Best of LessWrong: July 2012

1. Should you try to do good work on LW?
2. An Intuitive Explanation of Solomonoff Induction
3. Negative and Positive Selection
4. Neuroscience basics for LessWrongians
5. What Is Signaling, Really?
6. Game Theory As A Dark Art
7. The Mere Cable Channel Addition Paradox
8. Reply to Holden on The Singularity Institute
9. Interlude for Behavioral Economics
10. A Marriage Ceremony for Aspiring Rationalists
11. [Retracted] Simpson's paradox strikes again: there is no great stagnation?
12. Real World Solutions to Prisoners' Dilemmas
13. Magic players: "How do I lose?"
14. AI cooperation is already studied in academia as "program equilibrium"
15. Kurzweil's predictions: good accuracy, poor self-calibration
16. Notes on the Psychology of Power
17. Where to Intervene in a Human?
18. Imperfect Voting Systems
19. Article about LW: Faith, Hope, and Singularity: Entering the Matrix with New York's Futurist Set
20. Exploiting the Typical Mind Fallacy for more accurate questioning?
21. Revisiting SI's 2011 strategic plan: How are we doing?
22. CFAR website launched

Best of LessWrong: July 2012

1. Should you try to do good work on LW?
2. An Intuitive Explanation of Solomonoff Induction
3. Negative and Positive Selection
4. Neuroscience basics for LessWrongians
5. What Is Signaling, Really?
6. Game Theory As A Dark Art
7. The Mere Cable Channel Addition Paradox
8. Reply to Holden on The Singularity Institute
9. Interlude for Behavioral Economics
10. A Marriage Ceremony for Aspiring Rationalists
11. [Retracted] Simpson's paradox strikes again: there is no great stagnation?
12. Real World Solutions to Prisoners' Dilemmas
13. Magic players: "How do I lose?"
14. AI cooperation is already studied in academia as "program equilibrium"
15. Kurzweil's predictions: good accuracy, poor self-calibration
16. Notes on the Psychology of Power
17. Where to Intervene in a Human?
18. Imperfect Voting Systems
19. Article about LW: Faith, Hope, and Singularity: Entering the Matrix with New York’s Futurist Set
20. Exploiting the Typical Mind Fallacy for more accurate questioning?
21. Revisiting SI's 2011 strategic plan: How are we doing?
22. CFAR website launched

Should you try to do good work on LW?

I used to advocate trying to do good work on LW. Now I'm not sure, let me explain why.

It's certainly true that good work stays valuable no matter where you're doing it. Unfortunately, the standards of "good work" are largely defined by where you're doing it. If you're in academia, your work is good or bad by scientific standards. If you're on LW, your work is good or bad compared to other LW posts. Internalizing that standard may harm you if you're capable of more.

When you come to a place like Project Euler and solve some problems, or come to OpenStreetMap and upload some GPS tracks, or come to academia and publish a paper, that makes you a participant and you know exactly where you stand, relative to others. But LW is not a task-focused community and is unlikely to ever become one. LW evolved from the basic activity "let's comment on something Eliezer wrote". We inherited our standard of quality from that. As a result, when someone posts their work here, that doesn't necessarily help them improve.

For example, Yvain is a great contributor to LW and has the potential to be a star writer, but it seems to me that writing on LW doesn't test his limits, compared to trying new audiences. Likewise, my own work on decision theory math would've been held to a higher standard if the primary audience were mathematicians (though I hope to remedy that). Of course there have been many examples of seemingly good work posted to LW. Homestuck fandom also has a lot of nice-looking art, but it doesn't get fandoms of its own.

In conclusion, if you want to do important work, cross-post it if you must, but don't do it for LW exclusively. Big fish in a small pond always looks kinda sad.

An Intuitive Explanation of Solomonoff Induction

This is the completed article that Luke wrote the first half of. My thanks go to the following for reading, editing, and commenting; Luke Muehlhauser, Louie Helm, Benjamin Noble, and Francelle Wax.

People disagree about things. Some say that television makes you dumber; other say it makes you smarter. Some scientists believe life must exist elsewhere in the universe; others believe it must not. Some say that complicated financial derivatives are essential to a modern competitive economy; others think a nation's economy will do better without them. It's hard to know what is true.

And it's hard to know *how to figure out* what is true. Some argue that you should assume the things you are most certain about and then deduce all other beliefs from your original beliefs. Others think you should accept at face value the most intuitive explanations of personal experience. Still others think you should generally agree with the scientific consensus until it is disproved.

Wouldn't it be nice if determining what is true was like baking a cake? What if there was a *recipe* for finding out what is true? All you'd have to do is follow the written directions exactly, and after the last instruction you'd inevitably find yourself with some sweet, tasty *truth*!

In this tutorial, we'll explain the closest thing we've found so far to a recipe for finding truth: Solomonoff induction.

There are some qualifications to make. To describe just one: roughly speaking, you don't have time to follow the recipe. To find the truth to even a simple question using this recipe would require you to follow one step after another until long after the heat death of the universe, and you can't do that.

But we can find shortcuts. Suppose you know that the *exact* recipe for baking a cake asks you to count out one molecule of H_2O at a time until you have *exactly* 0.5 cups of

water. If you did that, you might not finish the cake before the heat death of the universe. But you could approximate that part of the recipe by measuring out something very close to 0.5 cups of water, and you'd probably still end up with a pretty good cake.

Similarly, once we know the exact recipe for finding truth, we can try to approximate it in a way that allows us to finish all the steps sometime before the sun burns out.

This tutorial explains that best-we've-got-so-far recipe for finding truth, Solomonoff induction. Don't worry, we won't be using any equations, just qualitative descriptions.

Like Eliezer Yudkowsky's Intuitive Explanation of Bayes' Theorem and Luke Muehlhauser's Crash Course in the Neuroscience of Human Motivation, this tutorial is *long*. You may not have time to read it; that's fine. But if you do read it, we recommend that you read it in sections.

Contents:

Algorithms

At an early age you learned a set of precisely-defined steps — a 'recipe' or, more formally, an *algorithm* — that you could use to find the largest number in a list of numbers like this:

21, 18, 4, 19, 55, 12, 30

The algorithm you learned probably looked something like this:

1. Look at the first item. Note that it is the largest you've seen on this list so far. If this is the only item on the list, output it as the largest number on the list. Otherwise, proceed to step 2.
2. Look at the next item. If it is larger than the largest item noted so far, note it as the largest you've seen in this list so far. Proceed to step 3.
3. If you have not reached the end of the list, return to step 2. Otherwise, output the last noted item as the largest number in the list.

Other algorithms could be used to solve the same problem. For example, you could work your way from right to left instead of from left to right. But the point is that if you follow this algorithm exactly, and you have enough time to complete the task, you can't *fail* to solve the problem. You can't get confused about what one of the steps means or what the next step is. Every instruction tells you exactly what to do next, all the way through to the answer.

You probably learned other algorithms, too, like how to find the greatest common divisor of any two integers (see image on right).

But not just any set of instructions is a precisely-defined algorithm. Sometimes, instructions are unclear or incomplete. Consider the following instructions based on an article about the scientific method:

1. Make an observation.
2. Form a hypothesis that explains the observation.
3. Conduct an experiment that will test the hypothesis.
4. If the experimental results disconfirm the hypothesis, return to step #2 and form a hypothesis not yet used. If the experimental results confirm the hypothesis, provisionally accept the hypothesis.

This is not an algorithm.

First, many of the terms are not clearly defined. What counts as an observation? What counts as a hypothesis? What would a hypothesis need to be like in order to 'explain' the observation? What counts as an experiment that will 'test' the hypothesis? What does it mean for experimental results to 'confirm' or 'disconfirm' a hypothesis?

