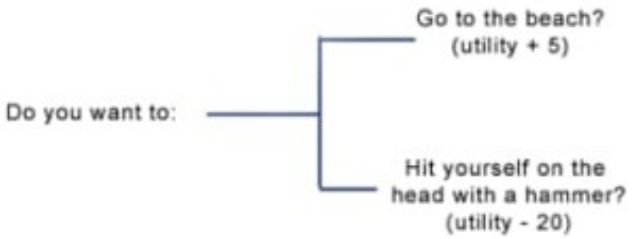

Do you want to:
Go to the beach?
(utility + 5)
Hit yourself on the
head with a hammer?
(utility - 20)

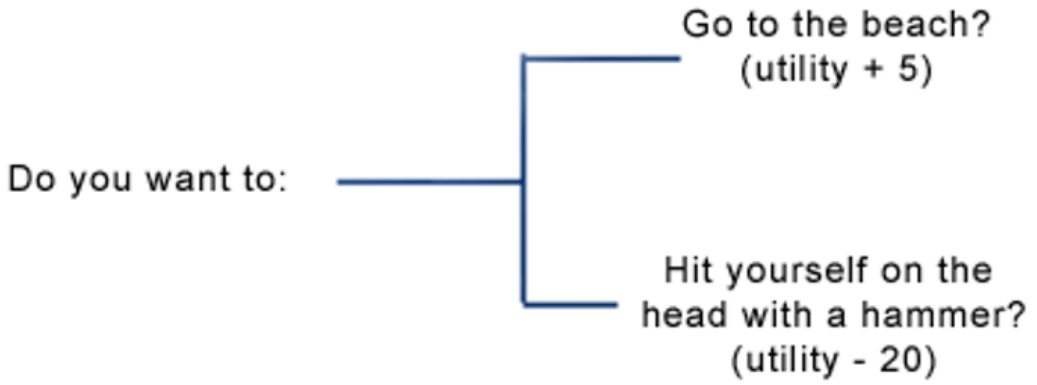
Introduction to Game Theory

Backward Reasoning Over Decision Trees

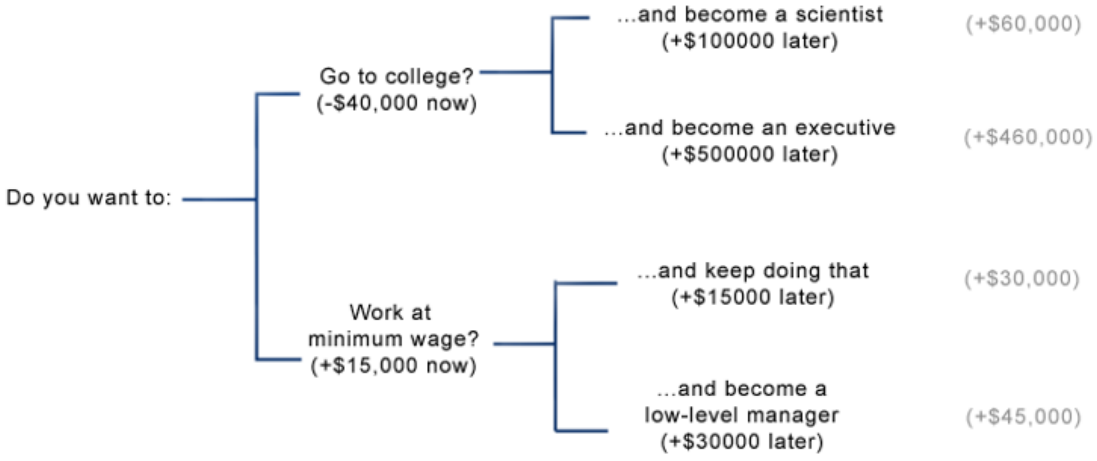
Game theory is the study of how rational actors interact to pursue incentives. It starts with the same questionable premises as economics: that everyone behaves rationally, that everyone is purely self-interested1, and that desires can be exactly quantified - and uses them to investigate situations of conflict and cooperation.

Here we will begin with some fairly obvious points about decision trees, but by the end we will have the tools necessary to explain a somewhat surprising finding: that giving a US president the additional power of line-item veto may in many cases make the president less able to enact her policies. Starting at the beginning:

The basic unit of game theory is the choice. Rational agents make choices in order to maximize their utility, which is sort of like a measure of how happy they are. In a one-person game, your choices affect yourself and maybe the natural environment, but nobody else. These are pretty simple to deal with:



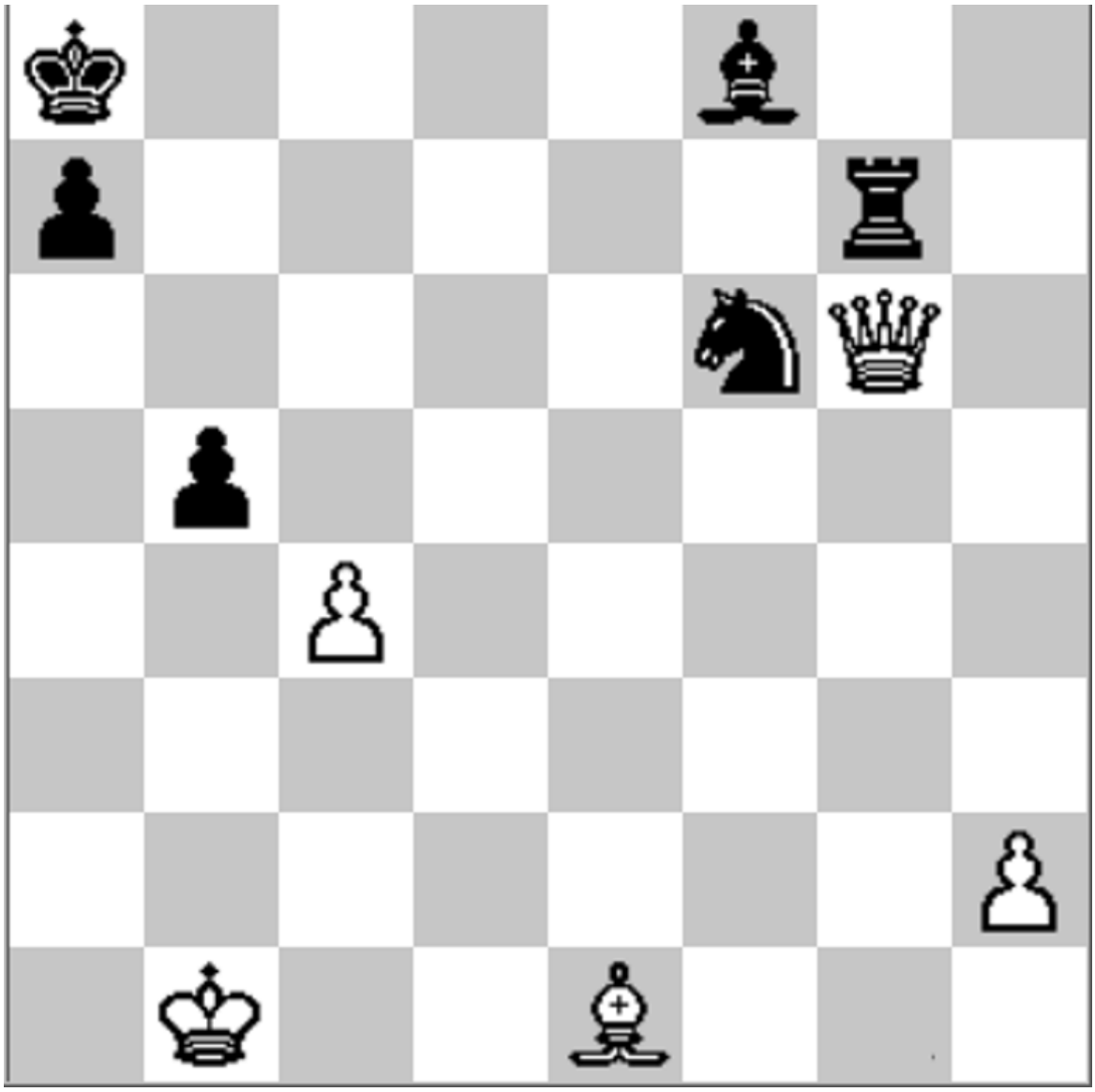
Here we visualize a choice as a branching tree. At each branch, we choose the option with higher utility; in this case, going to the beach. Since each outcome leads to new choices, sometimes the decision trees can be longer than this:



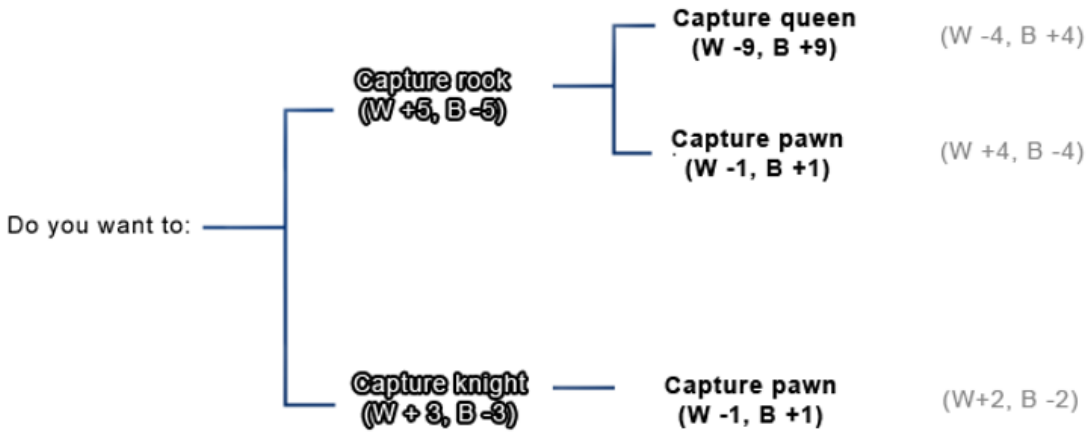
Here's a slightly more difficult decision, denominated in money instead of utility. If you want to make as much money as possible, then your first choice - going to college or starting a

minimum wage job right Now - seems to favor the more lucrative minimum wage job. But when you take Later into account, college opens up more lucrative future choices, as measured in the gray totals on the right-hand side. This illustrates the important principle of reasoning backward over decision trees. If you reason forward, taking the best option on the first choice and so on, you end up as a low-level manager. To get the real cash, you've got to start at the end - the total on the right - and then examine what choice at each branch will take you there.

This is all about as obvious as, well, not hitting yourself on the head with a hammer, so let's move on to where it really gets interesting: two-player games.

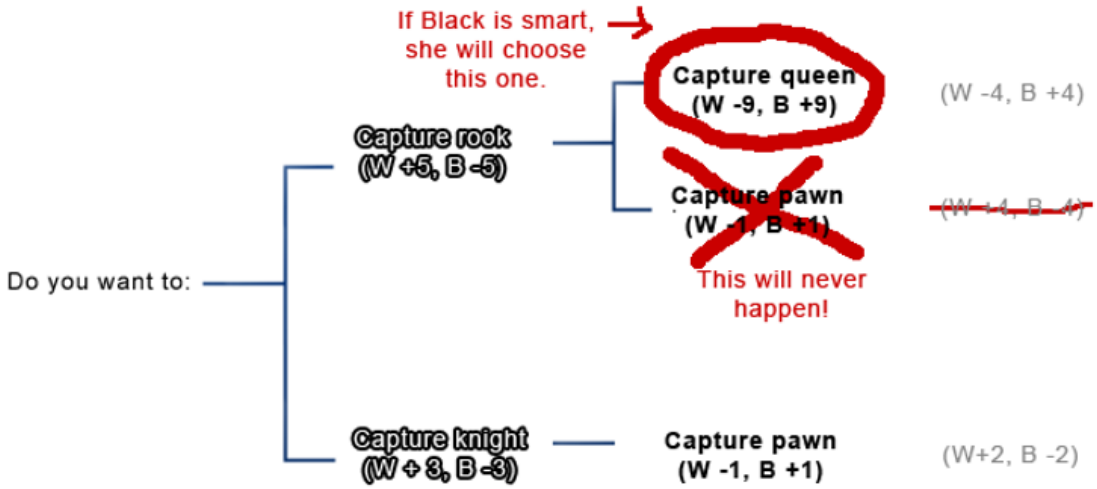


I'm playing White, and it's my move. For simplicity I consider only two options: queen takes knight and queen takes rook. The one chess book I've read values pieces in number of pawns: a knight is worth three pawns, a rook five, a queen nine. So at first glance, it looks like my best move is to take Black's rook. As for Black, I have arbitrarily singled out pawn takes pawn as her preferred move in the current position, but if I play queen takes rook, a new option opens up for her: bishop takes queen. Let's look at the decision tree:



If I foolishly play this two player game the same way I played the one-player go-to-college game, I note that the middle branch has the highest utility for White, so I take the choice that leads there: capture the rook. And then Black plays bishop takes queen, and I am left wailing and gnashing my teeth. What did I do wrong?

I should start by assuming Black will, whenever presented with a choice, take the option with the highest Black utility. Unless Black is stupid, I can cross out any branch that requires Black to play against her own interests. So now the tree looks like this:



The two realistic options are me playing queen takes rook and ending up without a queen and -4 utility, or me playing queen takes knight and ending up with a modest gain of 2 utility.

(my apologies if I've missed some obvious strategic possibility on this particular chessboard; I'm not so good at chess but hopefully the point of the example is clear.)

