the line and showing, by your willingness to be punished in defying an unjust law, the justice of your cause. They aimed at melting the hearts of their oppressors (and those who had not yet made up their minds).

King paid tribute to Gandhi as the first person in history to convert the Gold or Silver Rules into an effective instrument of social change. And Gandhi made it clear where his approach came from: "I learnt the lesson on nonviolence from my wife, when I tried to bend her to my will. Her determined resistance to my will on the one hand, and her quiet submission to the suffering of my stupidity involved on the other, ultimately made me ashamed of myself and cured me of my stupidity in thinking that I was born to rule over her."

Nonviolent civil disobedience has worked notable political change in this century—in prying India loose from British rule and stimulating the end of classic colonialism worldwide, and in providing some civil rights for African-Americans—although the threat of violence by others, however disavowed by Gandhi and King, may have also helped. The African National Congress (ANC) grew up in the Gandhian tradition. But by the 1950's it was clear that nonviolent noncooperation was making no progress whatever with the ruling white Nationalist Party. So in 1961 Nelson Mandela and his colleagues formed the military wing of the ANC, the Umkhonto we Sizwe, the Spear of the Nation, on the quite un-Gandhian grounds that the only thing whites understand is force.

Even Gandhi had trouble reconciling the rule of nonviolence with the necessities of defense against those with less lofty rules of conduct? "I have not the qualifications for teaching the philosophy of life. I have barely qualifications for practicing the philosophy I believe. I am but a poor struggling soul yearning to be . . . wholly truthful and wholly nonviolent in thought, word and deed, but ever failing to reach the ideal.

"Repay kindness with kindness," said Confucius, "but evil with justice." This might be called the Brass or Brazen Rule: **Do unto others as they do unto you.** It's the *lex talionis*, "an eye for an eye, and a tooth for a tooth," *plus* "one good turn deserves another." In actual human (and chimpanzee) behavior it's a familiar standard. "If the enemy inclines toward peace, do you also incline toward peace," President Bill Clinton quoted from the Qur'an at the Israeli-Palestinian peace accords. Without having to appeal

to anyone's better nature, we institute a kind of operant conditioning, rewarding them when they're nice to us and punishing them when they're not. We're not pushovers, but we're not unforgiving either. It sounds promising. Or is it true that "two wrongs don't make a right?"

Of baser coinage is the Iron Rule: **Do unto others as you like, before they do it unto you.** It is sometimes formulated as, "He who has the gold makes the rules," underscoring not just its departure from, but also its contempt for the Golden Rule. This is the secret maxim of many, if they can get away with it, and often the unspoken precept of the powerful.

Finally, I should mention two other rules, found throughout the living world. They explain a great deal: **Suck up to those above you, and abuse those below.** This is the motto of bullies and the norm in many nonhuman primate societies. It's really the Golden Rule for superiors, the Iron Rule for inferiors. Since there is no known alloy of gold and iron, we'll call it the Tin Rule for its flexibility. The other common rule is: **Give precedence in all things to close relatives, and do as you like to others.** This Nepotism Rule is known to evolutionary biologists as "kin selection."

Despite its apparent practicality, there's a fatal flaw in the Brazen Rule: unending vendetta. It hardly matters who starts the violence. Violence begets violence, and each side has reason to hate the other. "There is no way to peace," A. J. Muste said, "Peace is the way." But peace is hard and violence is easy. Even if almost everyone is for ending the vendetta, a single act of retribution can stir it up again: A dead relative's sobbing widow and grieving children are before us. Old men and women recall atrocities from their childhoods. The reasonable part of us tries to keep the peace, but the passionate part of us cries out for vengeance. Extremists in the two warring factions can count on one another. They are allied against the rest of us, contemptuous of appeals to understanding and loving-kindness. A few hotheads can force-march a legion of more prudent and rational people to brutality and war.

Many in the West have been so mesmerized by the appalling accords with Adolf Hitler in Munich in 1938 that they are unable to distinguish cooperation and appeasement. Rather than having to judge each gesture and approach on its own merits, we merely decide that the opponent is thoroughly evil, that all his concessions are offered in bad faith, and that force is the only thing he understands. Perhaps for Hitler this was the right judgement. But in general it is not the right judgment, as much as I wish that the invasion of the Rhineland had been forcibly opposed. It consolidates hostility on both sides and makes conflict much more likely. In a world with nuclear weapons, uncompromising hostility carries special and very dire dangers.

