Edge

Summer Reading from the Archive

INDIRECT RECIPROCITY, ASSESSMENT HARDWIRING, AND REPUTATION

A Conversation with Karl Sigmund

[December 2004]

These ideas fed into our work on indirect reciprocity, a concept that was first introduced by Robert Trivers in a famous paper in the 1970s. I recall that he mentioned this idea obliquely when he wrote about something he called "general altruism." Here, you give something back not to the person to whom you owe something, but to somebody else in society. He pointed out that this also works with regard to cooperation at a high level. Trivers didn't go into details, because at the time it was not really at the center of his thinking. He was mostly interested in animal behavior, and so far indirect reciprocity has not been proven to exist in animal behavior. It might exist in some cases, but ethologists are still debating the pros and cons.

In human societies, however, indirect reciprocity has a very striking effect. There is a famous anecdote about the American baseball player Yogi Berra, who said something to the effect of, "I make a point of going to other people's funerals because otherwise they won't come to mine." This is not as nonsensical as it seems. If a colleague of the university, for instance, goes faithfully to every faculty member's funeral, then the faculty will turn out strongly at his. Others reciprocate. It works. We think instinctively in terms of direct reciprocation—when I do something for you, you do something for me—but the same principle can apply in situations of indirect reciprocity. I do something for you and somebody else helps me in return.

Introduction

In an essay published in *Nature* ("Prisoners of the dilemma: When mathematics and biology met on a mountain," February 2004), Austrian mathematical biologist Martin Nowak, director of the Program for Evolutionary Dynamics at Harvard, wrote about his relationship with his fellow Austrian, the mathematician Karl Sigmund:

> Once a year the theoretical chemist Peter Schuster used to take his students from the University of Vienna to a small house in the Austrian mountains. During the day we skied, of course, but in the evening the emphasis was on science. I was a first-year PhD student looking for a project. The mathematician Karl Sigmund was there and gave a talk on what was a new topic for him: the prisoner's dilemma. At the end of the talk I asked a question, and the next day we traveled back to Vienna, endlessly debating this game. In subsequent days, I visited Karl's office and we started to do calculations. We had become prisoners of the dilemma.
>
> We often met in coffee houses, the genius loci of past glory. Here Kurt Gödel announced his incompleteness theorem, Ludwig Boltzmann worked on entropy, and Ludwig Wittgenstein challenged the Vienna Circle. Or we walked in the Vienna Forest, visiting a meadow called "Himmel" (Heaven), where a sign noted that here Sigmund Freud first understood the nature of dreams.
>
> Within a year, we had conceived an evolutionary description of probabilistic strategies in the prisoner's dilemma struggling for cooperation by natural selection The prevailing paradigm, tit-for-tat, an unforgiving retaliator, was replaced by generous tit-for-tat (which always cooperates when the other person has cooperated and sometimes even when the other person has

defected) and later by win—stay, lose—shift (which stays with its current choice if the score is above an aspiration level and changes otherwise). A byproduct of this work was adaptive dynamics, representing a new way to look at evolution of strategies in a continuous space.

My introduction to Karl Sigmund in the Austrian mountains was the turning point that brought together mathematics and biology.

~ ~

"I am often thinking about the different ways of cooperating," says Karl Sigmund, professor of mathematics at the University of Vienna, "and nowadays I'm mostly thinking about the strange aspects of indirect reciprocity. Right now, it turns out that economists are excited about this idea in the context of e-trading and e-commerce. In this case you also have a lot of anonymous interactions, not between the same two people but within a hugely mixed group where you are unlikely ever to meet the same person again. Here the question of trusting the other, the idea of reputation, is particularly important. Google page rankings, the reputation of eBay buyers and sellers, and the Amazon.com reader reviews are all based on trust, and there is a lot of moral hazard inherent in these interactions."