Second, the instructions may be incomplete. What do we do if we reach step 4 and the experimental results neither 'confirm' nor 'disconfirm' the hypothesis under consideration, but instead are in some sense 'neutral' toward the hypothesis? These instructions don't tell us what to do in that case.

An algorithm is a well-defined procedure that takes some value or values as input and, after a finite series of steps, generates some value or values as output.

For example, the 'find the largest number' algorithm above could take the input {21, 18, 4, 19, 55, 12, 30} and would, after 13 steps, produce the following output: {55}. Or it could take the input {34} and, after 1 step, produce the output: {34}.

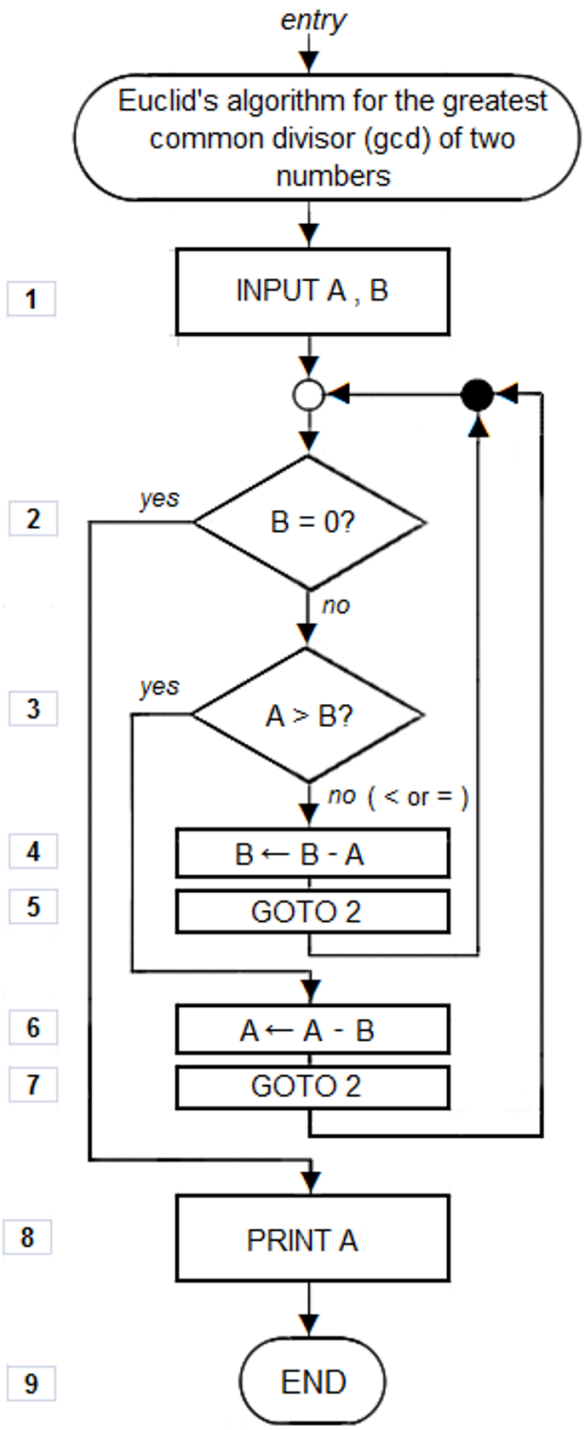
An algorithm is so well written, that we can construct machines that follow them. Today, the machines that follow algorithms are mostly computers. This is why all computer science students take a class in algorithms. If we construct our algorithm for truth, then we can make a computer program that finds truth—an Artificial Intelligence.

Induction

Let’s clarify what we mean. In movies, scientists will reveal “truth machines”. Input a statement, and the truth machine will tell you whether it is true or false. This is *not* what Solomonoff induction does. Instead, Solomonoff induction is our ultimate “induction machine”.

> Whether we are a detective trying to catch a thief, a scientist trying to discover a new physical law, or a businessman attempting to understand a recent change in demand, we are all in the process of collecting information and trying to infer the underlying causes.
>
> -Shane Legg

The problem of induction is this: We have a set of *observations* (or *data*), and we want to find the underlying causes of those observations. That is, we want to find *hypotheses* that explain our data. We’d like to know which hypothesis is correct, so we can use that knowledge to predict future events. Our algorithm for truth will not listen to questions and answer yes or no. Our algorithm will take in data (observations) and output the rule by which the data was created. That is, it will give us the explanation of the observations; the causes.



Suppose your data concern a large set of stock market changes and other events in the world. You’d like to know the processes responsible for the stock market price changes, because then you can predict what the stock market will do in the future, and make some money.

Or, suppose you are a parent. You come home from work to find a chair propped against the refrigerator, with the cookie jar atop the fridge a bit

emptier than before. You like cookies, and you don't want them to disappear, so you start thinking. One hypothesis that leaps to mind is that your young daughter used the chair to reach the cookies. However, many other hypotheses explain the data. Perhaps a very short thief broke into your home and stole some cookies. Perhaps your daughter put the chair in front of the fridge because the fridge door is broken and no longer stays shut, and you forgot that your friend ate a few cookies when he visited last night. Perhaps you moved the chair and ate the cookies yourself while sleepwalking the night before.

All these hypotheses are possible, but intuitively it seems like some hypotheses are more likely than others. If you've seen your daughter access the cookies this way before but have never been burgled, then the 'daughter hypothesis' seems more plausible. If some expensive things from your bedroom and living room are missing and there is hateful graffiti on your door at the eye level of a very short person, then the 'short thief' hypothesis becomes more plausible than before. If you suddenly remember that your friend ate a few cookies and broke the fridge door last night, the 'broken fridge door' hypothesis gains credibility. If you've never been burgled and your daughter is out of town and you have a habit of moving and eating things while sleepwalking, the 'sleepwalking' hypothesis becomes less bizarre.

So the weight you give to each hypothesis depends greatly on your prior knowledge. But what if you had just been hit on the head and lost all past memories, and for some reason the most urgent thing you wanted to do was to solve the mystery of the chair and cookies? Then how would you weigh the likelihood of the available hypotheses?

When you have very little data but want to compare hypotheses anyway, Occam's Razor comes to the rescue.

Occam's Razor

Consider a different inductive problem. A computer program outputs the following sequence of numbers:

1, 3, 5, 7

Which number comes next? If you guess correctly, you'll win $500.

In order to predict the next number in the sequence, you make a hypothesis about the process the computer is using to generate these numbers. One obvious hypothesis is that it is simply listing all the odd numbers in ascending order from 1. If that's true, you should guess that "9" will be the next number.

But perhaps the computer is using a different algorithm to generate the numbers. Suppose that n is the step in the sequence, so that n=1 when it generated '1', n=2 when it generated '3', and so on. Maybe the computer used this equation to calculate each number in the sequence:

$$2n - 1 + (n - 1)(n - 2)(n - 3)(n - 4)$$

If so, the next number in the sequence will be 33. (Go ahead, check the calculations.)

But doesn't the first hypothesis seem more likely?

The principle behind this intuition, which goes back to William of Occam, could be stated:

> Among all hypotheses consistent with the observations, the simplest is the most likely.

The principle is called Occam's razor because it 'shaves away' unnecessary assumptions.

For example, think about the case of the missing cookies again. In most cases, the 'daughter' hypothesis seems to make fewer unnecessary assumptions than the 'short thief' hypothesis does. You already know you have a daughter that likes cookies and knows how to move chairs to reach cookies. But in order for the short thief hypothesis to be plausible, you have to assume that (1) a thief found a way to break in, that (2) the thief wanted inexpensive cookies from your home, that (3) the thief was, unusually, too short to reach the top of the fridge without the help of a chair, and (4) many other unnecessary assumptions.