Breaking out of a long series of reprisals is, I claim, very hard. There are ethnic groups who have weakened themselves to the point of extinction because they had no machinery to escape from this cycle, the Kaingang of the Brazilian highlands, for example. The warring nationalities in the former Yugoslavia, in Rwanda, and elsewhere may provide further examples. The Brazen Rule seems too unforgiving. The Iron Rule promotes the advantage of a ruthless and powerful few against the interests of everybody else. The Golden and Silver Rules seem too complacent. They systematically fail to punish cruelty

and exploitation. They hope to coax people from evil to good by showing that kindness is possible. But there are sociopaths who do not much care about the feelings of others, and it is hard to imagine a Hitler or Stalin being shamed into redemption by good example. Is there a rule between the Golden and Silver on the one hand and the Brazen, Iron and Tin on the other which works better than any of them alone?

With so many different rules, how can you tell which to use, which will work? More than one rule may be operating even in the same person or nation. Are we doomed just to guess about this, or to rely on intuition, or just to parrot what we've been taught? Let's try to put aside, just for the moment, whatever rules we've been taught, and those we feel passionately—perhaps from a deeply rooted sense of justice—*must* be right.

Suppose we seek not to confirm or deny what we've been taught but to find out what really works. Is there a way to test competing codes of ethics? Granting that the real world may be much more complicated than any simulation, can we explore the matter scientifically?

\*\*\*

We're used to playing games in which somebody wins and somebody loses. Every point made by our opponent puts us that much farther behind. "Win-lose" games seem natural, and many people are hard pressed to think of a game that isn't win-lose. In win-lose games, the losses just balance the wins. That's why they're also called "zero-sum" games. There's no ambiguity about your opponent's intentions: Within the rules of the game, he will do anything he can to defeat you.

Many children are aghast the first time they really come face to face with the "lose" side of win-lose games. On the verge of bankruptcy in Monopoly, for example, they plead for special dispensation (forgoing rents, for example), and when this is not forthcoming may, in tears, denounce the game as heartless and unfeeling—which, of course, it is. (I've seen the board overturned, hotels and "Chance" cards and metal icons spilled onto the floor in spitting anger and humiliation—and not only by children.) Within the rules of Monopoly, there's no way for players to cooperate so that all benefit. That's not how the game is designed. The same is true for boxing, football, hockey, basketball, baseball, lacrosse, tennis, racquetball, chess, all Olympic events, yacht and car racing, pinochle, potsie, and partisan politics. In none of these games is there an opportunity to practice the Golden or Silver Rule, or even the Brazen. There is room only for the Rule of Iron and Tin. If we reverse the Golden Rule, why is it so rare in the games we teach our children?

After a million years of intermittently warring tribes we readily enough think in zero-sum mode, and treat every interaction as a contest or conflict. Nuclear war, though (and many conventional wars), economic depression and assaults on the global environment are all "lose-lose" propositions. Such vital human concerns as love, friendship, parenthood, music, art, and the pursuit of knowledge are "win-win" propositions. Our vision is dangerously narrow if all we know is "win-lose."

The scientific field that deals with such matters is called game theory, used in military tactics and strategy, trade policy, corporate competition, limiting of environmental pollution, and plans for nuclear war. The paradigmatic game is the Prisoner's Dilemma. It

is very much non-zero-sum. Win-win, win-lose and lose-lose outcomes all are possible. "Sacred" books carry few useful insights into strategy here. It is a wholly pragmatic game.

Imagine that you and a friend are arrested for committing a serious crime. For the purpose of the game, it doesn't matter whether either, neither, or both of you did it. What matters is that the police say they think you did. Before the two of you have any chance to compare stories or plan strategy, you are taken to separate interrogation cells. There, oblivious of your Miranda rights ("You have the right to remain silent..."), they try to make you confess. They tell you, as police sometimes do, that your friend has confessed and implicated you. (Some friend!) The police might be telling the truth. Or they might be lying. You're permitted only to plead innocent or guilty. If you're willing to say anything, what's your best tack to minimize punishment?

Here are the possible outcomes:

If you deny committing the crime and (unknown to you) your friend also denies it, the case might be hard to prove. In the plea bargain, both your sentences will be very light.

If you confess, and your friend does likewise, then the effort the State had to expend to solve the crime was small. In exchange you both may be given a fairly light sentence, although not so light as if you both had asserted your innocence.

But if you plead innocent, and your friend confesses, the State will ask for the maximum sentence for you and minimal punishment (maybe none) for your friend. Uh-oh. You are very vulnerable to a kind of double cross, what game theorists call "defection." So's he.

So, if you and your friend "cooperate" with one another—both pleading innocent (or both pleading guilty)—you both escape the worst. Should you play it safe and guarantee no worse than a middle range of punishment by confessing? Then, if your friend pleads innocent while you plead guilty, well, too bad for him, and you might get off scot-free.

When you think it through, you realize that, whatever your friend does you're better off defecting than cooperating. Maddeningly, the same holds true for your friend. But if you both defect, you are both worse off than if you had both cooperated. This is the Prisoner's Dilemma.