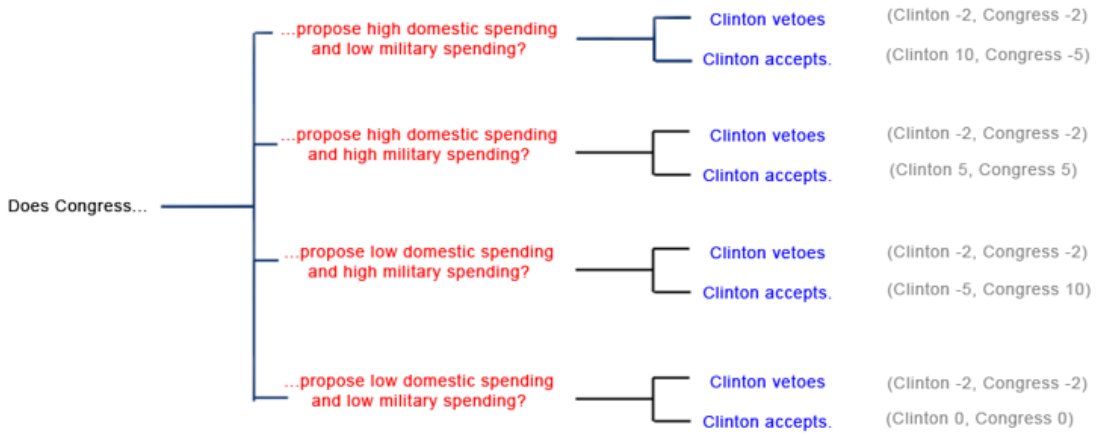
This method of alternating moves in a branching tree matches both our intuitive thought processes during a chess game ("Okay, if I do this, then Black's going to do this, and then I'd do this, and then...") and the foundation of some of the algorithms chess computers like Deep Blue use. In fact, it may seem pretty obvious, or even unnecessary. But it can be used to analyze some more complicated games with counterintuitive results.

Art of Strategy describes a debate from 1990s US politics revolving around so-called "line-item veto" power, the ability to veto only one part of a bill. For example, if Congress passed a bill declaring March to be National Game Theory Month and April to be National Branching

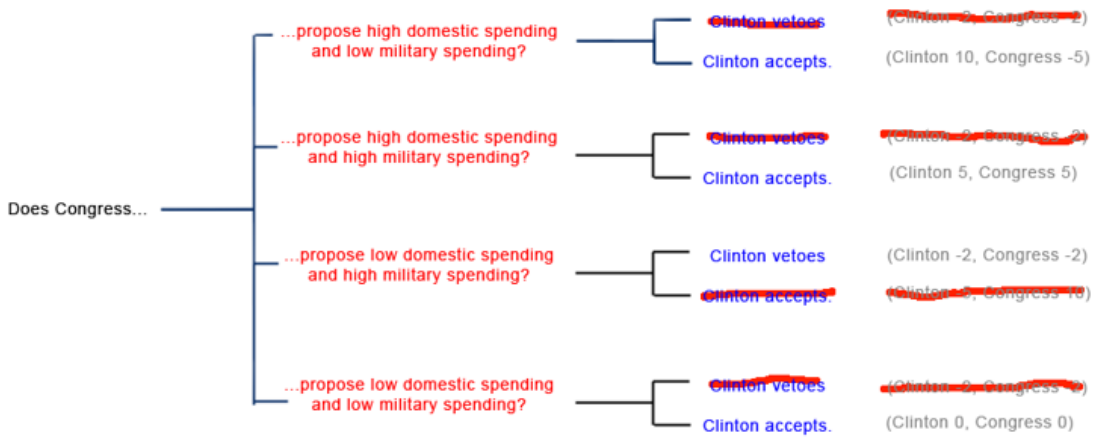
Tree Awareness Month, the President could veto only the part about April and leave March intact (as politics currently works, the President could only veto or accept the whole bill). During the '90s, President Clinton fought pretty hard for this power, which seems reasonable as it expands his options when dealing with the hostile Republican Congress.

But Dixit and Nalebuff explain that gaining line-item veto powers might hurt a President. How? Branching trees can explain.

Imagine Clinton and the Republican Congress are fighting over a budget. We can think of this as a game of sequential moves, much like chess. On its turn, Congress proposes a budget. On Clinton's turn, he either accepts or rejects the budget. A player “wins” if the budget contains their pet projects. In this game, we start with low domestic and military budgets. Clinton really wants to raise domestic spending (utility +10), and has a minor distaste for raised military spending (utility -5). Congress really wants to raise military spending (utility +10), but has a minor distaste for raised domestic spending (utility -5). The status quo is zero utility for both parties; if neither party can come to an agreement, voters get angry at them and they both lose 2 utility. Here's the tree when Clinton lacks line-item veto:

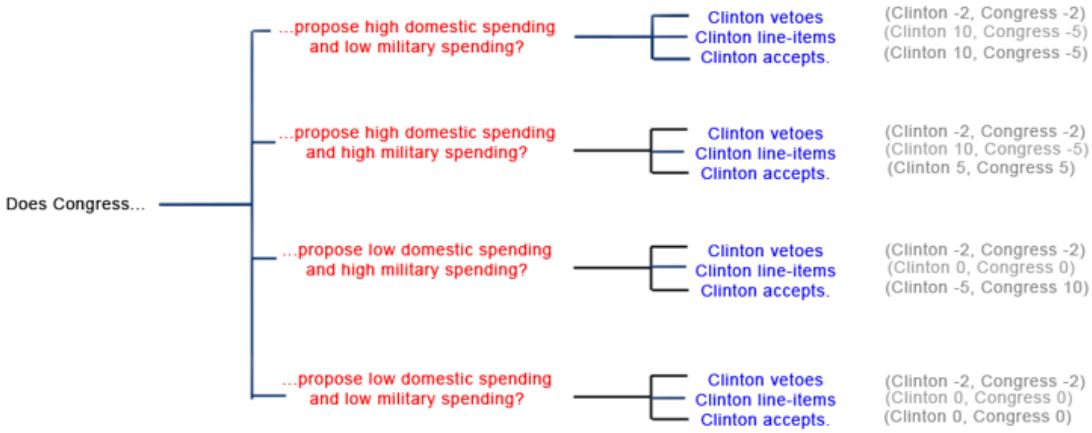


For any particular Republican choice, Clinton will never respond in a way that does not maximize his utility, so the the Republicans reason backward and arrive at something like this:

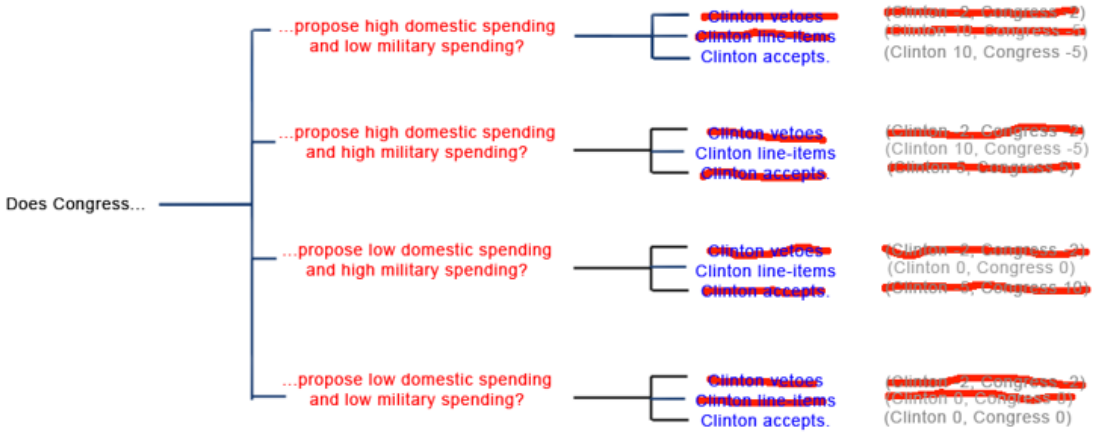


If Republicans are perfectly rational agents, they choose the second option, high domestic and high military spending, to give them their highest plausibly obtainable utility of 5.

But what if Clinton has the line-item veto? Now his options look like this:



If the Republicans stick to their previous choice of “high domestic and high military spending”, Clinton line-item vetoes the military spending, and we end up with a situation identical to the first choice: Clinton sitting on a pile of utility, and the Republicans wailing and gnashing their teeth. The Republicans need to come up with a new strategy, and their thought processes, based on Clinton as a utility-maximizer, look like this:



Here Congress's highest utility choice is to propose low domestic spending (it doesn't matter if they give more money to the military or not as this will get line-item vetoed). Let's say they propose low domestic and low military spending, and Clinton accepts. The utilities are (0, 0), and now there is much wailing and gnashing of teeth on both sides (game theorists call this a gnash equilibrium. Maybe you've heard of it.)

But now Clinton has a utility of 0, instead of a utility of 5. Giving him extra options has cost him utility! Why should this happen, and shouldn't he be able to avoid it?

This happened because Clinton's new abilities affect not only his own choices, but those of his opponents (compare Schelling: Strategies of Conflict). He may be able to deal with this if he can make the Republicans trust him.

In summary, simple sequential games can often be explored by reasoning backwards over decision trees representing the choices of the players involved. The next post will discuss simultaneous games and the concept of a Nash equilibrium.

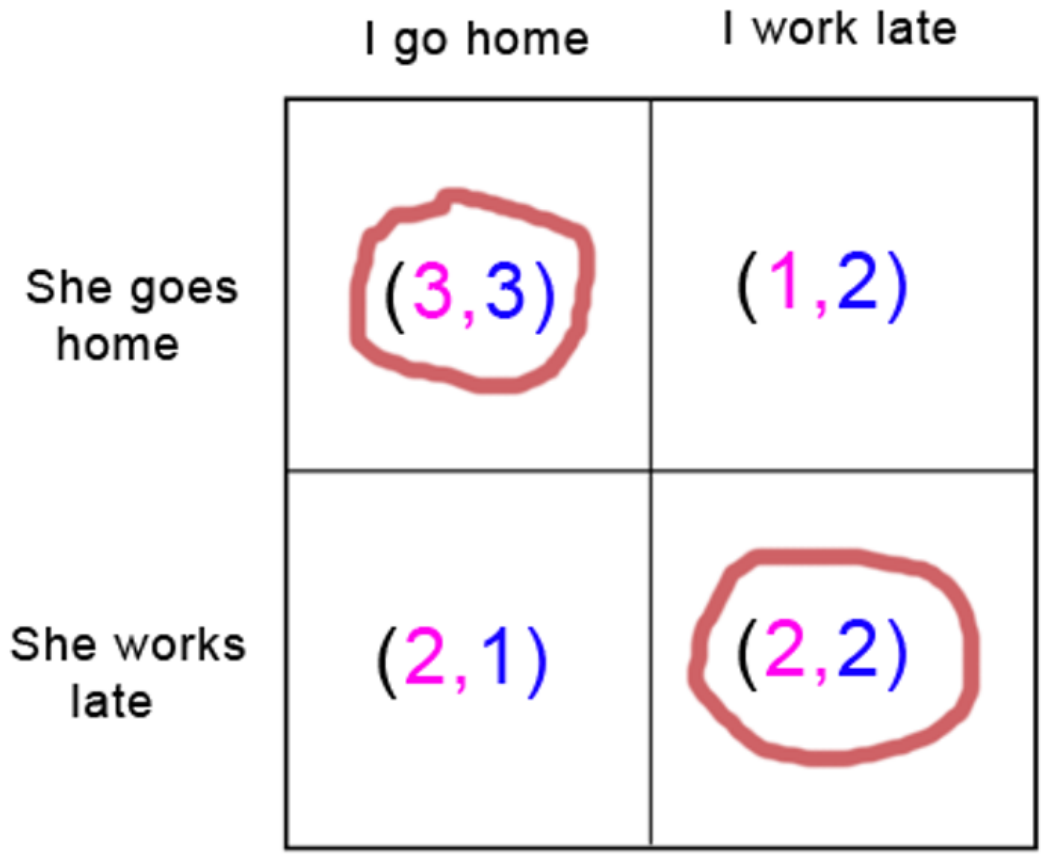
Footnotes:

1: Game theory requires self-interest in that all players' are driven solely by their desire to maximize their own payoff in the game currently being played without regard to the welfare of other players or any external standard of fairness. However, it can also be used to describe the behavior of altruistic agents so long as their altruistic concerns are represented in the evaluation of their payoff.

Nash Equilibria and Schelling Points

A Nash equilibrium is an outcome in which neither player is willing to unilaterally change her strategy, and they are often applied to games in which both players move simultaneously and where decision trees are less useful.

Suppose my girlfriend and I have both lost our cell phones and cannot contact each other. Both of us would really like to spend more time at home with each other (utility 3). But both of us also have a slight preference in favor of working late and earning some overtime (utility 2). If I go home and my girlfriend's there and I can spend time with her, great. If I stay at work and make some money, that would be pretty okay too. But if I go home and my girlfriend's not there and I have to sit around alone all night, that would be the worst possible outcome (utility 1). Meanwhile, my girlfriend has the same set of preferences: she wants to spend time with me, she'd be okay with working late, but she doesn't want to sit at home alone.



This “game” has two Nash equilibria. If we both go home, neither of us regrets it: we can spend time with each other and we've both got our highest utility. If we both stay at work, again, neither of us regrets it: since my girlfriend is at work, I am glad I stayed at work instead of going home, and since I am at work, my girlfriend is glad she stayed at work instead of going home. Although we both may wish that we had both gone home, neither of us specifically regrets our own choice, given our knowledge of how the other acted.

When all players in a game are reasonable, the (apparently) rational choice will be to go for a Nash equilibrium (why would you want to make a choice you'll regret when you know what the other player chose?) And since John Nash (remember that movie *A Beautiful Mind*?) proved that every game has at least one, all games between well-informed rationalists (who are not also being superrational in a sense to be discussed later) should end in one of these.

What if the game seems specifically designed to thwart Nash equilibria? Suppose you are a general invading an enemy country's heartland. You can attack one of two targets, East City or West City (you declared war on them because you were offended by their uncreative toponyms). The enemy general only has enough troops to defend one of the two cities. If you attack an undefended city, you can capture it easily, but if you attack the city with the enemy army, they will successfully fight you off.

	Attack East City	Attack West City
Defend East City	(0,1)	(1,0)
Defend West City	(1,0)	(0,1)

Here there is no Nash equilibrium without introducing randomness. If both you and your enemy choose to go to East City, you will regret your choice - you should have gone to West and taken it undefended. If you go to East and he goes to West, he will regret his choice - he should have gone East and stopped you in your tracks. Reverse the names, and the same is true of the branches where you go to West City. So every option has someone regretting their choice, and there is no simple Nash equilibrium. What do you do?