Occam's razor sounds right, but can it be made more precise, and can it be justified? How do we find *all* consistent hypotheses, and how do we judge their simplicity? We will return to those questions later. Before then, we'll describe the area of mathematics that usually deals with reasoning: probability.

Probability

You're a soldier in combat, crouching in a trench. You know for sure there is just one enemy soldier left on the battlefield, about 400 yards away. You also know that if the remaining enemy is a regular army troop, there's only a small chance he could hit you with one shot from that distance. But if the remaining enemy is a sniper, then there's a very good chance he can hit you with one shot from that distance. But snipers are rare, so it's probably just a regular army troop.

You peek your head out of the trench, trying to get a better look.

Bam! A bullet glances off your helmet and you duck down again.

"Okay," you think. "I know snipers are rare, but that guy just hit me with a bullet from 400 yards away. I suppose it might still be a regular army troop, but there's a seriously good chance it's a sniper, since he hit me from that far away."

After another minute, you dare to take another look, and peek your head out of the trench again.

Bam! Another bullet glances off your helmet! You duck down again.

"Whoa," you think. "It's definitely a sniper. No matter how rare snipers are, there's no way that guy just hit me twice in a row from that distance if he's a regular army troop. He's gotta be a sniper. I'd better call for support."

This is an example of reasoning under uncertainty, of updating uncertain beliefs in response to evidence. We do it all the time.

You start with some prior beliefs, and all of them are uncertain. You are 99.99% certain the Earth revolves around the sun, 90% confident your best friend will attend your

birthday party, and 40% sure that the song you're listening to on the radio was played by The Turtles.

Then, you encounter new evidence—new observations—and you update your beliefs in response.

Suppose you start out 85% confident that the one remaining enemy soldier is not a sniper. That leaves only 15% credence to the hypothesis that he *is* a sniper. But then, a bullet glances off your helmet — an event far more likely if the enemy soldier is a sniper than if he is not. So now you're only 40% confident he's a non-sniper, and 60% confident he is a sniper. Another bullet glances off your helmet, and you update again. Now you're only 2% confident he's a non-sniper, and 98% confident he is a sniper.

Probability theory is the mathematics of reasoning with uncertainty. The keystone of this subject is called Bayes' Theorem. It tells you how likely something is given some other knowledge. Understanding this simple theorem is more useful and important for most people than Solomonoff induction. If you haven't learned it already, you may want to read either tutorial #1, tutorial #2, tutorial #3, or tutorial #4 on Bayes' Theorem. The exact math of Bayes' Theorem is not required for this tutorial. We'll just describe its results qualitatively.

Bayes' Theorem can tell us how likely a hypothesis is, given evidence (or data, or observations). This is helpful because we want to know which model of the world is correct so that we can successfully predict the future. It calculates this probability based on the prior probability of the hypothesis alone, the probability of the evidence alone, and the probability of the evidence *given* the hypothesis. Now we just plug the numbers in.

Of course, it's not easy to "just plug the numbers in." You aren't an all-knowing god. You don't know *exactly* how likely it is that the enemy soldier would hit your helmet if he's a sniper, compared to how likely that is if he's not a sniper. But you can do your best. With enough evidence, it will become overwhelmingly clear which hypothesis is correct.

But guesses are not well-suited to an exact algorithm, and so our quest to find an algorithm for truth-finding must continue. For now, we turn to the problem of choosing priors.

The Problem of Priors

In the example above where you're a soldier in combat, I gave you your starting probabilities: 85% confidence that the enemy soldier was a sniper, and 15% confidence he was not. But what if you don't know your "priors"? What then?

Most situations in real life are complex, so that your "priors" (as used in Bayes' Theorem) are actually probabilities that have been updated several times with past evidence. You had an idea that snipers were rare because you saw many soldiers, but only a few of them were snipers. Or you read a reliable report saying that snipers were rare. But what would our ideal reasoning computer do before it knew anything? What would the probabilities be set to before we turned it on? How can we determine the probability of a hypothesis before seeing *any* data?

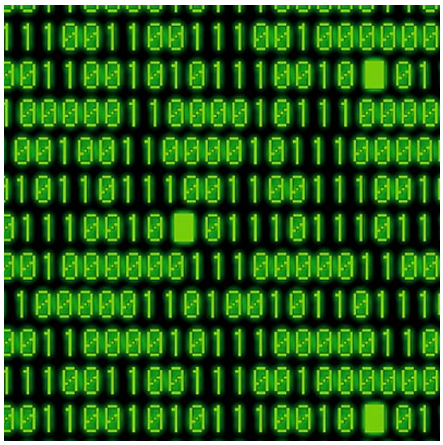
The general answer is Occam's razor; simpler hypotheses are more likely. But this isn't rigorous. It's usually difficult to find a measure of complexity, even for mathematical hypotheses. Is a normal curve simpler than an exponential curve? Bayesian probability

theory doesn’t have anything to say about choosing priors. Thus, many standard "prior distributions" have been developed. Generally, they distribute probability equally across hypotheses. Of course this is a good approach if all the hypotheses are equally likely. But as we saw above, it seems that some hypotheses are more complex than others, and this makes them less likely than the other hypotheses. So when distributing your probability across several hypotheses, you shouldn't necessarily distribute it evenly. There’s also a growing body of work around an idea called the Maximum Entropy Principle. This principle helps you choose a prior that makes the least assumptions given the constraints of the problem. But this principle can’t be used to handle all possible types of hypotheses, only ones for which “entropy” can be mathematically evaluated.

We need a method that everyone can agree provides the correct priors in all situations. This helps us perform induction correctly. It also helps everyone be more honest. Since priors partly determine what people believe, they can sometimes choose priors that help “prove” what they want to prove. This can happen intentionally or unintentionally. It can also happen in formal situations, such as an academic paper defending a proposed program.

To solve the problem of priors once and for all, we'd like to have an acceptable, *universal* prior distribution, so that there's no vagueness in the process of induction. We need a recipe, an *algorithm*, for selecting our priors. For that, we turn to the subject of binary sequences.

Binary Sequences



At this point, we have collected a lot of background material. We know about algorithms, and we know we need an algorithm that does induction. We know that induction also uses Occam’s razor and probability. We know that one of the problems in probability is selecting priors. Now we’re ready to formalize it.

To start, we need a language in which we can express all problems, all data, all hypotheses. *Binary* is the name for representing information using only the characters '0' and '1'. In a sense, binary is the simplest possible alphabet. A two-character alphabet is the smallest alphabet that can communicate a difference. If we had an alphabet of just one character, our “sentences” would be uniform. With two, we can begin to encode information. Each 0 or 1 in a binary sequence (e. g. 01001011) can be considered the answer to a yes-or-no question.

In the above example about sorting numbers, it's easy to convert it to binary just by writing a computer program to follow the algorithm. All programming languages are based on binary. This also applies to anything you’ve ever experienced using a computer. From the greatest movie you’ve ever seen to emotional instant messaging conversations, *all* of it was encoded in binary.

In principle, all information can be represented in binary sequences. If this seems like an extreme claim, consider the following...

All your experiences, whether from your eyes, ears, other senses or even memories and muscle movements, occur between neurons (your nerve cells and brain cells). And it was discovered that neurons communicate using a digital signal called the action potential. Because it is digital, a neuron either sends the signal, or it doesn't. There is no half-sending the action potential. This can be translated directly into binary. An action potential is a 1, no action potential is a 0. All your sensations, thoughts, and actions can be encoded as a binary sequence over time. A really long sequence.