I recently visited with Sigmund in Vienna and we discussed the implications of his ideas on Internet commerce. He became aware of the implications of these ideas for Internet commerce only a short time ago. He does not use eBay or buy books on Amazon. "It was my students who told me that reputation is now a very interesting issue in this context. When I tried for the nth time to write an introduction to a working paper on the topic of indirect reciprocity, they asked me, 'Why are you always looking to hominid evolution, to pre-history, when a similar thing is happening right now on the Internet?'

"Whether or not the pioneers behind Google are aware of these theories coming out of evolutionary biology would be an interesting question to ask," he says. "I have no contact with them, but claim a partial, subliminal credit for the term Google.

"When the Brin family first came to Vienna from Russia, Sergey Brin was three years old. The family stayed with us at our apartment for some time. His father, a friend and colleague, is the mathematician Michael Brin, who specializes in ergodic theory and dynamical systems, and that was also my field at that time. The first thing we offered the Brins when they entered our home was a 'Guglhupf,' the famous Austrian dessert...which greatly impressed the young Sergey. But I

bet he does not remember. Officially, Google comes from googol—a very large number—and the real reason for the name is lost in oblivion."

—JB

—

KARL SIGMUND is a professor of mathematics at the University of Vienna and one of the pioneers of evolutionary game theory. He is the author of *Exact Thinking in Demented Times: The Vienna Circle and the Epic Quest for the Foundations of Science.*

INDIRECT RECIPROCITY, ASSESSMENT HARDWIRING, AND REPUTATION

KARL SIGMUND: Direct reciprocity is an idea that we can see every day in every household. If my wife cooks, for instance, I will always wash the dishes. Otherwise, cooperation breaks down. This is a commonplace situation, and similar exchanges have been measured and studied for many years. Martin Nowak and I began working together on this idea of direct reciprocity after meeting at an Austrian mountain retreat in the late 1980s. I was there to deliver my lecture on Robert Axelrod, whose book, *The Evolution of Cooperation*, was already a classic. It did wonders for the study of direct reciprocity, presenting a lot of his own work, and also leading to more research. At the time that Axelrod wrote his book there were already hundreds of papers by psychologists, philosophers, and mathematicians on what has been called the prisoner's dilemma. Afterward there were thousands. It launched a big field.

The simplest instance of the prisoner's dilemma would be if you had two persons in separate rooms, each of whom has to decide whether or not to give a gift to the other player. According to the rules of the game, each gift-giver may give the other player $3, but doing so will cost him $1. Both players must make this decision at the same time. If both make the decision to give to the other player, both receive $3 while they have to pay $1, so both end up with a $2 net benefit. If both decline to give, it costs them nothing, and the payoff in both cases is zero. The rub is that if one player gives to the other player, and if the other player does not at the same time decide to give in return, then one player is exploited. He has given $1 but gets nothing in return, whereas the other player, the nasty exploiter, gets $3 and is best off.

In the simplest prisoner's dilemma, the game is played only once. If there is no future interaction it's obvious what you should do, because one only has to consider two possibilities. The other player can give a gift, in which case it's very good to accept it and not to give anything in return. You get $3 and it costs you nothing. If the other player does not give a gift, then it is again to your advantage not to give the other player anything because you would be foolish to pay him something when he doesn't return the favor. Under each circumstance, no matter what the other player does, you should defect by not giving a gift.

However, if the other player thinks the same thing, then you both end up with nothing. If you had both been generous you would both have earned $2 by the game. This means that pursuing your selfish interest in a rational way by thinking out the alternatives and doing what is best for you will actually lead to a solution that is bad for you. If you follow your instincts, which, if you don't know the other players, probably tell you to try cooperating and to be generous, then you will fare well. There is a distinction between the benefit of the group, so to speak (each of you gets $2) and a selfish benefit (if both you and the other player try hard for the selfish benefit, none gets anything).