Here the answer should be obvious: it doesn't matter. Flip a coin. If you flip a coin, and your opponent flips a coin, neither of you will regret your choice. Here we see a "mixed Nash equilibrium", an equilibrium reached with the help of randomness.

We can formalize this further. Suppose you are attacking a different country with two new potential targets: Metropolis and Podunk. Metropolis is a rich and strategically important city (utility: 10); Podunk is an out of the way hamlet barely worth the trouble of capturing it (utility: 1).

	Attack Metropolis	Attack Podunk
Defend Metropolis	0	1
Defend Podunk	10	0

A so-called first-level player thinks: “Well, Metropolis is a better prize, so I might as well attack that one. That way, if I win I get 10 utility instead of 1”

A second-level player thinks: “Obviously Metropolis is a better prize, so my enemy expects me to attack that one. So if I attack Podunk, he'll never see it coming and I can take the city undefended.”

A third-level player thinks: “Obviously Metropolis is a better prize, so anyone clever would never do something as obvious as attack there. They'd attack Podunk instead. But my opponent knows that, so, seeking to stay one step ahead of me, he has defended Podunk. He will never expect me to attack Metropolis, because that would be too obvious. Therefore, the city will actually be undefended, so I should take Metropolis.”

And so on ad infinitum, until you become hopelessly confused and have no choice but to spend years developing a resistance to iocane powder.

But surprisingly, there is a single best solution to this problem, even if you are playing against an opponent who, like Professor Quirrell, plays “one level higher than you.”

When the two cities were equally valuable, we solved our problem by flipping a coin. That won't be the best choice this time. Suppose we flipped a coin and attacked Metropolis when we got heads, and Podunk when we got tails. Since my opponent can predict my strategy, he would defend Metropolis every time; I am equally likely to attack Podunk and Metropolis, but taking Metropolis would cost them much more utility. My total expected utility from flipping the coin is 0.5: half the time I successfully take Podunk and gain 1 utility, and half the time I am defeated at Metropolis and gain 0.And this is not a Nash equilibrium: if I had known my opponent's strategy was to defend Metropolis every time, I would have skipped the coin flip and gone straight for Podunk.

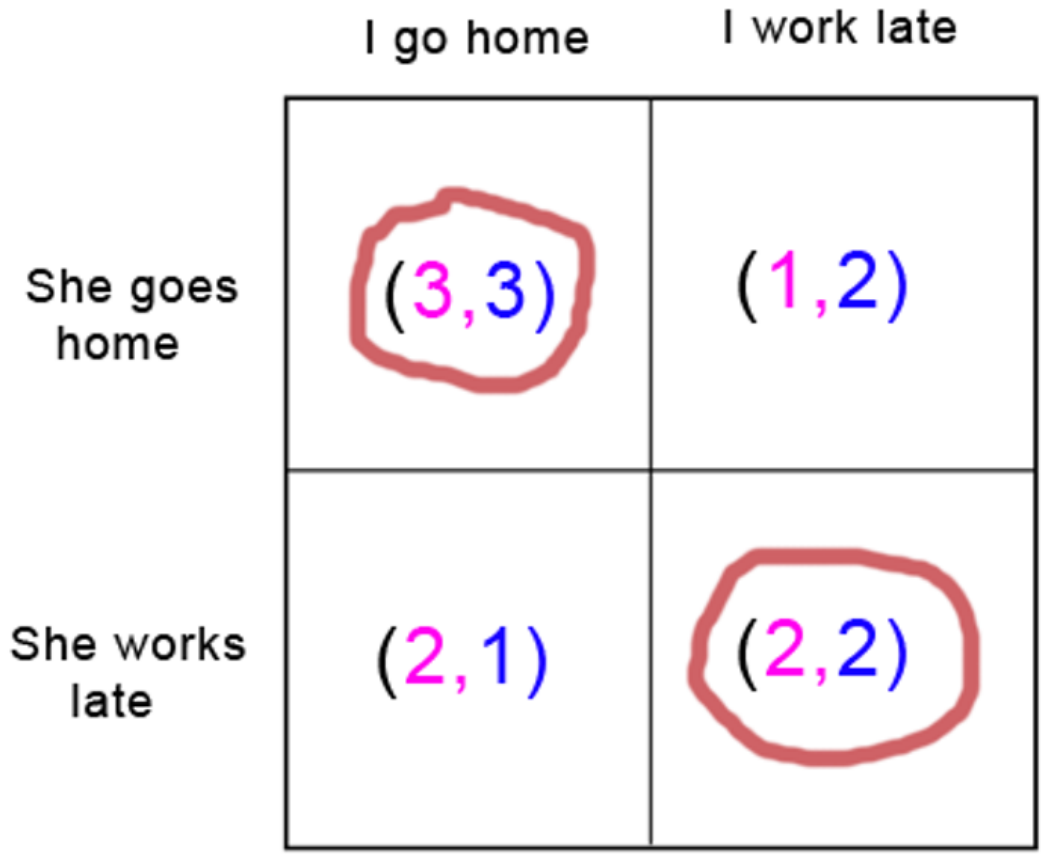
So how can I find a Nash equilibrium? In a Nash equilibrium, I don't regret my strategy when I learn my opponent's action. If I can come up with a strategy that pays exactly the same utility whether my opponent defends Podunk or Metropolis, it will have this useful property. We'll start by supposing I am flipping a *biased* coin that lands on Metropolis x percent of the time, and therefore on Podunk (1-x) percent of the time. To be truly indifferent which city my opponent defends, 10x (the utility my strategy earns when my opponent leaves Metropolis undefended) should equal 1(1-x) (the utility my strategy earns when my opponent leaves Podunk undefended). Some quick algebra finds that 10x = 1(1-x) is satisfied by x = 1/11. So I should attack Metropolis 1/11 of the time and Podunk 10/11 of the time.

My opponent, going through a similar process, comes up with the suspiciously similar result that he should defend Metropolis 10/11 of the time, and Podunk 1/11 of the time.

If we both pursue our chosen strategies, I gain an average 0.9090... utility each round, soundly beating my previous record of 0.5, and my opponent suspiciously loses an average -.9090 utility. It turns out there is no other strategy I can use to consistently do better than this when my opponent is playing optimally, and that even if I knew my opponent's strategy I would not be able to come up with a better strategy to beat it. It also turns out that there is no other strategy my opponent can use to consistently do better than this if I am playing optimally, and that my opponent, upon learning my strategy, doesn't regret his strategy either.

In *The Art of Strategy*, Dixit and Nalebuff cite a real-life application of the same principle in, of all things, penalty kicks in soccer. A right-footed kicker has a better chance of success if he kicks to the right, but a smart goalie can predict that and will defend to the right; a player expecting this can accept a less spectacular kick to the left if he thinks the left will be undefended, but a very smart goalie can predict this too, and so on. Economist Ignacio Palacios-Huerta laboriously analyzed the success rates of various kickers and goalies on the field, and found that they actually pursued a mixed strategy generally within 2% of the game theoretic ideal, proving that people are pretty good at doing these kinds of calculations unconsciously.

So every game really does have at least one Nash equilibrium, even if it's only a mixed strategy. But some games can have many, many more. Recall the situation between me and my girlfriend:



There are two Nash equilibria: both of us working late, and both of us going home. If there were only one equilibrium, and we were both confident in each other's rationality, we could choose that one and there would be no further problem. But in fact this game does present a problem: intuitively it seems like we might still make a mistake and end up in different places.

Here we might be tempted to just leave it to chance; after all, there's a 50% probability we'll both end up choosing the same activity. But other games might have thousands or millions of possible equilibria and so will require a more refined approach.

Art of Strategy describes a game show in which two strangers were separately taken to random places in New York and promised a prize if they could successfully meet up; they had no communication with one another and no clues about how such a meeting was to take place. Here there are a nearly infinite number of possible choices: they could both meet at the corner of First Street and First Avenue at 1 PM, they could both meet at First Street and Second Avenue at 1:05 PM, etc. Since neither party would regret their actions (if I went to First and First at 1 and found you there, I would be thrilled) these are all Nash equilibria.

Despite this mind-boggling array of possibilities, in fact all six episodes of this particular game ended with the two contestants meeting successfully after only a few days. The most popular meeting site was the Empire State Building at noon.

How did they do it? The world-famous Empire State Building is what game theorists call focal: it stands out as a natural and obvious target for coordination. Likewise noon, classically considered the very middle of the day, is a focal point in time. These focal points, also called

Schelling points after theorist Thomas Schelling who discovered them, provide an obvious target for coordination attempts.

What makes a Schelling point? The most important factor is that it be special. The Empire State Building, depending on when the show took place, may have been the tallest building in New York; noon is the only time that fits the criteria of “exactly in the middle of the day”, except maybe midnight when people would be expected to be too sleepy to meet up properly.

Of course, specialness, like beauty, is in the eye of the beholder. David Friedman writes:

> *Two people are separately confronted with the list of numbers [2, 5, 9, 25, 69, 73, 82, 96, 100, 126, 150] and offered a reward if they independently choose the same number. If the two are mathematicians, it is likely that they will both choose 2—the only even prime. Non-mathematicians are likely to choose 100—a number which seems, to the mathematicians, no more unique than the other two exact squares. Illiterates might agree on 69, because of its peculiar symmetry—as would, for a different reason, those whose interest in numbers is more prurient than mathematical.*

A recent open thread comment pointed out that you can justify anything with “for decision-theoretic reasons” or “due to meta-level concerns”. I humbly propose adding “as a Schelling point” to this list, except that the list is tongue-in-cheek and Schelling points really do explain almost everything - stock markets, national borders, marriages, private property, religions, fashion, political parties, peace treaties, social networks, software platforms and languages all involve or are based upon Schelling points. In fact, whenever something has “symbolic value” a Schelling point is likely to be involved in some way. I hope to expand on this point a bit more later.

Sequential games can include one more method of choosing between Nash equilibria: the idea of a subgame-perfect equilibrium, a special kind of Nash equlibrium that remains a Nash equilibrium for every subgame of the original game. In more intuitive terms, this equilibrium means that even in a long multiple-move game no one at any point makes a decision that goes against their best interests (remember the example from the last post, where we crossed out the branches in which Clinton made implausible choices that failed to maximize his utility?) Some games have multiple Nash equilibria but only one subgame-perfect one; we'll examine this idea further when we get to the iterated prisoners' dilemma and ultimatum game.

In conclusion, every game has at least one Nash equilibrium, a point at which neither player regrets her strategy even when she knows the other player's strategy. Some equilibria are simple choices, others involve plans to make choices randomly according to certain criteria. Purely rational players will always end up at a Nash equilibrium, but many games will have multiple possible equilibria. If players are trying to coordinate, they may land at a Schelling point, an equilibria which stands out as special in some way.

Introduction to Prisoners' Dilemma

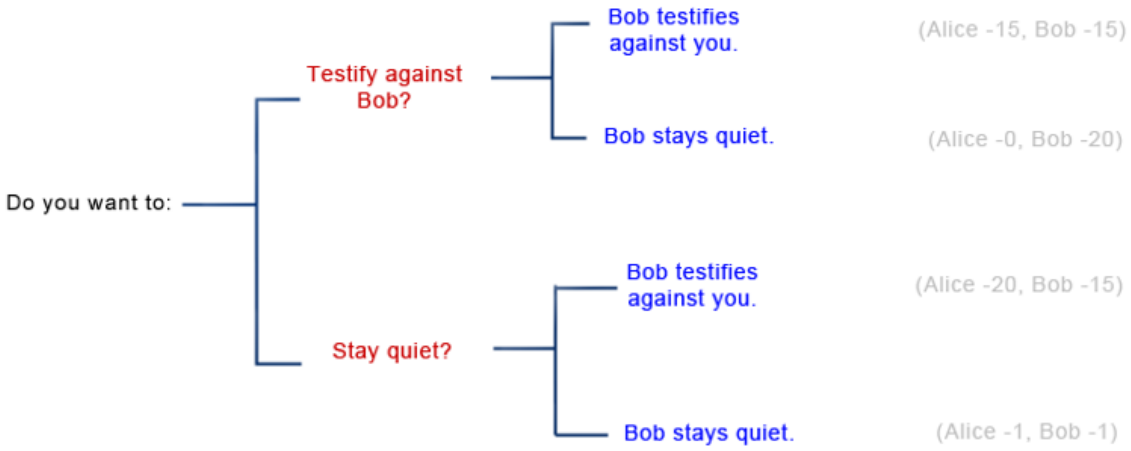
Related to: Previous posts on the Prisoners' Dilemma

Sometimes Nash equilibria just don't match our intuitive criteria for a good outcome. The classic example is the Prisoners' Dilemma.

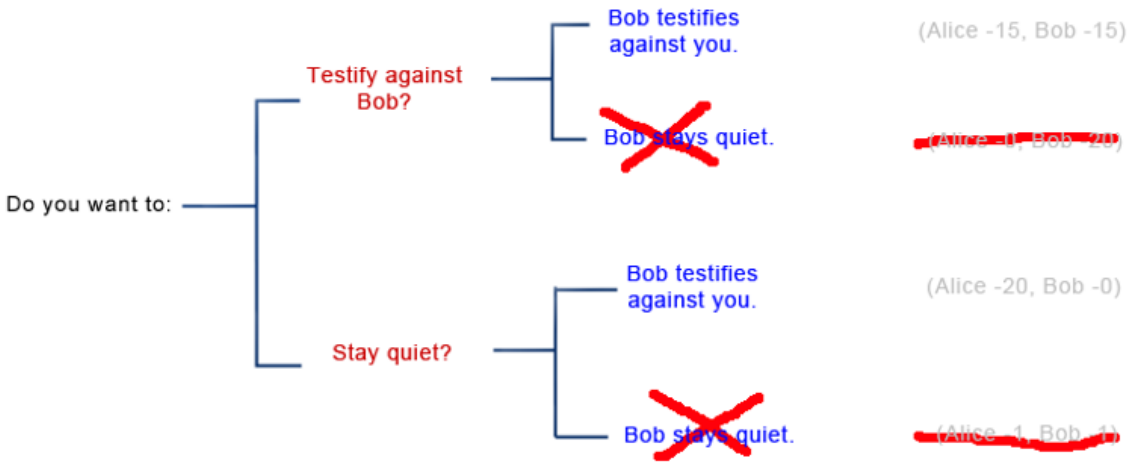
The police arrest two criminals, Alice and Bob, on suspicion of murder. The police admit they don't have enough evidence to convict the pair of murder, but they do have enough evidence to convict them of a lesser offence, possession of a firearm. They place Alice and Bob in separate cells and offer them the following deal:

“If neither of you confess, we'll have to charge you with possession, which will land you one year in jail. But if you turn state's witness against your partner, we can convict your partner of murder and give her the full twenty year sentence; in exchange, we will let you go free. Unless, that is, both of you testify against each other; in that case, we'll give you both fifteen years.”