Or, if neuron communication turns out to be more complicated than that, we can look to a deeper level. All events in the universe follow the laws of physics. We're not quite done discovering the true laws of physics; there are some inconsistencies and unexplained phenomena. But the currently proposed laws are incredibly accurate. And they can be represented as a single binary sequence.

You might be thinking, “But I see and do multiple things simultaneously, and in the universe there are trillions of stars all burning at the same time. How can parallel events turn into a single sequence of binary?”

This is a perfectly reasonable question. It turns out that, at least formally, this poses no problem at all. The machinery we will use to deal with binary sequences can turn multiple sequences into one just by dovetailing them together and adjusting how it processes them so the results are the same. Because it is easiest to deal with a single sequence, we do this in the formal recipe. Any good implementation of Solomonoff induction will use multiple sequences just to be faster.

A picture of your daughter can be represented as a sequence of ones and zeros. But a picture is not your daughter. A video of all your daughter’s actions can also be represented as a sequence of ones and zeros. But a video isn’t your daughter, either; we can’t necessarily tell if she’s thinking about cookies, or poetry. The position of all the subatomic particles that make up your daughter as she lives her entire life can be represented as a sequence of binary. And that really *is* your daughter.

Having a common and simple language can sometimes be the key to progress. The ancient Greek mathematician Archimedes discovered many *specific* results of calculus, but could not generalize the methods because he did not have the *language* of calculus. After this language was developed in the late 1600s, hundreds of mathematicians were able to produce new results in the field. Now, calculus forms an important base of our modern civilization.

Being able to do everything in the language of binary sequences simplifies things greatly, and gives us great power. Now we don't have to deal with complex concepts like “daughter” and “soldier." It's all still there in the data, only as a large sequence of 0s and 1s. We can treat it all the same.

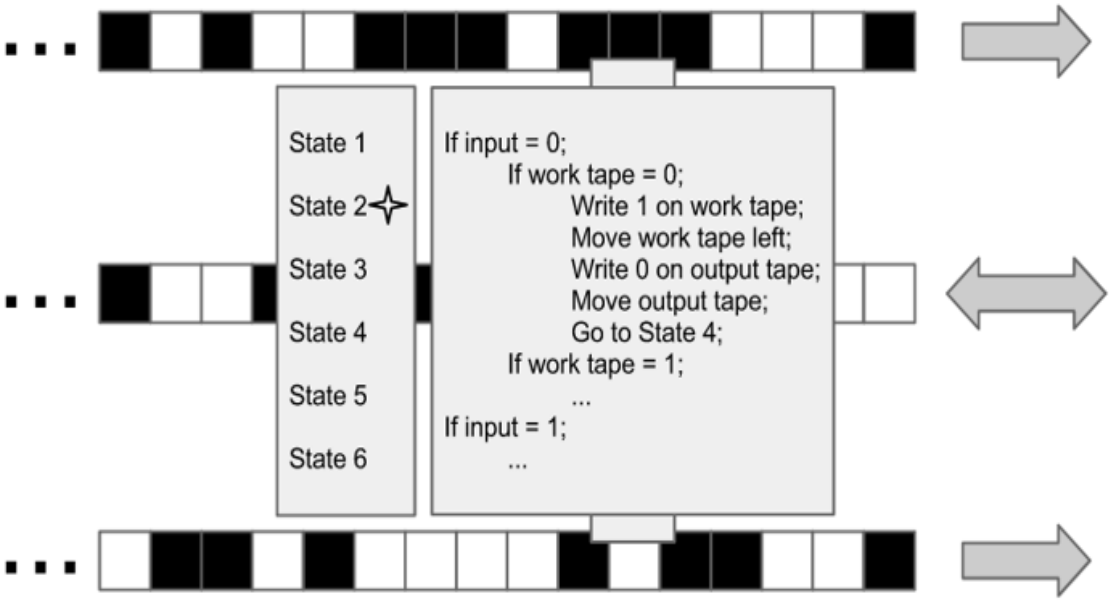
All Algorithms

Now that we have a simple way to deal with all types of data, let's look at hypotheses. Recall that we’re looking for a way to assign prior probabilities to hypotheses. (And then, when we encounter new data, we'll use Bayes' Theorem to update the probabilities we assign to those hypotheses). To be complete, and guarantee we find the *real* explanation for our data, we have to consider *all* possible hypotheses. But how could we ever find all possible explanations for our data? We could sit in a room for days, making a list of all the ways the cookies could be missing, and still not think of the possibility that our wife took some to work.

It turns out that mathematical abstraction can save the day. By using a well-tested model, and the language of binary, we can find all hypotheses.

This piece of the puzzle was discovered in 1936 by a man named Alan Turing. He created a simple, formal model of computers called “Turing machines” before anyone had ever built a computer.

In Turing's model, each machine's language is—you guessed it—binary. There is one binary sequence for the input, a second binary sequence that constantly gets worked on and re-written, and a third binary sequence for output. (This description is called a three-tape Turing machine, and is easiest to think about for Solomonoff induction. The normal description of Turing machines includes only one tape, but it turns out that they are equivalent.)



The rules that determine how the machine reacts to and changes the bits on these tapes are very simple. An example is shown on the diagram above. Basically, every Turing machine has a finite number of “states”, each of which is a little rule. These rules seem bland and boring at first, but in a few paragraphs, you’ll find out why they’re so exciting. First, the machine will start out in a certain state, with some binary on the input tape, all zeros on the work tape, and all zeros on the output tape. The rules for that first state will tell it to look at the input tape and the work tape. Depending on what binary number is on those two tapes, the rules will say to perform certain actions. It will say to;

1. feed the input tape (or not)
2. write 0 or 1 to the work tape
3. move the work tape left or right
4. write 0 or 1 to the output tape
5. feed the output tape (or not).

After that, the rules will say which state to move to next, and the process will repeat. Remember that the rules for these states are fixed; they could be written out on pieces of paper. All that changes are the tapes, and what rule the machine is currently following. The basic mathematics behind this is fairly simple to understand, and can be found in books on computational theory.

This model is *incredibly* powerful. Given the right rules, Turing machines can

- calculate the square of a number,
- run a spreadsheet program,
- compress large files,
- estimate the probability of rain tomorrow,
- control the flight of an airplane,
- play chess better than a human,
- and much, much more.

You may have noticed that this sounds like a list of what regular computers do. And you would be correct; the model of Turing machines came before, and served as an invaluable guide to, the invention of electronic computers. Everything they do is within the model of Turing machines.

Even more exciting is the fact that *all* attempts to formalize the intuitive idea of "algorithm" or "process" have been proven to be at most equally as powerful as Turing machines. If a system has this property, it is called *Turing complete*. For example, math equations using algebra have a huge range of algorithms they can express. Multiplying is an algorithm, finding the hypotenuse of a right triangle is an algorithm, and the quadratic formula is an algorithm. Turing machines can run all these algorithms, *and more*. That is, Turing machines can be used to calculate out all algebraic algorithms, but there are some algorithms Turing machines can run that can't be represented by algebra. This means algebra is not Turing complete. For another example, mathematicians often invent "games" where sequences of symbols can be rewritten using certain rules. (They will then try to prove things about what sequences these rules can create.) But no matter how creative their rules, every one of them can be simulated on a Turing machine. That is, for every set of re-writing rules, there is a binary sequence you can give a Turing machine so that the machine will rewrite the sequences in the same way.

Remember how limited the states of a Turing machine are; every machine has only a finite number of states with "if" rules like in the figure above. But somehow, using these and the tape as memory, they can simulate every set of rules, every algorithm ever thought up. Even the distinctively different theory of quantum computers is at most Turing complete. In the 80 years since Turing's paper, no superior systems have been found. The idea that Turing machines truly capture the idea of "algorithm" is called the Church-Turing thesis.