For the next ten years Martin and I milked the prisoner's dilemma with pleasure and success. At the time we were looking for other simple but interesting economic experiments and situations. Martin, for instance, introduced the spatial prisoner's dilemma. Here, you don't form bonds with just anyone in a population and play 200 rounds of the prisoner's dilemma with that person, but you interact only with your immediate neighbors. This is, of course, much more realistic, because usually you do not interact with just anybody in the town of Vienna, say, but only within a very small social network.

This study of the prisoner's dilemma went on for quite some time, but I must say I was happy when we reached an opening into what is now called indirect reciprocity about six years ago. Martin and I were the first to formalize the idea, creating a mathematical model to analyze this precisely. We wrote the first major paper that talked about how to analyze this question and then to set up experiments, and now there are dozens of groups that are actively involved.

There was one idea in particular in the literature already that was not taken very seriously. People thought the prisoner's dilemma game could be played according to a strategy called "tit-for-tat." This strategy says that whenever you meet a new person, in the first round you should cooperate, and from then on you should do whatever the other fellow has done in the previous round. But real life is actually subtler. When you meet a new partner, you possibly know a bit about his past. You have not interacted with him, but he has interacted with other people. If you

know that he defected against another person it would be good for you to start by expecting defection, and therefore to defect yourself. This has been called "observer tit-for-tat." It is just like tit-for-tat except that in the first round you do not necessarily cooperate; you cooperate only if you know that in interactions with other third parties this person has been nice.

Let me explain this very carefully. Tit-for-tat is a completely natural strategy. There is no inventor of this strategy, although it was submitted to prisoner's dilemma tournaments by Canadian game theorist Anatol Rappaport. He submitted the simplest of all strategies, which consisted of two lines of Fortran programming. The strategy is simply that in the first round, when you don't know the other player, you cooperate with the other player. You give him the benefit of the doubt, so to speak. From then onward, you mimic whatever the other player did in the previous round. If he was nasty, then you are nasty. If he was friendly and cooperative, you are friendly and cooperative. This strategy was extremely successful in computer tournaments.

In a computer tournament, however, you can practically exclude the possibilities of mistakes and errors. In real life it is quite likely that mistakes will occasionally happen. You can intend to do a nice thing and just happen not to have whatever it takes to do it. You can misunderstand an action of your partner. You can be in a bad mood because somebody else had hurt you during the day and you stupidly get your revenge on the wrong person. All these things can happen.

If you play tit-for-tat and are subject to these mistakes, then you are likely to get embroiled in a needless mutual punishment. If, for instance, the other player plays tit-for-tat and made a mistake by defecting in the previous round, you would punish him by defecting in this round. He will punish you, in turn, by defecting in the next round. And then you will punish him again, and so on, and so on, creating an endless vendetta. This continues until the next mistake puts a stop to it, although actually the next mistake can make things even worse. It could happen that from then on you keep punishing each other in every round mutually—not alternating, but simultaneously. At this point the situation is fairly hopeless.

Robert May suggested in an article in *Nature* that people should really use more generous strategies, allowing for occasional forgiveness. In our first paper in *Nature*, Martin and I presented a strategy called "generous tit-for-tat." Whenever the other player has done you a good turn, then there is a 100% chance that you will reply with a good action. But if the other player has done you a bad turn, then you will only return this bad action with certain probability, depending on the ratio of cost-to-benefit—say, only with a probability of 35%. Therefore, there is a rather

high probability that the cycle of mutual punishment which starts with an erroneous defection will be broken after one or two rounds. At that point, the players are again in a good mood and cooperate with each other until the next mistake happens. Generous tit-for-tat was a very obvious and very robust solution to this problem of what happens when mistakes occur.

Later we found another even more robust strategy than generous tit-for-tat. This was later called Pavlov's strategy, a name that is not the best possible, but that has stuck. Pavlov's strategy says that you should cooperate if and only if in the previous round you and your co-player have done the same thing. According to this strategy:

If you both cooperated, then you cooperate.

If you have both defected, then you should also cooperate.