Alice's decision tree looks like this (note that although Alice and Bob make their decisions simultaneously, I've represented it with Alice's decision first, which is a little sketchy but should illustrate the point):



If we use the same strategy we used as a chess player, we can cross out options where Bob decides to spend extra years in jail for no personal benefit, and we're left with this:



Seen like this, the choice is clear. If you stay quiet (“cooperate”), Bob turns on you, and you are left in jail alone for twenty years, wailing and gnashing your teeth. So instead, you both turn on each other (“defect”), and end up with a sentence barely any shorter.

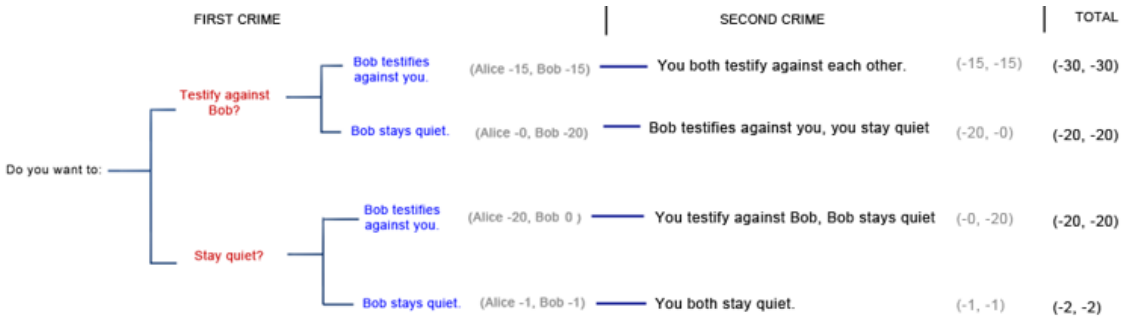
Another way to “prove” that defection is the “right” choice places Bob's decision first. What if you knew Bob would choose to cooperate with you? Then your choice would be between defecting and walking free, or cooperating and spending a whole year in jail - here defection wins out. But what if you knew Bob would choose to defect against you? Then your choice would be between defecting and losing fifteen years, or cooperating and losing twenty - again defection wins out. Since Bob can only either defect or cooperate, and since defection is better in both branches, “clearly” defection is the best option.

But a lot of things about this solution seem intuitively stupid. For example, when Bob goes through the same reasoning, your “rational” solution ends up with both of you in jail for fifteen years, but if you had both cooperated, you would have been out after a year. Both cooperating is better for both of you than both defecting, but you still both defect.

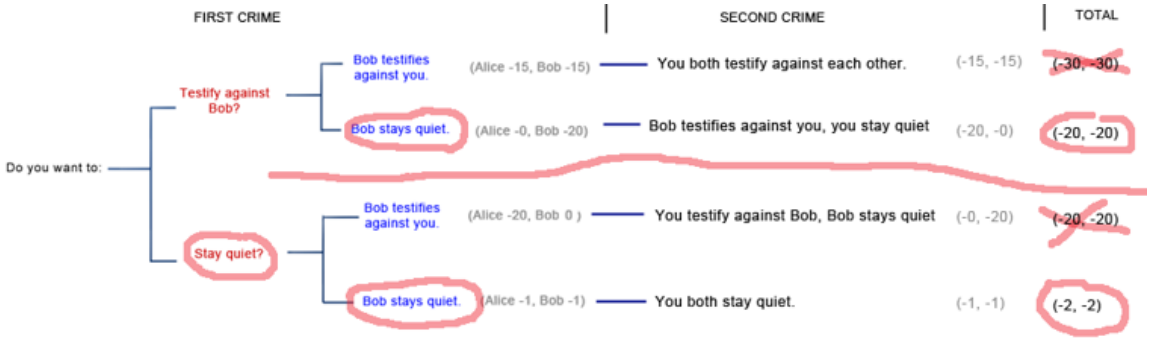
And if you still don't find that odd, imagine a different jurisdiction where the sentence for possession is only one day, and the police will only take a single day off your sentence for testifying against an accomplice. Now a pair of cooperators would end up with only a day in jail each, and a pair of defectors would end up with nineteen years, three hundred sixty four days each. Yet the math still tells you to defect!

Unfortunately, your cooperation only helps Bob, and Bob's cooperation only helps you. We can think of the Prisoner's Dilemma as a problem: both you and Bob prefer (cooperate, cooperate) to (defect, defect), but as it is, you're both going to end out with (defect, defect) and it doesn't seem like there's much you can do about it. To “solve” the Prisoner's Dilemma would be to come up with a way to make you and Bob pick the more desirable (cooperate, cooperate) outcome.

One proposed solution to the Prisoner's Dilemma is to iterate it - to assume it will happen multiple times in succession, as if Alice and Bob are going to commit new crimes as soon as they both get out of prison. In this case, you can threaten to reciprocate; to promise to reward cooperation with future cooperation and punish defection with future defection. Suppose Alice and Bob plan to commit two crimes, and before the first crime both promise to stay quiet on the second crime if and only if their partner stays quiet on the first. Now your decision tree as Alice looks like this:



And your calculation of Bob's thought processes go like this:



Remember that, despite how the graph looks, your first choice and Bob's first choice are simultaneous: they can't causally affect each other. So as Alice, you reason like this: On the top, Bob knows that if you testify against him, his choice will be either to testify against you (leading to the branch where you both testify against each other again next time) or to stay quiet (leading to a branch where next time he testifies against you but you stay quiet). So Bob reasons that if you testify against him, he should stay quiet this time.

On the bottom, Bob knows that if you don't testify against him, he can either testify against you (leading to the branch where you testify against him next time but he stays quiet) or stay quiet (leading to the branch where you both stay quiet again next time). Therefore, if you don't testify against him, Bob won't testify against you.

So you know that no matter what you do this time, Bob won't testify against you. That means your choice is between branches 2 and 4: Bob testifying against you next time or Bob not testifying against you next time. You prefer Branch 4, so you decide not to testify against Bob. The dilemma ends with neither of you testifying against each other in either crime, and both of you getting away with very light two year sentences.

The teeny tiny little flaw in this plan is that Bob may be a dirty rotten liar. Maybe he says he'll reciprocate, and so you both stay quiet after the first crime. Upon getting out of jail you continue your crime spree, predictably get re-arrested, and you stay quiet like you said you would to reward Bob's cooperation last time. But at the trial, you get a nasty surprise: Bob defects against you and walks free, and you end up with a twenty year sentence.

If we ratchet up to sprees of one hundred crimes and subsequent sentences (presumably committed by immortal criminals who stubbornly refuse to be cowed by the police's 100% conviction rate) on first glance it looks like we can successfully ensure Bob's cooperation on 99 of those crimes. After all, Bob won't want to defect on crime 50, because I could punish him on crime 51. He won't want to defect on crime 99, because I could punish him on crime

100. But he *will* want to defect on crime 100, because he gains either way and there's nothing I can do to punish him.

Here's where it gets weird. I assume Bob is a rational utility-maximizer and so will always defect on crime 100, since it benefits him and I can't punish him for it. So since I'm also rational, I might as well also defect on crime 100; my previous incentive to cooperate was to ensure Bob's good behavior, but since Bob won't show good behavior on crime 100 no matter what I do, I might as well look after my own interests.

But if we both know that we're both going to defect on crime 100 no matter what, then there's no incentive to cooperate on crime 99. After all, the only incentive to cooperate on crime 99 was to ensure my rival's cooperation on crime 100, and since that's out of the picture anyway, I might as well shorten my sentence a little.

Sadly, this generalizes by a sort of proof by induction. If crime N will always be (defect, defect), then crime N-1 should also always be (defect, defect), which means we should defect on all of the hundred crimes in our spree.

This feat of reasoning has limited value in real life, where perfectly rational immortal criminals rarely plot in smoke-filled rooms to commit exactly one hundred crimes together; criminals who are uncertain exactly when their crime sprees will come to a close still have incentive to cooperate. But it still looks like we're going to need a better solution than simply iterating the dilemma. The next post will discuss possibilities for such a solution.

Real World Solutions to Prisoners' Dilemmas

Why should there be real world solutions to Prisoners' Dilemmas? Because such dilemmas are a real-world problem.

If I am assigned to work on a school project with a group, I can either cooperate (work hard on the project) or defect (slack off while reaping the rewards of everyone else's hard work). If everyone defects, the project doesn't get done and we all fail - a bad outcome for everyone. If I defect but you cooperate, then I get to spend all day on the beach and still get a good grade - the best outcome for me, the worst for you. And if we all cooperate, then it's long hours in the library but at least we pass the class - a “good enough” outcome, though not quite as good as me defecting against everyone else's cooperation. This exactly mirrors the Prisoner's Dilemma.

Diplomacy - both the concept and the board game - involves Prisoners' Dilemmas. Suppose Ribbentrop of Germany and Molotov of Russia agree to a peace treaty that demilitarizes their mutual border. If both cooperate, they can move their forces to other theaters, and have moderate success there - a good enough outcome. If Russia cooperates but Germany defects, it can launch a surprise attack on an undefended Russian border and enjoy spectacular success there (for a while, at least!) - the best outcome for Germany and the worst for Russia. But if both defect, then neither has any advantage at the German-Russian border, and they lose the use of those troops in other theaters as well - a bad outcome for both. Again, the Prisoner's Dilemma.

Civilization - again, both the concept and the game - involves Prisoners' Dilemmas. If everyone follows the rules and creates a stable society (cooperates), we all do pretty well. If everyone else works hard and I turn barbarian and pillage you (defect), then I get all of your stuff without having to work for it and you get nothing - the best solution for me, the worst for you. If everyone becomes a barbarian, there's nothing to steal and we all lose out. Prisoner's Dilemma.

If everyone who worries about global warming cooperates in cutting emissions, climate change is averted and everyone is moderately happy. If everyone else cooperates in cutting emissions, but one country defects, climate change is still mostly averted, and the defector is at a significant economic advantage. If everyone defects and keeps polluting, the climate changes and everyone loses out. Again a Prisoner's Dilemma,

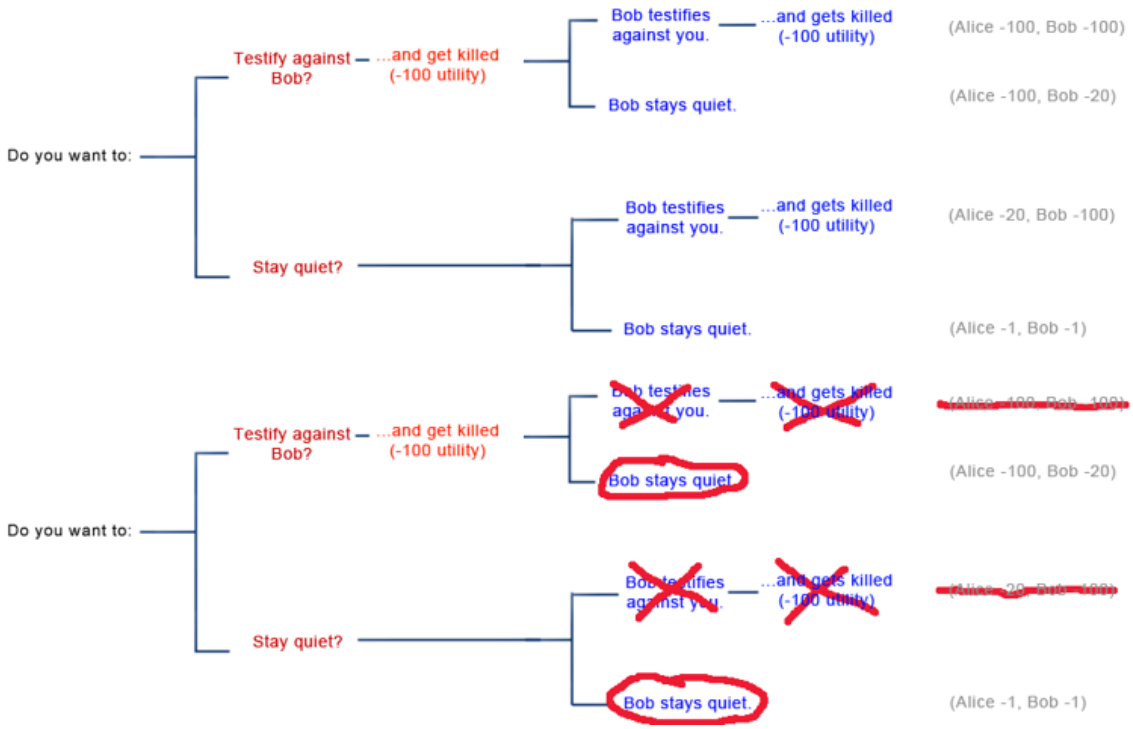
Prisoners' Dilemmas even come up in nature. In baboon tribes, when a female is in “heat”, males often compete for the chance to woo her. The most successful males are those who can get a friend to help fight off the other monkeys, and who then helps that friend find his own monkey loving. But these monkeys are tempted to take their friend's female as well. Two males who cooperate each seduce one female. If one cooperates and the other defects, he has a good chance at both females. But if the two can't cooperate at all, then they will be beaten off by other monkey alliances and won't get to have sex with anyone. Still a Prisoner's Dilemma!