So the model of Turing machines covers regular computers, but that is not all. As mentioned above, the current laws of physics can be represented as a binary sequence. That is, the laws of physics are an algorithm that can be fed into a Turing machine to compute the past, present and future of the universe. This includes stars burning, the climate of the earth, the action of cells, and even the actions and thoughts of humans. Most of the power here is in the laws of physics themselves. What Turing discovered is that these can be computed by a mathematically simple machine.

As if Turing's model wasn't amazing enough, he went on to prove that *one specific* set of these rules could simulate all *other* sets of rules.

The computer with this special rule set is called a universal Turing machine. We simulate another chosen machine by giving the universal machine a compiler binary sequence. A *compiler* is a short program that translates between computer languages or, in this case, between machines. Sometimes a compiler doesn't exist. For example, you couldn't translate Super Mario Brothers onto a computer that only plays tic-tac-toe. But there will always be a compiler to translate onto a universal Turing machine. We place this compiler in front of the input which would have been given to the chosen machine. From one perspective, we are just giving the universal machine a single, longer sequence. But from another perspective, the universal machine is using the compiler to set itself up to simulate the chosen machine. While the universal machine (using its own, fixed rules) is processing the compiler, it will write various things on the work tape. By the time it has passed the compiler and gets to the original input sequence, the work tape will have something written on it to help simulate the chosen machine. While processing the input, it will still follow its own, fixed rules, only the binary on the work tape will guide it down a different "path" through those rules than if we had only given it the original input.

For example, say that we want to calculate the square of 42 (or in binary, 101010). Assume we know the rule set for the Turing machine which squares numbers when given the number in binary. Given all the specifics, there is an algorithmic way to find the "compiler" sequence based on these rules. Let's say that the compiler is 1011000. Then, in order to compute the square of 42 on the universal Turing machine, we simply give it the input 1011000101010, which is just the compiler 1011000 next to the number 42 as 101010. If we want to calculate the square of 43, we just change the second part to 101011 (which is 101010 + 1). The compiler sequence doesn't change, because it is a property of the machine we want to simulate, e. g. the squaring machine, but not of the input to that simulated machine, e. g. 42.

In summary: algorithms are represented by Turing machines, and Turing machines are represented by inputs to the universal Turing machine. Therefore, algorithms are represented by inputs to the universal Turing machine.

In Solomonoff induction, the assumption we make about our data is that it was generated by some algorithm. That is, the hypothesis that explains the data is an algorithm. Therefore, a universal Turing machine can output the data, as long as you give the machine the correct hypothesis as input. Therefore, the set of all possible inputs to our universal Turing machine is the set of all possible hypotheses. This includes the hypothesis that the data is a list of the odd numbers, the hypothesis that the enemy soldier is a sniper, and the hypothesis that your daughter ate the cookies. This is the power of formalization and mathematics.

Solomonoff's Lightsaber

Now we can find all the hypotheses that would predict the data we have observed. This is much more powerful than the informal statement of Occam's razor. Because of its precision and completeness, this process has been jokingly dubbed "Solomonoff's Lightsaber". Given our data, we find potential hypotheses to explain it by running every hypothesis, one at a time, through the universal Turing machine. If the output matches our data, we keep it. Otherwise, we throw it away.

By the way, this is where Solomonoff induction becomes incomputable. It would take an infinite amount of time to check every algorithm. And even more problematic, some of the algorithms will run forever without producing output—*and we can't prove they will never stop running*. This is known as the halting problem, and it is a deep fact of

the theory of computation. It's the sheer number of algorithms, and these pesky non-halting algorithms that stop us from actually running Solomonoff induction.

The actual process above might seem a little underwhelming. We just check every single hypothesis? Really? Isn't that a little mindless and inefficient? This will certainly not be how the first true AI operates. But don't forget that before this, nobody had any idea how to do ideal induction, *even in principle*. Developing fundamental theories, like quantum mechanics, might seem abstract and wasteful. But history has proven that it doesn't take long before such theories and models change the world, as quantum mechanics did with modern electronics. In the future, men and women will develop ways to approximate Solomonoff induction in a second. Perhaps they will develop methods to eliminate large numbers of hypotheses all at once. Maybe hypotheses will be broken into distinct classes. Or maybe they'll use methods to statistically converge toward the right hypotheses.

So now, at least in theory, we have the whole list of hypotheses that might be the true cause behind our observations. These hypotheses, since they are algorithms, look like binary sequences. For example, the first few might be 01001101, 00110101100001101001001l0, and 1000111110111111000111010010100001. That is, for each of these three, when you give them to the universal Turing machine as input, the output is our data. Which of these three do you think is more likely to be the *true* hypothesis that generated our data in the first place?

We have a list, but we're trying to come up with a probability, not just a list of possible explanations. So how do we decide what the probability is of each of these hypotheses? Imagine that the true algorithm is produced in a most unbiased way: by flipping a coin. For each bit of the hypothesis, we flip a coin. Heads will be 0, and tails will be 1. In the example above, 01001101, the coin landed heads, tails, heads, heads, tails, and so on. Because each flip of the coin has a 50% probability, each bit contributes ½ to the final probability.

Therefore an algorithm that is one bit longer is half as likely to be the true algorithm. Notice that this intuitively fits Occam's razor; a hypothesis that is 8 bits long is much more likely than a hypothesis that is 34 bits long. Why bother with extra bits? We'd need *evidence* to show that they were necessary.

So, why not just take the shortest hypothesis, and call that the truth? Because all of the hypotheses predict the data we have so far, and in the future we might get data to rule out the shortest one. We keep all consistent hypotheses, but weigh the shorter ones with higher probability. So in our eight-bit example, the probability of 01001101 being the true algorithm is ½^8, or 1/256. It's important to say that this isn't an absolute probability in the normal sense. It hasn't been *normalized*—that is, the probabilities haven't been adjusted so that they add up to 1. This is computationally much more difficult, and might not be necessary in the final implementation of Solomonoff induction. These probabilities can still be used to compare how likely different hypotheses are.

To find the probability of the evidence alone, all we have to do is add up the probability of all these hypotheses consistent with the evidence. Since any of these hypotheses

could be the true one that generates the data, and they're mutually exclusive, adding them together doesn't double count any probability.

Formalized Science

Let’s go back to the process above that describes the scientific method. We’ll see that Solomonoff induction is *this process made into an algorithm*.

1. Make an observation.

Our observation is our binary sequence of data. Only binary sequences are data, and all binary sequences can qualify as data.

2. Form a hypothesis that explains the observation.

We use a universal Turing machine to find all possible hypotheses, no fuzziness included. The hypothesis “explains” the observation if the output of the machine matches the data exactly.

3. Conduct an experiment that will test the hypothesis.

The only “experiment” is to observe the data sequence for longer, and run the universal machine for longer. The hypotheses whose output continues to match the data are the ones that pass the test.

4. If the experimental results disconfirm the hypothesis, return to step #2 and form a hypothesis not yet used. If the experimental results confirm the hypothesis, provisionally accept the hypothesis.

This step tells us to repeat the “experiment” with each binary sequence in our matching collection. However, instead of “provisionally” accepting the hypothesis, we accept all matching hypotheses with a probability weight according to its length.

Now we’ve found the truth, as best as it can be found. We’ve excluded no possibilities. There’s no place for a scientist to be biased, and we don’t need to depend on our creativity to come up with the right hypothesis or experiment. We know how to measure how complex a hypothesis is. Our only problem left is to efficiently run it.