Now consider a repeated Prisoner's Dilemma, in which the two players go through a sequence of such games. At the end of each they figure out from their punishment how the other must have pled. They gain experience about each other's strategy (and character). Will they learn to cooperate game after game, both always denying that they committed any crime? Even if the reward for finking on the other is large?

You might try cooperating or defecting, depending on how the previous game or games have gone. If you cooperate overmuch, the other player may exploit your good nature. If you defect overmuch, your friend is likely to defect often, and this is bad for both of you. You know your defection pattern is data being fed to the other player. What is the right mix of cooperation and defection? How to behave then becomes, like any other question in Nature, a subject to be investigated experimentally.

This matter has been explored in a continuing round-robin computer tournament by the University of Michigan sociologist Robert Axelrod, in his remarkable book *The Evolution of Cooperation*. Various codes of behavior confront one another, and at the end we see who wins (who gets the lightest cumulative prison term). The simplest strategies might be to cooperate all the time, no matter how much advantage is taken of you; or never to cooperate, no matter what benefits might accrue from cooperation. These are the Golden Rule and the Iron Rule. They always lose, the one from a superfluity of kindness, the other from an overabundance of ruthlessness. Strategies slow to punish defection lose—in part because they send a signal that noncooperation can win. The Golden Rule is not only an unsuccessful strategy; it is also dangerous for other players, who may succeed in the short term only to be mowed down by exploiters in the long term.

Should you defect at first, but if your opponent cooperates even once, cooperate in all future games? Should you cooperate at first, but if your opponent defects even once, defect in all future games? These strategies also lose. Unlike sports, you cannot rely on your opponent to be always out to get you.

The most effective strategy in many such tournaments is called "Tit-for-Tat." It's very simple: You start out cooperating, and in each subsequent round simply do what your opponent did the last time. You punish defections, but once the other player cooperates, you're willing to let bygones be bygones. At first it seems to garner only mediocre success. But as time goes on, the other strategies defeat themselves, from too much kindness or too much cruelty—and this middle way pulls ahead. Except for always being nice on the first move, Tit-for-Tat is identical to the Brazen Rule. It promptly (in the very next game) rewards cooperation and punishes defection, and has the great virtue that it makes your strategy absolutely clear to your opponent. (Strategic ambiguity can be lethal.)

### **TABLE OF PROPOSED RULES TO LIVE BY**

<b>The Golden Rule</b>	Do unto others as you would have them do unto you.
<b>The Silver Rule</b>	Do not do unto others what you would not have them do unto you.
<b>The Brazen Rule</b>	Do unto others as they do unto you.
<b>The Iron Rule</b>	Do unto others as you like, before they do it unto you.
<b>The Tit-for-Tat Rule</b>	Cooperate with others first, then do unto them as they do unto you.

Once there get to be several players employing Tit-for-Tat, they rise in the standings together. To succeed, Tit-for-Tat strategists must find others who are willing to reciprocate, with whom they can cooperate. After the first tournament in which the Brazen Rule unexpectedly won, some experts thought the strategy too forgiving. Next tournament, they tried to exploit it by defecting more often. They always lost. Even experienced strategists tended to underestimate the power of forgiveness and reconciliation. Tit-for-Tat involves an interesting mix of proclivities: initial friendliness,

willingness to forgive, and fearless retaliation. The superiority of the Tit-for-Tat Rule in such tournaments was recounted by Axelrod.

Something like it can be found throughout the animal kingdom and has been well-studied in our closest relatives, the chimps. Described and named "reciprocal altruism" by the biologist Robert Trivers, animals may do favors for others in expectation of having the favors returned—not every time, but often enough to be useful. This is hardly an invariable moral strategy, but is not uncommon either. So there is no need to debate the antiquity of the Golden, Silver, and Brazen Rules, or Tit-for-Tat, and the priority of the moral prescriptives in the Book of Leviticus. Ethical rules of this sort were not originally invented by some enlightened human lawgiver. They go deep into our evolutionary past. They were with our ancestral line from a time before we were human.

The Prisoner's Dilemma is a very simple game. Real life is considerably more complex. If he gives our apple to the pencil man, is my father more likely to get an apple back? Not from the pencil man; we'll never see him again. But might widespread acts of charity improve the economy and give my father a raise? Or do we give the apple for emotional, not economic rewards? Also, unlike the players in an ideal Prisoner's Dilemma game, human beings and nations come to their interactions with predispositions, both hereditary and cultural.

But the central lessons in a not very prolonged round-robin of Prisoner's Dilemma are about strategic clarity; about the self-defeating nature of envy; about the importance of long-term over short-term goals; about the dangers of both tyranny and patsydom; and especially about approaching the whole issue of rules to live by as an experimental question. Game theory also suggests that a broad knowledge of history is a key survival tool.