So one might expect the real world to have produced some practical solutions to Prisoners' Dilemmas.

One of the best known such systems is called “society”. You may have heard of it. It boasts a series of norms, laws, and authority figures who will punish you when those norms and laws are broken.

Imagine that the two criminals in the original example were part of a criminal society - let's say the Mafia. The Godfather makes Alice and Bob an offer they can't refuse: turn against

one another, and they will end up “sleeping with the fishes” (this concludes my knowledge of the Mafia). Now the incentives are changed: defecting against a cooperator doesn't mean walking free, it means getting murdered.



Both prisoners cooperate, and amazingly the threat of murder ends up making them both better off (this is also the gist of some of the strongest arguments against libertarianism: in Prisoner's Dilemmas, threatening force against rational agents can increase the utility of all of them!)

Even when there is no godfather, society binds people by concern about their “reputation”. If Bob got a reputation as a snitch, he might never be able to work as a criminal again. If a student gets a reputation for slacking off on projects, she might get ostracized on the playground. If a country gets a reputation for backstabbing, others might refuse to make treaties with them. If a person gets a reputation as a bandit, she might incur the hostility of those around her. If a country gets a reputation for not doing enough to fight global warming, it might...well, no one ever said it was a perfect system.

Aside from humans in society, evolution is also strongly motivated to develop a solution to the Prisoner's Dilemma. The Dilemma troubles not only lovestruck baboons, but ants, minnows, bats, and even viruses. Here the payoff is denominated not in years of jail time, nor in dollars, but in reproductive fitness and number of potential offspring - so evolution will certainly take note.

Most people, when they hear the rational arguments in favor of defecting every single time on the iterated 100-crime Prisoner's Dilemma, will feel some kind of emotional resistance. Thoughts like “Well, maybe I'll try cooperating anyway a few times, see if it works”, or “If I promised to cooperate with my opponent, then it would be dishonorable for me to defect on the last turn, even if it helps me out., or even “Bob is my friend! Think of all the good times we've had together, robbing banks and running straight into waiting police cordons. I could never betray him!”

And if two people with these sorts of emotional hangups play the Prisoner's Dilemma together, they'll end up cooperating on all hundred crimes, getting out of jail in a mere

century and leaving rational utility maximizers to sit back and wonder how they did it.

Here's how: imagine you are a supervillain designing a robotic criminal (who's that go-to supervillain Kaj always uses for situations like this? Dr. Zany? Okay, let's say you're him). You expect to build several copies of this robot to work as a team, and expect they might end up playing the Prisoner's Dilemma against each other. You want them out of jail as fast as possible so they can get back to furthering your nefarious plots. So rather than have them bumble through the whole rational utility maximizing thing, you just insert an extra line of code: “in a Prisoner's Dilemma, always cooperate with other robots”. Problem solved.

Evolution followed the same strategy (no it didn't; this is a massive oversimplification). The emotions we feel around friendship, trust, altruism, and betrayal are partly a built-in hack to succeed in cooperating on Prisoner's Dilemmas where a rational utility-maximizer would defect a hundred times and fail miserably. The evolutionarily dominant strategy is commonly called “Tit-for-tat” - basically, cooperate if and only if your opponent did so last time.

This so-called "superrationality” appears even more clearly in the Ultimatum Game. Two players are given $100 to distribute among themselves in the following way: the first player proposes a distribution (for example, “Fifty for me, fifty for you”) and then the second player either accepts or rejects the distribution. If the second player accepts, the players get the money in that particular ratio. If the second player refuses, no one gets any money at all.

The first player's reasoning goes like this: “If I propose $99 for myself and $1 for my opponent, that means I get a lot of money and my opponent still has to accept. After all, she prefers $1 to $0, which is what she'll get if she refuses.

In the Prisoner's Dilemma, when players were able to communicate beforehand they could settle upon a winning strategy of precommiting to reciprocate: to take an action beneficial to their opponent if and only if their opponent took an action beneficial to them. Here, the second player should consider the same strategy: precommit to an ultimatum (hence the name) that unless Player 1 distributes the money 50-50, she will reject the offer.

But as in the Prisoner's Dilemma, this fails when you have no reason to expect your opponent to follow through on her precommitment. Imagine you're Player 2, playing a single Ultimatum Game against an opponent you never expect to meet again. You dutifully promise Player 1 that you will reject any offer less than 50-50. Player 1 offers 80-20 anyway. You reason “Well, my ultimatum failed. If I stick to it anyway, I walk away with nothing. I might as well admit it was a good try, give in, and take the $20. After all, rejecting the offer won't magically bring my chance at $50 back, and there aren't any other dealings with this Player 1 guy for it to influence.”

This is seemingly a rational way to think, but if Player 1 knows you're going to think that way, she offers 99-1, same as before, no matter how sincere your ultimatum sounds.

Notice all the similarities to the Prisoner's Dilemma: playing as a "rational economic agent" gets you a bad result, it looks like you can escape that bad result by making precommitments, but since the other player can't trust your precommitments, you're right back where you started

If evolutionary solutions to the Prisoners' Dilemma look like trust or friendship or altruism, solutions to the Ultimatum Game involve different emotions entirely. The Sultan presumably does not want you to elope with his daughter. He makes an ultimatum: “Touch my daughter, and I will kill you.” You elope with her anyway, and when his guards drag you back to his palace, you argue: “Killing me isn't going to reverse what happened. Your ultimatum has failed. All you can do now by beheading me is get blood all over your beautiful palace carpet, which hurts you as well as me - the equivalent of pointlessly passing up the last dollar in an Ultimatum Game where you've just been offered a 99-1 split.”

The Sultan might counter with an argument from social institutions: “If I let you go, I will look dishonorable. I will gain a reputation as someone people can mess with without any consequences. My choice isn't between bloody carpet and clean carpet, it's between bloody carpet and people respecting my orders, or clean carpet and people continuing to defy me.”

But he's much more likely to just shout an incoherent stream of dreadful Arabic curse words. Because just as friendship is the evolutionary solution to a Prisoner's Dilemma, so anger is the evolutionary solution to an Ultimatum Game. As various gurus and psychologists have observed, anger makes us irrational. But this is the good kind of irrationality; it's the kind of irrationality that makes us pass up a 99-1 split even though the decision costs us a dollar.

And if we know that humans are the kind of life-form that tends to experience anger, then if we're playing an Ultimatum Game against a human, and that human precommits to rejecting any offer less than 50-50, we're much more likely to believe her than if we were playing against a rational utility-maximizing agent - and so much more likely to give the human a fair offer.

It is distasteful and a little bit contradictory to the spirit of rationality to believe it should lose out so badly to simple emotion, and the problem might be correctable. Here we risk crossing the poorly charted border between game theory and decision theory and reaching ideas like timeless decision theory: that one should act as if one's choices determined the output of the algorithm one instantiates (or more simply, you should assume everyone like you will make the same choice you do, and take that into account when choosing.)

More practically, however, most real-world solutions to Prisoner's Dilemmas and Ultimatum Games still hinge on one of three things: threats of reciprocation when the length of the game is unknown, social institutions and reputation systems that make defection less attractive, and emotions ranging from cooperation to anger that are hard-wired into us by evolution. In the next post, we'll look at how these play out in practice.

Interlude for Behavioral Economics

The so-called “rational” solutions to the Prisoners' Dilemma and Ultimatum Game are suboptimal to say the least. Humans have various kludges added by both nature or nurture to do better, but they're not perfect and they're certainly not simple. They leave entirely open the question of what real people will actually do in these situations, a question which can only be addressed by hard data.

As in so many other areas, our most important information comes from reality television. *The Art of Strategy* discusses a US game show “Friend or Foe” where a team of two contestants earned money by answering trivia questions. At the end of the show, the team used a sort-of Prisoner's Dilemma to split their winnings: each team member chose “Friend” (cooperate) or “Foe” (defect). If one player cooperated and the other defected, the defector kept 100% of the pot. If both cooperated, each kept 50%. And if both defected, neither kept anything (this is a significant difference from the standard dilemma, where a player is a little better off defecting than cooperating if her opponent defects).

Players chose “Friend” about 45% of the time. Significantly, this number remained constant despite the size of the pot: they were no more likely to cooperate when splitting small amounts of money than large.

Players seemed to want to play “Friend” if and only if they expected their opponents to do so. This is not rational, but it accords with the “Tit-for-Tat” strategy hypothesized to be the evolutionary solution to Prisoner's Dilemma. This played out on the show in a surprising way: players' choices started off random, but as the show went on and contestants began participating who had seen previous episodes, they began to base their decision on observable characteristics about their opponents. For example, in the first season women cooperated more often than men, so by the second season a player was cooperating more often if their opponent was a woman - whether or not that player was a man or woman themselves.

Among the superficial characteristics used, the only one to reach statistical significance according to the study was age: players below the median age of 27 played “Foe” more often than those over it (65% vs. 39%, $p < .001$). Other nonsignificant tendencies were for men to defect more than women (53% vs. 46%, $p=.34$) and for black people to defect more than white people (58% vs. 48%, $p=.33$). These nonsignificant tendencies became important because the players themselves attributed significance to them: for example, by the second season women were playing “Foe” 60% of the time against men but only 45% of the time against women ($p<.01$) presumably because women were perceived to be more likely to play “Friend” back; also during the second season, white people would play “Foe” 75% against black people, but only 54% of the time against other white people.

(This risks self-fulfilling prophecies. If I am a black man playing a white woman, I expect she will expect me to play “Foe” against her, and she will “reciprocate” by playing “Foe” herself. Therefore, I may choose to “reciprocate” against her by playing “Foe” myself, even if I wasn't originally intending to do so, and other white women might observe this, thus creating a vicious cycle.)

In any case, these attempts at coordinated play worked, but only imperfectly. By the second season, 57% of pairs chose the same option - either (C, C) or (D, D).

Art of Strategy included another great Prisoner's Dilemma experiment. In this one, the experimenters spoiled the game: they told both players that they would be deciding simultaneously, but in fact, they let Player 1 decide first, and then secretly approached Player 2 and told her Player 1's decision, letting Player 2 consider this information when making her own choice.

Why should this be interesting? From the previous data, we know that humans play “tit-for-expected-tat”: they will generally cooperate if they believe their opponent will cooperate too. We can come up with two hypotheses to explain this behavior. First, this could be a folk version of Timeless Decision Theory or Hofstadter's superrationality; a belief that their own decision literally determines their opponent's decision. Second, it could be based on a belief in fairness: if I think my opponent cooperated, it's only decent that I do the same.

The “researchers spoil the setup” experiment can distinguish between these two hypotheses. If people believe their choice determines that of their opponent, then once they know their opponent's choice they no longer have to worry and can freely defect to maximize their own winnings. But if people want to cooperate to reward their opponent, then learning that their opponent cooperated for sure should only increase their willingness to reciprocate.

The results: If you tell the second player that the first player defected, 3% still cooperate (apparently 3% of people are Jesus). If you tell the second player that the first player cooperated.....only 16% cooperate. When the same researchers in the same lab didn't tell the second player anything, 37% cooperated.

This is a pretty resounding victory for the “folk version of superrationality” hypothesis. 21% of people wouldn't cooperate if they heard their opponent defected, wouldn't cooperate if they heard their opponent cooperated, but will cooperate if they don't know which of those two their opponent played.

Moving on to the Ultimatum Game: very broadly, the first player usually offers between 30 and 50 percent, and the second player tends to accept. If the first player offers less than about 20 percent, the second player tends to reject it.

Like the Prisoner's Dilemma, the amount of money at stake doesn't seem to matter. This is really surprising! Imagine you played an Ultimatum Game for a billion dollars. The first player proposes $990 million for herself, $10 million for you. On the one hand, this is a 99-1 split, just as unfair as $99 versus $1. On the other hand, ten million dollars!

Although tycoons have yet to donate a billion dollars to use for Ultimatum Game experiments, researchers have done the next best thing and flown out to Third World countries where even $100 can be an impressive amount of money. In games in Indonesia played for a pot containing a sixth of Indonesians' average yearly income, Indonesians still rejected unfair offers. In fact, at these levels the first player tended to propose fairer deals than at lower stakes - maybe because it would be a disaster if her offer get rejected.

It was originally believed that results in the Ultimatum Game were mostly independent of culture. Groups in the US, Israel, Japan, Eastern Europe, and Indonesia all got more or less the same results. But this elegant simplicity was, like so many other things, ruined by the Machiguenga Indians of eastern Peru. They tend to make

offers around 25%, and will accept pretty much anything.

One more interesting finding: people who accept low offers in the Ultimatum Game have lower testosterone than those who reject them.

There is a certain degenerate form of the Ultimatum Game called the Dictator Game. In the Dictator Game, the second player doesn't have the option of vetoing the first player's distribution. In fact, the second player doesn't do anything at all; the first player distributes the money, both players receive the amount of money the first player decided upon, and the game ends. A perfectly selfish first player would take 100% of the money in the Dictator Game, leaving the second player with nothing.

In a metaanalysis of 129 papers consisting of over 41,000 individual games, the average amount the first player gave the second player was 28.35%. 36% of first players take everything, 17% divide the pot equally, and 5% give everything to the second player, nearly doubling our previous estimate of what percent of people are Jesus.

The meta-analysis checks many different results, most of which are insignificant, but a few stand out. Subjects playing the dictator game “against” a charity are much more generous; up to a quarter give everything. When the experimenter promises to “match” each dollar given away (eg the dictator gets $100, but if she gives it to the second player the second player gets $200), the dictator gives much more (somewhat surprising, as this might be an excuse to keep $66 for yourself and get away with it by claiming that both players still got equal money). On the other hand, if the experimenters give the second player a free $100, so that they start off richer than the dictator, the dictator compensates by not giving them nearly as much money.

Old people give more than young people, and non-students give more than students. People from “primitive” societies give more than people from more developed societies, and the more primitive the society, the stronger the effect. The most important factor, though? As always, sex. Women both give more and get more in dictator games.