Approximations

As mentioned before, we actually can’t run all the hypotheses to see which ones match. There are infinitely many, and some of them will never halt. So, just like our cake recipe, we need some very helpful approximations that still deliver very close to true outputs. Technically, all prediction methods are approximations of Solomonoff induction, because Solomonoff induction tries all possibilities. But most methods use a *very* small set of hypotheses, and many don't use good methods for estimating their probabilities. How can we more directly approximate our recipe? At present, there aren't any outstandingly fast and accurate approximations to Solomonoff induction. If there were, we would be well on our way to true AI. Below are some ideas that have been published.

In *Universal Artificial Intelligence*, Marcus Hutter provides a full formal approximation of Solomonoff induction which he calls AIXI-*tl*. This model is optimal in a technical and subtle way. Also it can always return an answer in a finite amount of time, but this time is usually extremely long and *doubles* with every bit of data. It would still take longer

than the life of the universe before we got an answer for most questions. How can we do even better?

One general method would be to use a Turing machine that you know will always halt. Because of the halting problem, this means that you won't be testing some halting algorithms that might be the correct hypotheses. But you could still conceivably find a large set of hypotheses that all halted.

Another popular method of approximating any intractable algorithm is to use randomness. This is called a Monte Carlo method. We can't test *all* hypotheses, so we have to select a subset. But we don't want our selection process to bias the result. Therefore we could randomly generate a bunch of hypotheses to test. We could use an evolutionary algorithm, where we test this seed set of hypotheses, and keep the ones that generate data closest to ours. Then we would vary these hypotheses, run them again through the Turing machine, keep the closest fits, and continue this process until the hypotheses actually predicted our data exactly.

The mathematician Jürgen Schmidhuber proposes a different probability weighing which also gives high probability to hypotheses that can be quickly computed. He demonstrates how this is "near-optimal". This is vastly faster, but risks making another assumption, that faster algorithms are inherently more likely.

Unresolved Details

Solomonoff induction is an active area of research in modern mathematics. While it is universal in a broad and impressive sense, some choices can still be made in the mathematical definition. Many people also have philosophical concerns or objections to the claim that Solomonoff induction is ideal and universal.

The first question many mathematicians ask is, "Which universal Turing machine?" I have written this tutorial as if there is only one, but there are in fact infinitely many sets of rules that can simulate all other sets of rules. Just as the length of a program will depend on which programming language you write it in, the length of the hypothesis as a binary sequence will depend on which universal Turing machine you use. This means the probabilities will be different. The change, however, is very limited. There are theorems that well-define this effect and it is generally agreed not to be a concern. Specifically, going between two universal machines cannot increase the hypothesis length any more than the length of the compiler from one machine to the other. This length is fixed, independent of the hypothesis, so the more data you use, the less this difference matters.

Another concern is that the true hypothesis may be incomputable. There are known definitions of binary sequences which make sense, but which no Turing machine can output. Solomonoff induction would converge to this sequence, but would never predict it exactly. It is also generally agreed that *nothing* could ever predict this sequence, because all known predictors are equivalent to Turing machines. If this is a problem, it is similar to the problem of a finite universe; there is nothing that can be done about it.

Lastly, many people, mathematicians included, reject the ideas behind the model, such as that the universe can be represented as a binary sequence. This often delves into complex philosophical arguments, and often revolves around consciousness.

Many more details can be found the more one studies. Many mathematicians work on modified versions and extensions where the computer learns how to act as well as predict. However these open areas are resolved, Solomonoff induction has provided an invaluable model and perspective for research into solving the problem of how to find truth. (Also see Open Problems Related to Solomonoff Induction.)

Fundamental theories have an effect of unifying previously separate ideas. After Turing discovered the basics of computability theory, it was only a matter of time before these ideas made their way across philosophy and mathematics. Statisticians could only find exact probabilities in simplified situations; now they know how to find them in all situations. Scientists wondered how to really know which hypothesis was simpler; now they can find a number. Philosophers wondered how induction could be justified. Solomonoff induction gives them an answer.

So, what’s next? To build our AI truth-machine, or even to find the precise truth to a single real-world problem, we need considerable work on approximating our recipe. Obviously a scientist cannot presently download a program to tell him the complexity of his hypothesis regarding deep-sea life behavior. But now that we know how to find truth in principle, mathematicians and computer scientists will work on finding it in practice.

(How does this fit with Bayes' theorem? A followup.)

Negative and Positive Selection

(Originally posted to my blog, The Rationalist Conspiracy; cross-posted here on request of Lukeprog.)

You're the captain of a team, and you want to select really good players. How do you do it?

One way is through what I call *positive selection*. You devise a test – say, who can run the fastest – and pick the people who do best. If you want to be really strict, like if you're selecting for the Olympics, you only pick the top fraction of a percent. If you're a player, and you want to get selected, you have to train to do better on the test.

The opposite method is *negative selection.* Instead of one test to pick out winners, you design many tests to pick out losers. You test, say, who can't run very well when it's hot out, and get rid of them. Then you test who can't run very well when it's cold out, and get rid of them. Then you test running in the rain, and get rid of the losers there. And so on and so forth. When you're strict with negative selection, you have lots and lots of tests, so that it's very hard for any one person to pass through all the filters.

I think a big part of where American society's gone wrong over the last hundred years is the ubiquitous use of *negative selection* over *positive selection*. (Athletics is one of the only exceptions. It's apparently so important that people really care about performance – as opposed to, say, in medicine, where we exclude brilliant doctors if they don't have the stamina to work ninety hours a week.) A single test can always be flawed; for example, IQ tests and SATs have many flaws. However, with negative selection, how badly you do is determined by the failure rate of every test *combined.* If you have twenty tests, and even one of them is so flawed it excludes good players, then your team will suck.

Elite college admissions is an example of a negative selection test. There's no one way you can do really, really well, and thereby be admitted to Harvard. Instead, you have to pass a bunch of different selection filters: Are your SATs good enough? Are your grades good enough? Is your essay good enough? Are your extracurriculars good enough? Are your recommendations good enough? Failure on any one step usually means not getting admitted. And as competition has intensified, colleges have added more and more filters, like the supplemental applications top schools now require (in addition to the Common Application). It wasn't always this way – Harvard used to admit primarily based on an entrance exam – until they discovered this let too many Jews in (no, seriously). More recently, the negative selection has been intensified by eliminating the SAT's high ceiling.

Academia is another example of negative selection. To get tenure, first you have to get into a top PhD program. Then you have to graduate. Then you have to get a good recommendation from your advisor. Then you have to get a good postdoc. Then you have to get another good postdoc. Then you have to get a good assistant professorship. Then you have to get approved by the tenure committee. For the most part, if even one of those steps goes wrong – if you went to a second-tier PhD program, say – there's no way to recover. Once you're off the "track", you're off, and there's no getting back on. It's fail once, fail forever.

Grades are another example – A is a good grade, but there's no *excellent* grade. There's no grade that you only get if you're in the top 0.1%. Hence, getting a really good GPA doesn't mean excelling, so much as it means never failing. If you're in high school and are taking six classes, if you fail one, your GPA is now 3.3 or less, regardless of how good you are otherwise.

In any field, at the top end, you tend to get a lot of variance. (Insert tales of the mad artist and mad mathematician.) Negative selection suppresses variance, by eliminating many of the dimensions on which people vary. Students at Yale are, for the most part, all strikingly similar – same socioeconomic class, same interests, same pursuits, same life goals, even the same style of dress. A lot of people tend to assume performance follows a bell curve, but in some cases, it's more like a Pareto distribution: the top people do hundreds or thousands of times better than average. Hence, if you eliminate the small fraction of people at the very top, your performance is hosed. Fortunately for VC funds, the startup world is still positive selection.