If you have cooperated and the other player has defected, then you should defect in the next round.

If you defected and the other player has cooperated, then you should again defect in the next round.

At first glance the strategy looks bizarre, but in our computer simulation it turned out that it always won in an environment where mistakes were likely. In the end, it was almost always the dominating strategy in the population. Almost everyone was playing Pavlov's strategy, and it was very stable; it was much better than tit-for-tat.

Later we understood that this strategy is actually not so strange. It is the simplest learning mechanism that you can imagine. This is a win-stay, lose-shift learning mechanism that has already been studied in animals—for training horses and so on—for 100 years. It says simply that if a person gets a bad result, then he is less likely to repeat the former move. And if a person has a good outcome and wins, then he is more likely to repeat the former move because it was successful. It is simply reward and punishment in action. If one studies this in the context of the prisoners' dilemma, one gets exactly Pavlov's strategy. For instance, if you have defected and the other player has cooperated, then according to the prisoners' dilemma, you exploit the other player and your payoff is very high. You are very happy, and so you repeat your move, therefore defecting again in the next round. However, if you have cooperated and the other player has defected, then you

have been exploited. You are very unhappy, and you are going to switch to another move. You have cooperated in the past, but now you are going to defect.

These are theoretical experiments, but students of animal behavior also did lots of real-life experiments. They even turned to humans with this series of experiments. People like Manfred Milinski, who is director of a Max Planck Institute in Germany and a very straight, very hard-nosed student of animal behavior, started a new career in studying human nature. He used students from Switzerland and Germany to set up experiments based on the prisoner's dilemma game to check whether people were likely to play this Pavlov strategy or not. He found that, indeed, there is strong evidence that Pavlov's strategy is quite widespread in humans.

These ideas fed into our work on indirect reciprocity, a concept that was first introduced by Robert Trivers in a famous paper in the 1970s. I recall that he mentioned this idea obliquely when he wrote about something he called "general altruism." Here, you give something back not to the person to whom you owe something, but to somebody else in society. He pointed out that this also works with regard to cooperation at a high level. Trivers didn't go into details, because at the time it was not really at the center of his thinking. He was mostly interested in animal behavior, and so far indirect reciprocity has not been proven to exist in animal behavior. It might exist in some cases, but ethologists are still debating the pros and cons.

In human societies, however, indirect reciprocity has a very striking effect. There is a famous anecdote about the American baseball player Yogi Berra, who said something to the effect of, "I make a point of going to other people's funerals because otherwise they won't come to mine." This is not as nonsensical as it seems. If a colleague of the university, for instance, goes faithfully to every faculty member's funeral, then the faculty will turn out strongly at his. Others reciprocate. It works. We think instinctively in terms of direct reciprocation—when I do something for you, you do something for me—but the same principle can apply in situations of indirect reciprocity. I do something for you and somebody else helps me in return.

Balzac wrote that behind every large fortune there is a crime. This is a very romantic, absurd, and completely outdated idea. In fact, it very often happens that behind a great fortune or a great success is some action that is particularly generous. In one of my research projects, I'm collecting such cases. For example, the French branch of the Rothschild family protected the money of their English clients during the Napoleonic Wars. They were under extreme political pressure to give it up, but they kept the interests of their English clients at heart. Afterwards,

of course, they became extraordinarily rich because everybody knew that they could be trusted.

From Trivers' work others derived models about indirect reciprocity, but they were the wrong types of models. People had been reading Axelrod and there were some abortive attempts at modeling indirect reciprocity and explaining it through game theory. Their conclusion was that reciprocity could not work except in groups of two, which have to interact for a long time. One idea was that the principle behind indirect reciprocity is that if I receive something, I'm more likely to give the next person who comes along. There might be something true about it, but there have been experiments showing that this principle by itself would not suffice with regard to explaining the stability of indirect reciprocity.