It is somewhat inspiring that so many people give so much in this game, but before we become too excited about the fundamental goodness of humanity, Art of Strategy mentions a great experiment by Dana, Cain, and Dawes. The subjects were offered a choice: either play the Dictator Game with a second player for $10, or get $9 and the second subject is sent home and never even knows what the experiment is about. A third of participants took the second option.

So generosity in the Dictator Game isn't always about wanting to help other people. It seems to be about knowing, deep down, that some anonymous person who probably doesn't even know your name and who will never see you again is disappointed in you. Remove the little problem of the other person knowing what you did, and they will not only keep the money, but even be willing to pay the experiment a dollar to keep them quiet.

What Is Signaling, Really?

The most commonly used introduction to signaling, promoted both by Robin Hanson and in *The Art of Strategy*, starts with college degrees. Suppose, there are two kinds of people, smart people and stupid people; and suppose, with wild starry-eyed optimism, that the populace is split 50-50 between them. Smart people would add enough value to a company to be worth a $100,000 salary each year, but stupid people would only be worth $40,000. And employers, no matter how hard they try to come up with silly lateral-thinking interview questions like “How many ping-pong balls could fit in the Sistine Chapel?”, can't tell the difference between them.

Now suppose a certain college course, which costs $50,000, passes all smart people but flunks half the stupid people. A strategic employer might declare a policy of hiring (for a one year job; let's keep this model simple) graduates at $100,000 and non-graduates at $40,000.

Why? Consider the thought process of a smart person when deciding whether or not to take the course. She thinks “I am smart, so if I take the course, I will certainly pass. Then I will make an extra $60,000 at this job. So my costs are $50,000, and my benefits are $60,000. Sounds like a good deal.”

The stupid person, on the other hand, thinks: “As a stupid person, if I take the course, I have a 50% chance of passing and making $60,000 extra, and a 50% chance of failing and making $0 extra. My expected benefit is $30,000, but my expected cost is $50,000. I'll stay out of school and take the $40,000 salary for non-graduates.”

...assuming that stupid people all know they're stupid, and that they're all perfectly rational experts at game theory, to name two of several dubious premises here. Yet despite its flaws, this model does give some interesting results. For example, it suggests that rational employers will base decisions upon - and rational employees enroll in - college courses, even if those courses teach nothing of any value. So an investment bank might reject someone who had no college education, even while hiring someone who studied Art History, not known for its relevance to derivative trading.

We'll return to the specific example of education later, but for now it is more important to focus on the general definition that X signals Y if X is more likely to be true when Y is true than when Y is false. Amoral self-interested agents after the $60,000 salary bonus for intelligence, whether they are smart or stupid, will always say “Yes, I'm smart” if you ask them. So saying “I am smart” is not a signal of intelligence. Having a college degree is a signal of intelligence, because a smart person is more likely to get one than a stupid person.

Life frequently throws us into situations where we want to convince other people of something. If we are employees, we want to convince bosses we are skillful, honest, and hard-working. If we run the company, we want to convince customers we have superior products. If we are on the dating scene, we want to show potential mates that we are charming, funny, wealthy, interesting, you name it.

In some of these cases, mere assertion goes a long way. If I tell my employer at a job interview that I speak fluent Spanish, I'll probably get asked to talk to a Spanish-speaker at my job, will either succeed or fail, and if I fail will have a lot of questions to

answer and probably get fired - or at the very least be in more trouble than if I'd just admitted I didn't speak Spanish to begin with. Here society and its system of reputational penalties help turn mere assertion into a credible signal: asserting I speak Spanish is costlier if I don't speak Spanish than if I do, and so is believable.

In other cases, mere assertion doesn't work. If I'm at a seedy bar looking for a one-night stand, I can tell a girl I'm totally a multimillionaire and feel relatively sure I won't be found out until after that one night - and so in this she would be naive to believe me, unless I did something only a real multimillionaire could, like give her an expensive diamond necklace.

How expensive a diamond necklace, exactly? To absolutely prove I am a millionaire, only a million dollars worth of diamonds will do; $10,000 worth of diamonds could in theory come from anyone with at least $10,000. But in practice, people only care so much about impressing a girl at a seedy bar; if everyone cares about the same amount, the amount they'll spend on the signal depends mostly on their marginal utility of money, which in turn depends mostly on how much they have. Both a millionaire and a tenthousandaire can afford to buy $10,000 worth of diamonds, but only the millionaire can afford to buy $10,000 worth of diamonds on a whim. If in general people are only willing to spend 1% of their money on an impulse gift, then $10,000 is sufficient evidence that I am a millionaire.

But when the stakes are high, signals can get prohibitively costly. If a dozen millionaires are wooing Helen of Troy, the most beautiful woman in the world, and willing to spend arbitrarily much money on her - and if they all believe Helen will choose the richest among them - then if I only spend $10,000 on her I'll be outshone by a millionaire who spends the full million. Thus, if I want any chance with her at all, then even if I am genuinely the richest man around I might have to squander my entire fortune on diamonds.

This raises an important point: *signaling can be really horrible*. What if none of us are entirely sure how much Helen's other suitors have? It might be rational for all of us to spend everything we have on diamonds for her. Then twelve millionaires lose their fortunes, eleven of them for nothing. And this isn't some kind of wealth transfer - for all we know, Helen might not even like diamonds; maybe she locks them in her jewelry box after the wedding and never thinks about them again. It's about as economically productive as digging a big hole and throwing money into it.

If all twelve millionaires could get together beforehand and compare their wealth, and agree that only the wealthiest one would woo Helen, then they could all save their fortunes and the result would be exactly the same: Helen marries the wealthiest. If all twelve millionaires are remarkably trustworthy, maybe they can pull it off. But if any of them believe the others might lie about their wealth, or that one of the poorer men might covertly break their pact and woo Helen with gifts, then they've got to go through with the whole awful “everyone wastes everything they have on shiny rocks” ordeal.

Examples of destructive signaling are not limited to hypotheticals. Even if one does not believe Jared Diamond's hypothesis that Easter Island civilization collapsed after chieftains expended all of their resources trying to out-signal each other by building larger and larger stone heads, one can look at Nikolai Roussanov's study on how the dynamics of signaling games in US minority communities encourage conspicuous consumption and prevent members of those communities from investing in education and other important goods.

The Art of Strategy even advances the surprising hypothesis that corporate advertising can be a form of signaling. When a company advertises during the Super Bowl or some other high-visibility event, it costs a lot of money. To be able to afford the commercial, the company must be pretty wealthy; which in turn means it probably sells popular products and isn't going to collapse and leave its customers in the lurch. And to want to afford the commercial, the company must be pretty confident in its product: advertising that you should shop at Wal-Mart is more profitable if you shop at Wal-Mart, love it, and keep coming back than if you're likely to go to Wal-Mart, hate it, and leave without buying anything. This signaling, too, can become destructive: if every other company in your industry is buying Super Bowl commercials, then none of them have a comparative advantage and they're in exactly the same relative position as if none of them bought Super Bowl commercials - throwing money away just as in the diamond example.

Most of us cannot afford a Super Bowl commercial or a diamond necklace, and less people may build giant stone heads than during Easter Island's golden age, but a surprising amount of everyday life can be explained by signaling. For example, why did about 50% of readers get a mental flinch and an overpowering urge to correct me when I used “less” instead of “fewer” in the sentence above? According to Paul Fussell's “Guide Through The American Class System” (ht SIAI mailing list), nitpicky attention to good grammar, even when a sentence is perfectly clear without it, can be a way to signal education, and hence intelligence and probably social class. I would not dare to summarize Fussell's guide here, but it shattered my illusion that I mostly avoid thinking about class signals, and instead convinced me that pretty much everything I do from waking up in the morning to going to bed at night is a class signal. On flowers:

> *Anyone imagining that just any sort of flowers can be presented in the front of a house without status jeopardy would be wrong. Upper-middle-class flowers are rhododendrons, tiger lilies, amaryllis, columbine, clematis, and roses, except for bright-red ones. One way to learn which flowers are vulgar is to notice the varieties favored on Sunday-morning TV religious programs like Rex Humbard's or Robert Schuller's. There you will see primarily geraniums (red are lower than pink), poinsettias, and chrysanthemums, and you will know instantly, without even attending to the quality of the discourse, that you are looking at a high-prole setup. Other prole flowers include anything too vividly red, like red tulips. Declassed also are phlox, zinnias, salvia, gladioli, begonias, dahlias, fuchsias, and petunias. Members of the middle class will sometimes hope to mitigate the vulgarity of bright-red flowers by planting them in a rotting wheelbarrow or rowboat displayed on the front lawn, but seldom with success.*

Seriously, read the essay.

In conclusion, a signal is a method of conveying information among not-necessarily-trustworthy parties by performing an action which is more likely or less costly if the information is true than if it is not true. Because signals are often costly, they can sometimes lead to a depressing waste of resources, but in other cases they may be the only way to believably convey important information.

Bargaining and Auctions

Some people have things. Other people want them. Economists agree that the eventual price will be set by supply and demand, but both parties have tragically misplaced their copies of the *Big Book Of Levels Of Supply And Demand For All Goods*. They're going to have to decide on a price by themselves.

When the transaction can be modeled by the interaction of one seller and one buyer, this kind of decision usually looks like bargaining. When it's best modeled as one seller and multiple buyers (or vice versa), the decision usually looks like an auction. Many buyers and many sellers produce a marketplace, but this is complicated and we'll stick to bargains and auctions for now.

Simple bargains bear some similarity to the Ultimatum Game. Suppose an antique dealer has a table she values at $50, and I go to the antique store and fall in love with it, believing it will add $400 worth of classiness to my room. The dealer should never sell for less than $50, and I should never buy for more than $400, but any value in between would benefit both of us. More specifically, it would give us a combined $350 profit. The remaining question is how to divide that $350 pot.

If I make an offer to buy at $60, I'm proposing to split the pot "$10 for you, $340 for me". If the dealer makes a counter-offer of $225, she's offering "$175 for you, $175 for me" - or an even split.

Each round of bargaining resembles the Ultimatum Game because one player proposes to split a pot, and the other player accepts or rejects. If the other player rejects the offer (for example, the dealer refuses to sell it for $60) then the deal falls through and neither of us gets any money.

But bargaining is unlike the Ultimatum Game for several reasons. First, neither player is the designated "offer-maker"; either player may begin by making an offer. Second, the game doesn't end after one round; if the dealer rejects my offer, she can make a counter-offer of her own. Third, and maybe most important, neither player is exactly sure about the size of the pot: I don't walk in knowing that the dealer bought the table for $50, and I may not really be sure I value the table at $400.

Our intuition tells us that the fairest method is to split the profits evenly at a price of $225. This number forms a useful Schelling point (remember those?) that prevents the hassle of further bargaining.

The *Art of Strategy* (see the beginning of Ch. 11) includes a proof that an even split is the rational choice under certain artificial assumptions. Imagine a store selling souvenirs for the 2012 Olympics. They make $1000/day each of the sixteen days the Olympics are going on. Unfortunately, the day before the Olympics, the workers decide to strike; the store will make no money without workers, and they don't have enough time to hire scabs.

Suppose Britain has some very strange labor laws that mandate the following negotiation procedure: on each odd numbered day of the Olympics, the labor union representative will approach the boss and make an offer; the boss can either accept it or reject it. On each even numbered day, the boss makes the offer to the labor union.

So if the negotiations were to drag on to the sixteenth and last day of the Olympics, on that even-numbered day the boss would approach the labor union rep. They're both the sort of straw man rationalists who would take 99-1 splits on the Ultimatum Game, so she offers the labor union rep $1 of the $1000. Since it's the last day of the Olympics and she's a straw man rationalist, the rep accepts.

But on the fifteenth day of the Olympics, the labor union rep will approach the boss. She knows that if no deal is struck today, she'll end out with $1 and the boss will end out with $999. She has to convince the boss to accept a deal on the fifteenth day instead of waiting until the sixteenth. So she offers $1 of the profits from the fifteenth day to the boss, with the labor union keeping the rest; now their totals are $1000 for the workers, $1000 for the boss. Since $1000 is better than $999, the boss agrees to these terms and the strike is ended on the fifteenth day.

We can see by this logic that on odd numbered days the boss and workers get the same amount, and on even numbered days the boss gets more than the workers but the ratio converges to 1:1 as the length of the negotiations increase. If they were negotiating an indefinite contract, then even if the boss made the first move we might expect her to offer an even split.

So both some intuitive and some mathematical arguments lead us to converge on this idea of an even split of the sort that gives us the table for $225. But if I want to be a “hard bargainer” - the kind of person who manages to get the table for less than $225 - I have a couple of things I could try.

I could deceive the seller as to how much I valued the table. This is a pretty traditional bargaining tactic: “That old piece of junk? I'd be doing you a favor for taking it off your hands.” Here I'm implicitly claiming that the dealer must have paid less than $50, and that I would get less than $400 worth of value. If the dealer paid $20 and I'd only value it to the tune of $300, then splitting the profit evenly would mean a final price of $160. The dealer could then be expected to counter my move with his own claim as to the table's value: “$160? Do I look like I was born yesterday? This table was old in the time of the Norman Conquest! Its wood comes from a tree that grows on an enchanted island in the Freptane Sea which appears for only one day every seven years!” The final price might be determined by how plausible we each considered the other's claims.

Or I could rig the Ultimatum Game. Used car dealerships are notorious for adding on “extras” after you've agreed on a price over the phone (“Well yes, we agreed the car was $5999, but if you want a steering wheel, that costs another $200.”) Somebody (possibly an LWer?) proposed showing up to the car dealership without any cash or credit cards, just a check made out for the agreed-upon amount; the dealer now has no choice but to either take the money or forget about the whole deal. In theory, I could go to the antique dealer with a check made out for $60 and he wouldn't have a lot of options (though do remember that people usually reject ultimata of below about 70-30). The classic bargaining tactic of “I am but a poor chimney sweep with only a few dollars to my name and seven small children to feed and I could never afford a price above $60” seems closely related to this strategy.