Less obviously, a world with lots of negative selection might be a nasty one to live in. If you think of yourself as trying to *eliminate bad*, rather than *encourage good*, you start operating on the purity vs. contamination moral axis. Any tiny amount of bad, anywhere, must be gotten rid of, and that can lead to all sorts of nastiness. "When you are a Guardian of the Truth, all you can do is try to stave off the inevitable slide into entropy by zapping anything that departs from the Truth. If there's some way to pump against entropy, generate new true beliefs along with a little waste heat, that same pump can keep the truth alive without secret police."

Neuroscience basics for LessWrongians

The origins of this article are in my partial transcript of the live June 2011 debate between Robin Hanson and Eliezer Yudkowsky. While I still feel like I don't entirely understand his arguments, a few of his comments about neuroscience made me strongly go, "no, that's not right."

Furthermore, I've noticed that while LessWrong in general seems to be very strong on the psychological or "black box" side of cognitive science, there isn't as much discussion of neuroscience here. This is somewhat understandable. Our current understanding of neuroscience is frustratingly incomplete, and too much journalism on neuroscience is sensationalistic nonsense. However, I think what we do know is worth knowing. (And part of what makes much neuroscience journalism annoying is that it makes a big deal out of things that are *totally unsurprising,* given what we *already know.)*

My qualifications to do this: while my degrees are in philosophy, for awhile in undergrad I was a neuroscience major, and ended up taking quite a bit of neuroscience as a result. This means I can assure you that most of what I say here is standard neuroscience which could be found in an introductory textbook like Nichols, Martin, Wallace, & Fuchs' *From Neuron to Brain* (one of the text books I used as an undergraduate). The only things that might not be totally standard are the conjecture I make about how complex currently-poorly-understood areas of the brain are likely to be, and also some of the points I make in criticism of Eliezer at the end (though I believe these are not a very big jump from current textbook neuroscience.)

One of the main themes of this article will be specialization within the brain. In particular, we know that the brain is divided into specialized areas at the macro level, and we understand some (though not very much) of the micro-level wiring that supports this specialization. It seems likely that each region of the brain has its own micro-level wiring to support its specialized function, and in some regions the wiring is likely to be quite complex.

1. Specialization of brain regions

One of the best-established facts about the brain is that specific regions handle specific functions. And it isn’t just that in each individual, specific brain regions handle specific functions. It’s also that which regions handle which functions is consistent across individuals. This is an extremely well-established finding, but it’s worth briefly summarizing some of the evidence for it.

One kind of evidence comes from experiments involving direct electrical stimulation of the brain. This cannot ethically be done on humans without a sound medical reason, but it is used with epileptic patients in order to determine the source of the problem, which is necessary in order to treat epilepsy surgically.

In epileptic patients, stimulating certain regions of the brain (known as the primary sensory areas) causes the patient to report sensations: sights, sounds, feelings, smells, and tastes. Which sensations are caused by stimulating which regions of the

brain is consistent across patients. This is the source of the “Penfield homunculus,” a map of brain regions which, when stimulated, result in touch sensations which patients describe as feeling like they come from particular parts of the body. Stimulating one region, for example, might consistently lead to a patient reporting a feeling in his left foot.

Regions of the brain associated with sensations are known as sensory areas or sensory cortex. Other regions of the brain, when stimulated, lead to involuntary muscle movements. Those areas are known as motor areas or motor cortex, and again, which areas correspond to which muscles is consistent across patients. The consistency of the mapping of brain regions across patients is important, because it’s evidence of an innate structure to the brain.

An even more significant kind of evidence comes from studies of patients with brain damage. Brain damage can produce very specific ability losses, and patients with damage to the same areas will typically have similar ability losses. For example, the rear most part of the human cerebral cortex is the primary visual cortex, and damage to it results in a phenomenon known as cortical blindness. That is to say, the patient is blind in spite of having perfectly good eyes. Their other mental abilities may be unaffected.

That much is not surprising, given what we know from studies involving electrical stimulation, but ability losses from brain damage can be strangely specific. For example, neuroscientists now believe that one function of the temporal lobe is to recognize objects and faces. A key line of evidence for this is that patient with damage to certain parts of the temporal lobe will be unable to identify those things by sight, even though they may be able to describe the objects in great detail. Here is neurologist Oliver Sacks’ description of an interaction with one such patient, the titular patient in Sacks’ book *The Man Who Mistook His Wife For a Hat:*

> ‘What is this?’ I asked, holding up a glove.
>
> ‘May I examine it?’ he asked, and, taking it from me, he proceeded to examine it as he had examined the geometrical shapes.
>
> ‘A continuous surface,’ he announced at last, ‘infolded on itself. It appears to have’—he hesitated—’five outpouchings, if this is the word.’
>
> ‘Yes,’ I said cautiously. You have given me a description. Now tell me what it is.’
>
> ‘A container of some sort?’
>
> ‘Yes,’ I said, ‘and what would it contain?’
>
> ‘It would contain its contents!’ said Dr P., with a laugh. ‘There are many possibilities. It could be a change purse, for example, for coins of five sizes. It could ...’
>
> I interrupted the barmy flow. ‘Does it not look familiar? Do you think it might contain, might fit, a part of your body?’
>
> No light of recognition dawned on his face. (Later, by accident, he got it on, and exclaimed, ‘My God, it’s a glove!’)

The fact that damage to certain parts of the temporal lobe results in an inability to recognize objects contains an extremely important lesson. For most of us, recognizing objects requires no effort or thought as long as we can see the object clearly. Because it's easy for us, it might be tempting to think it's an inherently easy task, one that shouldn't require hardly any brain matter to perform. Certainly it never occurred to me before I studied neuroscience that object recognition might require a special brain region. But it turns out that tasks that seem easy to us can in fact require such a specialized region.

Another example of this fact comes from two brain regions involved in language, Broca's area and Wernicke's area. Damage to each area leads to distinct types of difficulties with language, known as Broca's aphasia and Wernicke's aphasia, respectively. Both are strange conditions, and my description of them may not give a full sense of what they are like. Readers might consider searching online for videos of interviews Broca's aphasia and Wernicke's aphasia patients to get a better idea of what the conditions entail.

Broca's aphasia is a loss of ability to produce language. In one extreme case, one of the original cases studied by Paul Broca, a patient was only able to say the word "tan." Other patients may have a less limited vocabulary, but still struggle to come up with words for what they want to say. And even when they can come up with individual words, they may be unable to put them into sentences. However, patients with Broca's aphasia appear to have no difficulty understanding speech, and show awareness of their disability.

Wernicke's aphasia is even stranger. It is often described as an inability to understand language while still being able to produce language. However, while patients with Wernicke's aphasia may have little difficulty producing complete, grammatically correct sentences, the sentences tend to be nonsensical. And Wernicke's patients often act as if they are completely unaware of their condition. A former professor of mine once described a Wernicke's patient as sounding "like a politician," and from watching a video of an interview with the patient, I agreed: I was impressed by his ability to confidently utter nonsense.

The fact of these two forms of aphasia suggest that Broca's area and Wernicke's area have two very important and distinct roles in our ability to produce and understand language. And I find this fact strange to write about. Like object recognition, language comes naturally to us. As I write this, my intuitive feeling is that the work I am doing comes mainly in the ideas, plus making a few subtle stylistic decisions. I know from neuroscience that I would be unable to write this if I had significant damage to either region. Yet I am totally unconscious of the work they are doing for me.

2. Complex, specialized wiring within regions

"Wiring" is a hard metaphor to avoid when talking about the brain, but it is also a potentially misleading one. People often talk about "electrical signals" in the brain, but unlike electrical signals in human technology, which involves movement of electrons between the atomics of the conductor, signals in the human brain involve movement of ions and small molecules across cell membranes and between cells.