Then a famous scientist in Chicago named Richard Alexander, the director of the Museum of Natural History, wrote a book about the Darwinian evolution of morals. In this book he asked questions like, what is moral? And how do we start to form our ideas about what is good and bad? We look at what people do for society. We are always assessing the reputations of others, and are more likely to give to somebody who has a high reputation, someone who has in her or his past, given help t others—not necessarily to me, though, but to somebody. If I only give to a person with a high reputation, I channel my help to those who have proved their value for cooperation.

Martin picked up on this work and started with a very simple model. It was a kind of numerical score, a label that says how often a person has given in the past. The idea is that my decision of whether to give to that person or not will depend on that score. I give more freely to persons with a high score. I refuse to help persons with a low score. This extremely simple model worked very well and inspired a lot of economists to do many experiments based on it.

In this type of experiment a group of ten people don't know each other and remain anonymous. All they know is that each person in this group has a number, 1 though 10. Occasionally, two people are chosen randomly. One of them is assigned the role of the donor, and the other is assigned the role of the recipient. The donor can give $3 to the recipient at the cost of $1 dollar to himself. If one assumes this person is selfish and rational, she should not give anything and should keep this $1. She is not going to be punished anyway. In many of these experiments, under conditions of complete anonymity, at the beginning people tend to give once or twice. But when they see there is no immediate return, they stop giving.

However, it's different if the donor knows that next to the number of each participant is a sign indicating how often this participant has given in the past. If a recipient has given five times and refused to give only once, for example, then the recipient has a respectably high score, and it turned out that the donors have a tendency to give preferentially to those with a high score. Manfred Milinski and others have studied this example, and it has become almost a trade by now. There are dozens of papers on this simple setup that show that indirect reciprocity works under very simple conditions.

At the same time, though, there have been theoreticians who have said that this model cannot work for a very simple reason: If you are discriminating and see that a recipient is a defector who has not given, then you will not give to that person. But at the same time that you punish him by not giving your own score will be diminished. Even if this act of not giving is fully justified in your eyes, the next person who sees only whether you have given or not in the past will infer, "Ah ha! You have not given, and therefore you are a bad guy, and you will not get anything from me." By punishing somebody you lower your own score and therefore you risk not receiving a benefit from a third person in the next round. Punishing somebody is a costly business, because it costs you future benefits.

The theoreticians then said this cannot work. Why should you engage in this act of punishment when it costs you something? This has been called a social dilemma. Punishing others is altruistic in the sense that if you didn't have this possibility of punishment, cooperation would vanish from the group. But this altruism costs you something. Your own reasoning should tell you that it's better to always give, because then your score will always be at a maximum. Therefore, your chances of receiving something will also be maximal.

I'm often thinking about the different ways of cooperating, and nowadays I'm mostly thinking about these strange aspects of indirect reciprocity. Right now it turns out that economists are excited about this idea in the context of e-trading and e-commerce. In this case you also have a lot of anonymous interactions, not between the same two people but within a hugely mixed group where you are unlikely ever to meet the same person again. Here the question of trusting the other, the idea of reputation, is particularly important. Google page rankings, the reputation of eBay buyers and sellers, and the Amazon.com reader reviews are all based on trust, and there is a lot of moral hazard inherent in these interactions.

Before I get into specific examples, let's first talk about what we have been accustomed to in the past. In a marketplace in ancient Egypt or in a medieval town, you saw the same person day after day. Either he sold you something or you sold him something, and you grew old together. This is no longer the case. In

Internet commerce you buy something from somebody you have never seen, and you put your trust in an agency when you might only see its e-mail address. Lots of money is involved in these interactions, and there are lots of possibilities for cheating. You have to be sure that you trust the right e-commerce company, or that an offer on eBay is serious and will not give you worthless junk. You have to trust someone to send money before you get the goods, and when they arrive there's the question of determining if the goods are worth it.