And although we're still technically talking about transactions with only one buyer and seller, the mere threat of another seller can change the balance of power drastically. Suppose I tell the dealer I know of another dealer who sells modern art for a fixed price of $300, and that the modern art would add exactly as much classiness to my room as this antique table - that is, I only want one of the two and I'm indifferent

between them. Now we're no longer talking about coming up with a price between $50 and $400 - anything over $300 and I'll reject it and go to the other guy. Now we're talking about splitting the $250 profit between $50 and $300, and if we split it evenly I should expect to pay $175.

(why not $299? After all, the dealer knows $299 is better than my other offer. Because we're still playing the Ultimatum Game, that's why. And if it was $299, then having a second option - art that I like as much as the table - would actually make my bargaining position worse - after all, I was getting it for $225 before.)

Negotiation gurus call this backup option the BATNA ("Best Alternative To Negotiated Agreement") and consider it a useful thing to have. If only one participant in the negotiation has a BATNA greater than zero, that person is less desperate, needs the agreement less, and can hold out for a better deal - just as my $300 art allowed me to lower the asking price of the table from $225 to $175.

This "one buyer, one seller" model is artificial, but from here we can start to see how the real world existence of other buyers and sellers serve as BATNAs for both parties and how such negotiations eventually create the supply and demand of the marketplace.

The remaining case is one seller and multiple buyers (or vice versa). Here the seller's BATNA is "sell it to the other guy", and so a successful buyer must beat the other guy's price. In practice, this takes the form of an auction (why is this different than the previous example? Partly because in the previous example, we were comparing a negotiable commodity - the table - to a fixed price commodity - the art.)

How much should you bid at an auction? In the so-called English auction (the classic auction where a crazy man stands at the front shouting "Eighty!!! Eighty!!! We have eighty!!! Do I hear eighty-five?!? Eighty-five?!? Eighty-five to the man in the straw hat!!! Do I hear ninety?!?) the answer should be pretty obvious: keep bidding infinitesmally more than the last guy until you reach your value for the product, then stop. For example, with the $400 table, keep bidding until the price approaches $400.

But what about a sealed-bid auction, where everyone hands the auctioneer their bid and the auctioneer gives the product to the highest? Or what about the so-called "Dutch auction" where the auctioneer starts high and goes lower until someone bites ("A hundred?!? Anyone for a hundred?!? No?!? Ninety-five?!? Anyone for...yes?!? Sold for ninety-five to the man in the straw hat!!!).

The rookie mistake is to bid the amount you value the product. Remember, economists define "the amount you value the product" as "the price at which you would be indifferent between having the product and just keeping the money". If you go to an auction planning to bid your true value, you should expect to get absolutely zero benefit out of the experience. Instead, you should bid infinitesimally more than what you predict the next highest bidder will pay, as long as this is below your value.

Thus, the auction beloved by economists as perhaps the purest example of auction forms is the Vickrey, in which everyone submits a sealed bid, the highest bidder wins, and she pays the amount of the second-highest bid. This auction has a certain very elegant property, which is that here the dominant strategy is to bid your true value. Why?

Suppose you value a table at $400. If you try to game the system by bidding $350

instead of $400, you may lose out and can at best break even. Why? Because if the highest other bid was above $400, you wouldn't win the table in either case, and your ploy profits you nothing. And if the highest other bid was between $350 and $400 (let's say $375), now you lose the table and make $0 profit, as opposed to the $25 profit you would have made if you had bid your true value of $400, won, and paid the second-highest bid of $375. And if everyone else is below $350 (let's say $300) then you would have paid $300 in either case, and again your ploy profits you nothing. Bid above your true valuation (let's say $450) and you face similar consequences: either you wouldn't have gotten the table anyway, you get the table for the same amount as before, or you get the table for a value between $400 and $450 and now you're taking a loss.

In the real world, English, Dutch, sealed-bid and Vickrey auctions all differ a little in ways like how much information they give the bidders about each other, or whether people get caught up in the excitement of bidding, or what to do when you don't really know your true valuation. But in simplified rational models, they all end at an identical price: the true valuation of the second-highest bidder.

In conclusion, the gentlemanly way to bargain is to split the difference in profits between your and your partner's best alternative to an agreement, and gentlemanly auctions tend to end at the value of the second-highest participant. Some less gentlemanly alternatives are also available and will be discussed later.

Imperfect Voting Systems

Stalin once (supposedly) said that “He who casts the votes determines nothing; he who counts the votes determines everything “ But he was being insufficiently cynical. He who chooses the voting system may determine just as much as the other two players.

The Art of Strategy gives some good examples of this principle: here's an adaptation of one of them. Three managers are debating whether to give a Distinguished Employee Award to a certain worker. If the worker gets the award, she must receive one of two prizes: a $50 gift certificate, or a $10,000 bonus.

One manager loves the employee and wants her to get the $10,000; if she can't get the $10,000, she should at least get a gift certificate. A second manager acknowledges her contribution but is mostly driven by cost-cutting; she'd be happiest giving her the gift certificate, but would rather refuse to recognize her entirely than lose $10,000. And the third manager dislikes her and doesn't want to recognize her at all - but she also doesn't want the company to gain a reputation for stinginess, so if she gets recognized she'd rather give her the $10,000 than be so pathetic as to give her the cheap certificate.

The managers arrange a meeting to determine the employee's fate. If the agenda tells them to vote for or against giving her an award, and then proceed to determine the prize afterwards if she wins, then things will not go well for the employee. Why not? Because the managers reason as follows: if she gets the award, Manager 1 and Manager 3 will vote for the $10,000 prize, and Manager 2 will vote for the certificate. Therefore, voting for her to get the award is practically the same as voting for her to get the $10,000 prize. That means Manager 1, who wants her to get the prize, will vote yes on the award, but Managers 2 and 3, who both prefer no award to the $10,000, will strategically vote not to give her the award. Result: she doesn't get recognized for her distinguished service.

But suppose the employee involved happens to be the secretary arranging the meeting where the vote will take place. She makes a seemingly trivial change to the agenda: the managers will vote for what the prize should be first, and then vote on whether to give it to her.

If the managers decide the appropriate prize is $10,000, then the motion to give the award will fail for exactly the same reasons it did above. But if the managers decide the certificate is appropriate, then Manager 1 and 2, who both prefer the certificate to nothing, will vote in favor of giving the award. So the three managers, thinking strategically, realize that the decision before them, which looks like “$10 grand or certificate”, is really “No award or certificate”. Since 1 and 2 both prefer the certificate to nothing, they vote that the certificate is the appropriate prize (even though Manager 1 doesn't really believe this) and the employee ends out with the gift certificate.

But if the secretary is really smart, she may set the agenda as follows: The managers first vote whether or not to give $10,000, and if that fails, they next vote whether or not to give the certificate; if both votes fail the employee gets nothing. Here the managers realize that if the first vote (for $10,000) fails, the next vote (certificate or nothing) will pass, since two managers prefer certificate to nothing as mentioned

before. So the true choice in the first vote is “$10,000 versus certificate”. Since two managers (1 and 3) prefer the $10,000 to the certificate, those two start by voting to give the full $10,000, and this is what the employee gets.

So we see that all three options are possible outcomes, and that the true power rests not in the hands of any individual manager, but in the secretary who determines how the voting takes place.

Americans have a head start in understanding the pitfalls of voting systems thanks to the so-called two party system. Every four years, they face quandaries like "If leftists like me vote for Nader instead of Gore just because we like him better, are we going to end up electing Bush because we've split the leftist vote?"

Empirically, yes. The 60,000 Florida citizens who voted Green in 2000 didn't elect Nader. However, they did make Gore lose to Bush by a mere 500 votes. The last post discussed a Vickrey auction, a style of auction in which you have have no incentive to bid anything except your true value. Wouldn't it be nice if we had an electoral system with the same property: one where you should always vote for the candidate you actually support? If such a system existed, we would have ample reason to institute it and could rest assured that no modern-day Stalin was manipulating us via the choice of voting system we used.

Some countries do claim to have better systems than the simple winner-takes-all approach of the United States. My own adopted homeland of Ireland uses a system called “single transferable vote” (also called instant-runoff vote), in which voters rank the X candidates from 1 to X. If a candidate has the majority of first preference votes (or a number of first preference votes greater than the number of positions to fill divided by the number of candidates, in elections with multiple potential winners like legislative elections), then that candidate wins and any surplus votes go to their voters' next preference. If no one meets the quota, then the least popular candidate is eliminated and their second preference votes become first preferences. The system continues until all available seats are full.

For example, suppose I voted (1: Nader), (2: Gore), (3: Bush). The election officials tally all the votes and find that Gore has 49 million first preferences, Bush has 50 million, and Nader has 5 million. There's only one presidency, so a candidate would have to have a majority of votes (greater than 52 million out of 104 million) to win. Since no one meets that quota, the lowest ranked candidate gets eliminated - in this case, Nader. My vote now goes to my second preference, Gore. If 4 million Nader voters put Gore second versus 1 million who put Bush second, the tally's now at 53 million Gore, 51 million Bush. Gore has greater than 52 million and wins the election - the opposite result from if we'd elected a president the traditional way.

Another system called Condorcet voting also uses a list of all candidates ranked in order, but uses the information to run mock runoffs between each of them. So a Condorcet system would use the ballots to run a Gore/Nader match (which Gore would win), a Gore/Bush match (which Gore would win), and a Bush/Nader match (which Bush would win). Since Gore won all of his matches, he becomes President. This becomes complicated when no candidate wins all of his matches (imagine Gore beating Nader, Bush beating Gore, but Nader beating Bush in a sort of Presidential rock-paper-scissors.) Condorcet voting has various options to resolve this; some systems give victory to the candidate whose greatest loss was by the smallest margin, and others to candidates who defeated the greatest number of other candidates.

Do these systems avoid the strategic voting that plagues American elections? No. For example, both Single Transferable Vote and Condorcet voting sometimes provide incentives to rank a candidate with a greater chance of winning higher than a candidate you prefer - that is, the same "vote Gore instead of Nader" dilemma you get in traditional first-past-the-post.

There are many other electoral systems in use around the world, including several more with ranking of candidates, a few that do different sorts of runoffs, and even some that ask you to give a numerical rating to each candidate (for example “Nader 10, Gore 6, Bush -100000”). Some of them even manage to eliminate the temptation to rank a non-preferred candidate first. But these work only at the expense of incentivizing other strategic manuevers, like defining “approved candidate” differently or exaggerating the difference between two candidates.

So is there any voting system that automatically reflects the will of the populace in every way without encouraging tactical voting? No. Various proofs, including the Gibbard-Satterthwaite Theorem and the better-known Arrow Impossibility Theorem show that many of the criteria by which we would naturally judge voting systems are mutually incompatible and that all reasonable systems must contain at least some small element of tactics (one example of an unreasonable system that eliminates tactical voting is picking one ballot at random and determining the results based solely on its preferences; the precise text of the theorem rules out “nondeterministic or dictatorial” methods).

This means that each voting system has its own benefits and drawbacks, and that which one people use is largely a matter of preference. Some of these preferences reflect genuine concern about the differences between voting systems: for example, is it better to make sure your system always elects the Condorcet winner, even if that means the system penalizes candidates who are too similar to other candidates? Is it better to have a system where you can guarantee that participating in the election always makes your candidate more likely to win, or one where you can be sure that everyone voting exactly the opposite will never elect the same candidate?

But in practice, these preferences tend to be political and self-interested. This was recently apparent in Britain, which voted last year on a referendum to change the voting system. The Liberal Democrats, who were perpetually stuck in the same third-place situation as Nader in the States, supported a change to a form of instant runoff voting which would have made voting Lib Dem a much more palatable option; the two major parties opposed it probably for exactly that reason.

Although no single voting system is mathematically perfect, several do seem to do better on the criteria that real people care about; look over Wikipedia's section on the strengths and weaknesses of different voting systems to see which one looks best.

Game Theory As A Dark Art

One of the most charming features of game theory is the almost limitless depths of evil to which it can sink.

Your garden-variety evils act against your values. Your better class of evil, like Voldemort and the folk-tale version of Satan, use your greed to trick *you* into acting against *your own* values, then grab away the promised reward at the last moment. But even demons and dark wizards can only do this once or twice before most victims wise up and decide that taking their advice is a bad idea. Game theory can force you to betray your deepest principles for no lasting benefit again and again, and still leave you convinced that your behavior was rational.

Some of the examples in this post probably wouldn't work in reality; they're more of a *reductio ad absurdum* of the so-called *homo economicus* who acts free from any feelings of altruism or trust. But others are lifted directly from real life where seemingly intelligent people genuinely fall for them. And even the ones that don't work with real people might be valuable in modeling institutions or governments.

Of the following examples, the first three are from *The Art of Strategy*; the second three are relatively classic problems taken from around the Internet. A few have been mentioned in the comments here already and are reposted for people who didn't catch them the first time.

The Evil Plutocrat

You are an evil plutocrat who wants to get your pet bill - let's say a law that makes evil plutocrats tax-exempt - through the US Congress. Your usual strategy would be to bribe the Congressmen involved, but that would be pretty costly - Congressmen no longer come cheap. Assume all Congressmen act in their own financial self-interest, but that absent any financial self-interest they will grudgingly default to honestly representing their constituents, who hate your bill (and you personally). Is there any way to ensure Congress passes your bill, without spending any money on bribes at all?

Yes. Simply tell all Congressmen that *if* your bill fails, you will donate some stupendous amount of money to whichever party gave the greatest percent of their votes in favor.

Suppose the Democrats try to coordinate among themselves. They say “If we all oppose the bill, then if even one Republican supports the bill, the Republicans will get lots of money they can spend on campaigning against us. If only one of us supports the bill, the Republicans may anticipate this strategy and two of them may support it. The only way to ensure the Republicans don't gain a massive windfall and wipe the floor with us next election is for most of us to vote for the bill.”

Meanwhile, in their meeting, the Republicans think the same thing. The vote ends with most members of Congress supporting your bill, and you don't end up having to pay any money at all.

The Hostile Takeover

You are a ruthless businessman who wants to take over a competitor. The competitor's stock costs $100 a share, and there are 1000 shares, distributed among a hundred investors who each own ten. That means the company ought to cost $100,000, but you don't have $100,000. You only have $98,000. Worse, another competitor with $101,000 has made an offer for greater than the value of the company: they will pay $101 per share if they end up getting all of the shares. Can you still manage to take over the company?

Yes. You can make what is called a two-tiered offer. Suppose all investors get a chance to sell shares simultaneously. You will pay $105 for 500 shares - better than they could get from your competitor - but only pay $90 for the other 500. If you get fewer than 500 shares, all will sell for $105; if you get more than 500, you will start by distributing the $105 shares evenly among all investors who sold to you, and then distribute out as many of the $90 shares as necessary (leaving some $90 shares behind except when all investors sell to you) . And you will do this whether or not you succeed in taking over the company - if only one person sells you her share, then that one person gets $105.

Suppose an investor believes you're not going to succeed in taking over the company. That means you're not going to get over 50% of shares. That means the offer to buy 500 shares for $105 will still be open. That means the investor can either sell her share to you (for $105) or to your competitor (for $101). Clearly, it's in this investor's self-interest to sell to you.

Suppose the investor believes you will succeed in taking over the company. That means your competitor will not take over the company, and its $101 offer will not apply. That means that the new value of the shares will be $90, the offer you've made for the second half of shares. So they will get $90 if they don't sell to you. How much will they get if they do sell to you? They can expect half of their ten shares to go for $105 and half to go for $90; they will get a total of $97.50 per share. $97.50 is better than $90, so their incentive is to sell to you.

Suppose the investor believes you are right on the cusp of taking over the company, and her decision will determine the outcome. In that case, you have at most 499 shares. When the investor gives you her 10 shares, you will end up with 509 - 500 of which are $105 shares and 9 of which are $90 shares. If these are distributed randomly, investors can expect to make on average $104.73 per share, compared to $101 if your competitor buys the company.

Since all investors are thinking along these lines, they all choose to buy shares from you instead of your competitor. You pay out an average of $97.50 per share, and take over the company for $97,500, leaving $500 to spend on the victory party.

The stockholders, meanwhile, are left wondering why they just all sold shares for $97.50 when there was someone else who was promising them $101.

The Hostile Takeover, Part II

Your next target is a small family-owned corporation that has instituted what they consider to be invincible protection against hostile takeovers. All decisions are made by the Board of Directors, who serve for life. Although shareholders vote in the new members of the Board after one of them dies or retires, Board members can hang on

for decades. And all decisions about the Board, impeachment of its members, and enforcement of its bylaws are made by the Board itself, with members voting from newest to most senior.

So you go about buying up 51% of the stock in the company, and sure enough, a Board member retires and is replaced by one of your lackeys. This lackey can propose procedural changes to the Board, but they have to be approved by majority vote. And at the moment the other four directors hate you with a vengeance, and anything you propose is likely to be defeated 4-1. You need those other four windbags out of there, and soon, but they're all young and healthy and unlikely to retire of their own accord.

The obvious next step is to start looking for a good assassin. But if you can't find one, is there any way you can propose mass forced retirement to the Board and get them to approve it by majority vote? Even better, is there any way you can get them to approve it unanimously, as a big “f#@& you” to whoever made up this stupid system?

Yes. Your lackey proposes as follows: “I move that we vote upon the following: that if this motion passes unanimously, all members of the of the Board resign immediately and are given a reasonable compensation; that if this motion passes 4-1 that the Director who voted against it must retire without compensation, and the four directors who voted in favor may stay on the Board; and that if the motion passes 3-2, then the two 'no' voters get no compensation and the three 'yes' voters may remain on the board and will also get a spectacular prize - to wit, our company's 51% share in your company divided up evenly among them.”

Your lackey then votes “yes”. The second newest director uses backward reasoning as follows:

Suppose that the vote were tied 2-2. The most senior director would prefer to vote “yes”, because then she gets to stay on the Board and gets a bunch of free stocks.

But knowing that, the second most senior director (SMSD) will also vote 'yes'. After all, when the issue reaches the SMSD, there will be one of the following cases:

1. If there is only one yes vote (your lackey's), the SMSD stands to gain from voting yes, knowing that will produce a 2-2 tie and make the most senior director vote yes to get her spectacular compensation. This means the motion will pass 3-2, and the SMSD will also remain on the board and get spectacular compensation if she votes yes, compared to a best case scenario of remaining on the board if she votes no.

2. If there are two yes votes, the SMSD must vote yes - otherwise, it will go 2-2 to the most senior director, who will vote yes, the motion will pass 3-2, and the SMSD will be forced to retire without compensation.

3. And if there are three yes votes, then the motion has already passed, and in all cases where the second most senior director votes “no”, she is forced to retire without compensation. Therefore, the second most senior director will always vote “yes”.

Since your lackey, the most senior director, and the second most senior director will always vote "yes", we can see that the other two directors, knowing the motion will pass, must vote "yes" as well in order to get any compensation at all. Therefore, the motion passes unanimously and you take over the company at minimal cost.

The Dollar Auction

You are an economics professor who forgot to go to the ATM before leaving for work, and who has only $20 in your pocket. You have a lunch meeting at a very expensive French restaurant, but you're stuck teaching classes until lunchtime and have no way to get money. Can you trick your students into giving you enough money for lunch in exchange for your $20, without lying to them in any way?

Yes. You can use what's called an all-pay auction, in which several people bid for an item, as in a traditional auction, but everyone pays their bid regardless of whether they win or lose (in a common variant, only the top two bidders pay their bids).

Suppose one student, Alice, bids $1. This seems reasonable - paying $1 to win $20 is a pretty good deal. A second student, Bob, bids $2. Still a good deal if you can get a twenty for a tenth that amount.

The bidding keeps going higher, spurred on by the knowledge that getting a $20 for a bid of less than $20 would be pretty cool. At some point, maybe Alice has bid $18 and Bob has bid $19.

Alice thinks: “What if I raise my bid to $20? Then certainly I would win, since Bob would not pay more than $20 to get $20, but I would only break even. However, breaking even is better than what I'm doing now, since if I stay where I am Bob wins the auction and I pay $18 without getting anything.” Therefore Alice bids $20.

Bob thinks “Well, it sounds pretty silly to bid $21 for a twenty dollar bill. But if I do that and win, I only lose a dollar, as opposed to bowing out now and losing my $19 bid.” So Bob bids $21.

Alice thinks “If I give up now, I'll lose a whole dollar. I know it seems stupid to keep going, but surely Bob has the same intuition and he'll give up soon. So I'll bid $22 and just lose two dollars...”

It's easy to see that the bidding could in theory go up with no limits but the players' funds, but in practice it rarely goes above $200.

...yes, $200. Economist Max Bazerman claims that of about 180 such auctions, seven have made him more than $100 (ie $50 from both players) and his highest take was $407 (ie over $200 from both players).

In any case, you're probably set for lunch. If you're not, take another $20 from your earnings and try again until you are - the auction gains even more money from people who have seen it before than it does from naive bidders (!) Bazerman, for his part, says he's made a total of $17,000 from the exercise.

At that point you're starting to wonder why no one has tried to build a corporation around this, and unsurprisingly, the online auction site Swoopo appears to be exactly that. More surprisingly, they seem to have gone bankrupt last year, suggesting that maybe H.L. Mencken was wrong and someone *has* gone broke underestimating people's intelligence.

The Bloodthirsty Pirates

You are a pirate captain who has just stolen $17,000, denominated entirely in $20

bills, from a very smug-looking game theorist. By the Pirate Code, you as the captain may choose how the treasure gets distributed among your men. But your first mate, second mate, third mate, and fourth mate all want a share of the treasure, and demand on threat of mutiny the right to approve or reject any distribution you choose.You expect they'll reject anything too lopsided in your favor, which is too bad, because that was totally what you were planning on.

You remember one fact that might help you - your crew, being bloodthirsty pirates, all hate each other and actively want one another dead. Unfortunately, their greed seems to have overcome their bloodlust for the moment, and as long as there are advantages to coordinating with one another, you won't be able to turn them against their fellow sailors. Doubly unfortunately, they also actively want you dead.

You think quick. “Aye,” you tell your men with a scowl that could turn blood to ice, “ye can have yer votin' system, ye scurvy dogs” (you're that kind of pirate). “But here's the rules: I propose a distribution. Then you all vote on whether or not to take it. If a majority of you, or even half of you, vote 'yes', then that's how we distribute the treasure. But if you vote 'no', then I walk the plank to punish me for my presumption, and the first mate is the new captain. He proposes a new distribution, and again you vote on it, and if you accept then that's final, and if you reject it he walks the plank and the second mate becomes the new captain. And so on.”

Your four mates agree to this proposal. What distribution should you propose? Will it be enough to ensure your comfortable retirement in Jamaica full of rum and wenches?

Yes. Surprisingly, you can get away with proposing that you get $16,960, your first mate gets nothing, your second mate gets $20, your third mate gets nothing, and your fourth mate gets $20 - and you will still win 3 -2.

The fourth mate uses backward reasoning like so: Suppose there were only two pirates left, me and the third mate. The third mate wouldn't have to promise me anything, because if he proposed all $17,000 for himself and none for me, the vote would be 1-1 and according to the original rules a tie passes. Therefore this is a better deal than I would get if it were just me and the third mate.

But suppose there were three pirates left, me, the third mate, and the second mate. Then the second mate would be the new captain, and he could propose $16,980 for himself, $0 for the third mate, and $20 for me. If I vote no, then it reduces to the previous case in which I get nothing. Therefore, I should vote yes and get $20. Therefore, the final vote is 2-1 in favor.

But suppose there were four pirates left: me, the third mate, the second mate, and the first mate. Then the first mate would be the new captain, and he could propose $16,980 for himself, $20 for the third mate, $0 for the second mate, and $0 for me. The third mate knows that if he votes no, this reduces to the previous case, in which he gets nothing. Therefore, he should vote yes and get $20. Therefore, the final vote is 2-2, and ties pass.

(He might also propose $16980 for himself, $0 for the second mate, $0 for the third mate, and $20 for me. But since he knows I am a bloodthirsty pirate who all else being equal wants him dead, I would vote no since I could get a similar deal from the third mate and make the first mate walk the plank in the bargain. Therefore, he would offer the $20 to the third mate.)

But in fact there are five pirates left: me, the third mate, the second mate, the first mate, and the captain. The captain has proposed $16,960 for himself, $20 for the second mate, and $20 for me. If I vote no, this reduces to the previous case, in which I get nothing. Therefore, I should vote yes and get $20.

(The captain would avoid giving the $20s to the third and fourth rather than to the second and fourth mates for a similar reason to the one given in the previous example - all else being equal, the pirates would prefer to watch him die.)

The second mate thinks along the same lines and realizes that if he votes no, this reduces to the case with the first mate, in which the second mate also gets nothing. Therefore, he too votes yes.

Since you, as the captain, obviously vote yes as well, the distribution passes 3-2. You end up with $16,980, and your crew, who were so certain of their ability to threaten you into sharing the treasure, each end up with either a single $20 or nothing.

The Prisoners' Dilemma, Redux

This sequence previously mentioned the popularity of Prisoners' Dilemmas as gimmicks on TV game shows. In one program, Golden Balls, contestants do various tasks that add money to a central “pot”. By the end of the game, only two contestants are left, and are offered a Prisoners' Dilemma situation to split the pot between them. If both players choose to “Split”, the pot is divided 50-50. If one player “Splits” and the other player “Steals”, the stealer gets the entire pot. If both players choose to “Steal”, then no one gets anything. The two players are allowed to talk to each other before making a decision, but like all Prisoner's Dilemmas, the final choice is made simultaneously and in secret.

You are a contestant on this show. You are actually not all that evil - you would prefer to split the pot rather than to steal all of it for yourself - but you certainly don't want to trust the other guy to have the same preference. In fact, the other guy looks a bit greedy. You would prefer to be able to rely on the other guy's rational self-interest rather than on his altruism. Is there any tactic you can use before the choice, when you're allowed to communicate freely, in order to make it rational for him to cooperate?

Yes. In one episode of Golden Balls, a player named Nick successfully meta-games the game by transforming it from the Prisoner's Dilemma (where defection is rational) to the Ultimatum Game (where cooperation is rational)

Nick tells his opponent: “I am going to choose 'Steal' on this round.” (He then immediately pressed his button; although the show hid which button he pressed, he only needed to demonstrate that he had committed and his mind could no longer be changed) “If you also choose 'Steal', then for certain neither of us gets any money. If you choose 'Split', then I get all the money, but immediately after the game, I will give you half of it. You may not trust me on this, and that's understandable, but think it through. First, there's no less reason to think I'm trustworthy than if I had just told you I pressed 'Split' to begin with, the way everyone else on this show does. And second, now if there's any chance whatsoever that I'm trustworthy, then that's some chance of getting the money - as opposed to the zero chance you have of getting the money if you choose 'Steal'.”

Nick's evaluation is correct. His opponent can either press 'Steal', with a certainty of

getting zero, or press 'Split', with a nonzero probability of getting his half of the pot depending on Nick's trustworthiness.

But this solution is not quite perfect, in that one can imagine Nick's opponent being very convinced that Nick will cheat him, and deciding he values punishing this defection more than the tiny chance that Nick will play fair. That's why I was so impressed to see cousin_it propose what I think is an even better solution on the Less Wrong thread on the matter:

> This game has multiple Nash equilibria and cheap talk is allowed, so correlated equilibria are possible. Here's how you implement a correlated equilibrium if your opponent is smart enough:
>
> "We have two minutes to talk, right? I'm going to ask you to flip a coin (visibly to both of us) at the last possible moment, the exact second where we must cease talking. If the coin comes up heads, I promise I'll cooperate, you can just go ahead and claim the whole prize. If the coin comes up tails, I promise I'll defect. Please cooperate in this case, because you have nothing to gain by defecting, and anyway the arrangement is fair, isn't it?"

This sort of clever thinking is, in my opinion, the best that game theory has to offer. It shows that game theory need not be only a tool of evil for classical figures of villainy like bloodthirsty pirate captains or corporate raiders or economists, but can also be used to create trust and ensure cooperation between parties with common interests.