Of course, in this kind of e-commerce there is also the possibility of rating your partner and reporting whether you were satisfied with him the last time or not. This builds up the reputation of every agent in this game, creating a modern version of reputation in a society where you are not meeting the same person day after day, but meeting a stranger.

I became aware of the implications of these ideas for Internet commerce only a short time ago, because I myself don't use eBay, or buy books on Amazon. It was my students who told me that reputation is now a very interesting issue in this context. When I tried for the nth time to write an introduction to a working paper on the topic of indirect reciprocity, they asked me, "Why are you always looking to hominid evolution, to pre-history, when a similar thing is happening right now on the Internet?" There are now at least ten papers by economists in the works on these topics that I could mention here, but all of them are in the preparatory stage and none has been published.

Whether or not the pioneers behind Google are aware of these theories coming out of evolutionary biology would be an interesting question to ask. I have no contact with them. But I claim a subliminal credit for the term Google.

When the Brin family first came to Vienna from Russia, Sergey Brin was three years old. The family stayed with us at our appartement for some time. His father, a friend and colleague, is the mathematician Michael Brin, who specializes in ergodic theory and dynamical systems, and that was also my field at that time. The first thing we offered the Brins when they entered our home was a "Guglhupf," the famous Austrian dessert...which greatly impressed the young Sergey. But I bet he does not remember. Officially, the name Google comes from googol—a very large number—and the real reason is lost in oblivion.

Of course, in the early 1980s I was not thinking about the Internet. I was reading Richard Alexander's work about the evolution of morals and how people first started assessing each other. Alexander knew of course that different cultures have very different morals. Similarly, different cultures have very different languages and nevertheless many linguists now generally assume that there is a

universal language instinct. Similarly, it is probable that although different cultures have different morals, they are based on a universal moral instinct in the sense that there is a tendency to assess people all the time, even if this occurs according to different models. Maybe these depend on the culture or the civilization, but basically our tendency for assessing others is hardwired.

Assessment hardwiring is something that could also be implemented in the context of e-commerce, because if you see two people and one of them refuses to give a gift to the other person, what do you infer? If you only see one isolated act, you will probably think that a potential donor who did not give is a bad guy. But if the potential recipient has a bad reputation, then you would probably say it is justified not to give to that person.

Something being studied experimentally right now is whether people really go into the fine details when they observe an interaction. Do they observe it in isolation, and do they really care about the standing of the persons involved? Of course, it would be more sophisticated for an assessment to consider both the moral standing of the donor and the moral standing of the recipient before this interaction took place.

In e-commerce you would want to quantify this moral standing with only minimal parameters. This would ideally rely on a binary label that is either 0 or 1, because this is easiest to implement. But if one starts thinking about different ways to assess an action between two people, there are many possible combinations of what you consider good and bad. It could be that you consider it good to refuse to help a bad guy, or that you consider it bad to help a bad guy. If you analyze all these possibilities you come to an amazingly high number of possible morals. When we calculated them, we found that there are actually 4,096 possibilities! Which one of them is really working at a given time? This is a question for experimentation, and experiments in this area are in full swing. I know several groups who are doing this. While I am not an experimentalist, I am becoming increasingly interested in that area.

It would be interesting to ask whether or not the pioneers behind Google, eBay, and Amazon are aware of these theories coming out of evolutionary biology. Reputation is something that is profoundly embedded in our mentality, and we—not just old professors, but everyone—care enormously about it. I have read that the moment when people really get desperate and start running amok is when they feel that they are considered completely worthless in their society.

I should stress that we have been talking here essentially about human nature. The more or less official idea that human beings are selfish and rational—an idea

that nobody except economists really took seriously, and now even economists say that they never did—this idea has now been totally discredited. There are many experiments that show that spontaneous impulses like the tendency for fairness or acts of sympathy or generosity play a huge role in human life.

___

Edge.org is a nonprofit private operating foundation under Section 501(c)(3) of the Internal Revenue Code.
Copyright © 2021 By Edge Foundation, Inc All Rights Reserved.

##