Furthermore, the first thing most people who know a little bit about neuroscience will think of when they hear the word “wiring” is axons and dendrites, the long skinny projections along which signals are transmitted from neuron to neuron. But it isn’t just the layout of axons and dendrites that matters. Ion channels, and the structures that transport neurotransmitters across cell membranes, are also important.

These can vary a lot at the synapse, the place where two neurons touch. For example, synapses vary in strength, that is to say, the strength of the one neuron’s effect on the other neuron. Synapses can also be excitatory (activity in one leads increased activity in the other) or inhibitory (activity in one leads to decreased activity in the other). And this is just a couple of the ways synapses can vary; the details can be somewhat complicated, and I’ll give one example of how the details can be complicated later.

I say all this just to make clear what I mean when I talk about how the brain’s “wiring.” By “wiring,” I mean all the features of the physical structures that connect neurons to each other and which are relevant for understanding how the brain works. I mean to all the things I’ve mentioned above, and anything I may have omitted. It’s important to have a word to talk about this wiring, because what (admittedly little) we understand about how the brain works we understand in terms of this wiring.

For example, the nervous system actually first begins processing visual information in the retina (the part of the eye at the back where our light receptors are). This is done by what’s known as the center-surround system: a patch of light receptors, when activated, excites one neuron, but nearby patches of light receptors, when activated, inhibit that same neuron (sometimes, the excitatory roles and inhibitory roles are reversed).

The effect of this is that what the neurons are sensitive to is not light itself, but contrast. They’re contrast detectors. And what allows them to detect contrast isn’t anything magical, it’s just the wiring, the way the neurons are connected together.

This technique of getting neurons to serve specific functions based on how they are wired together shows up in more complicated ways in the brain itself. There’s one line of evidence for specialization of brain regions that I saved for this section, because it also tells us about the details of how the brain is wired. That line of evidence is recordings from the brain using electrodes.

For example, during the 50’s David Hubel and Torsten Weisel did experiments where they paralyzed the eye muscles of animals, stuck electrodes in primary visual areas of the animals’ brains, and then showed the animals various images to see which images would cause electrical signals in the animals’ primary visual areas. It turned out that the main thing that causes electrical signals in the primary visual areas is lines.

In particular, a given cell in the primary visual area will have a particular orientation of line which it responds to. It appears that the way these line-orientation detecting cells work is that they receive input from several contrast detecting cells which, themselves, correspond to regions of the retina that are themselves all in a line. A line in the right position and orientation will activate all of the contrast-detecting cells, which in turn activates the line-orientation detecting cell. A line in the right position but wrong orientation will activate only one or a few contrast-detecting cells, not enough to activate the line-orientation detecting cell.

[If this is unclear, a diagram like the one on Wikipedia may be helpful, though Wikipedia's diagram may not be the best.]

Another example of a trick the brain does with neural wiring is locating sounds using what are called “interaural time differences.” The idea is this: there is a group of neurons that receives input from both ears, and specifically responds to *simultaneous* input from both ears. However, the axons running from the ears to the neurons in this group of cells vary in length, and therefore they vary in how long it takes them to get a signal from each ear.

This means that which cells in this group respond to a sound depends on whether or not the sound reaches reaches the ears at the same time or at different times, and (if at different times) on how big the time difference is. If there’s no difference, that means the sound came from directly ahead or behind (or above or below). A big difference, with the sound reaching the left ear first, indicates the sound came from the left. A big difference, with the sound reaching the right ear first, indicates the sound came from the left. Small differences indicate something in between.

[A diagram might be helpful here too, but I'm not sure where to find a good one online.]

I’ve made a point to mention these bits of wiring because they’re cases where neuroscientists have a clear understanding of how it is that a particular neuron is able to fire only in response to a particular kind of stimulus. Unfortunately, cases like this are relatively rare. In other cases, however, we at least know that particular neurons respond specifically to more complex stimuli, even though we don’t know why. In rats, for example, there are cells in the hippocampus that activate only when the rat is in a particular location; apparently their purpose is to keep track of the rat’s location.

The visual system gives us some especially interesting cases of this sort. We know that the primary visual cortex sends information to other parts of the brain in two broadly-defined pathways, the dorsal pathway and the ventral pathway. The dorsal pathway appears to be responsible for processing information related to position and movement. Some cells in the dorsal pathway, for example, fire only when an animal sees an object moving in a particular direction.

The most interesting cells of this sort that neuroscientists have found so far, though, are probably some of the cells in the medial temporal lobe, which is part of the ventral pathway. In one study (Quiroga et al. 2005), researchers took epileptic patients who had had electrodes implanted in their brains in order to locate the source of their epilepsy and showed them pictures of various people, objects, and landmarks. What the researchers found is that the neuron or small group of neurons a given electrode was reading from typically only responded to pictures of one person or thing.

Furthermore, a particular electrode often got readings from very different pictures of a single person or thing, but not similar pictures of different people or things. In one notorious example, they found a neuron that they could only get to respond to either pictures of actress Halle Berry or the text “Halle Berry.” This included drawings of the actress, as well as pictures of her dressed as Catwoman (a role which she had recently performed at the time the study was performed), but not other drawings or other pictures of Catwoman.

What’s going on here? Based on what we know about the wiring of contrast-detectors and orientation-detectors the following conjecture seems highly likely: if we were to map out the brain completely and then follow the path along which visual information is transmitted, we would find that neurons gradually come to be wired together in more and more complex ways, to allow them to gradually become specific to more

and more complex features of visual images. This, I think, is an extremely important inference.

We know that experience can impact the way the brain is wired. In fact, some aspects of the brain's wiring seem to have evolved specifically to be able to change in response to experience (the main wiring of that sort we know about is called Hebbian synapses, but the details aren't important here). And it is actually somewhat difficult to draw a clear line between features of the brain that are innate and features of the brain that are the product of learning, because some fairly basic features of the brain depend on outside cues in order to develop.

Here, though, I'll use the word "innate" to refer to features of the brain that will develop given the overwhelming majority of the conditions animals of a given species actually develop under. Under that definition, a "Halle Berry neuron" is highly unlikely to be innate, because there isn't enough room in the brain to have a neuron specific to every person a person might possibly learn about. Such neural wiring is almost certainly the result of learning.

But importantly, the underlying structure that makes such learning possible is probably at least somewhat complicated, and also specialized for that particular kind of learning. This is because such person-specific and object-specific neurons are not found in all regions of the brain, there must be something special about the medial temporal lobe that allows such learning to happen there.

Similar reasoning applies to regions of the brain that we know even less about. For example, it seems likely that Broca's area and Wernicke's area both contain specialized wiring for handling language, though we have little idea how that wiring might perform its function. Given that humans seem to have a considerable innate knack for learning language (Pinker 2007), it again seems likely that the wiring is somewhat complicated.

3. On some problematic comments by Eliezer

I agree with Singularity Institute positions on a great deal. After all, I recently made my first donation to the Singularity Institute. But here, I want to point out some problematic neuroscience-related comments in Eliezer's debate with Robin Hanson:

> If you actually look at the genome, we've got about 30,000 genes in here. Most of our 750 megabytes of DNA is repetitive and almost certainly junk, as best we understand it. And the brain is simply not a very complicated artifact by comparison to, say, Windows Vista. Now the complexity that it does have it uses a lot more effectively than Windows Vista does. It probably contains a number of design principles which Microsoft knows not. (And I'm not saying it's that small because it's 750 megabytes, I'm saying it's gotta be that small because at least 90% of the 750 megabytes is junk and there's only 30,000 genes for the whole body, never mind the brain.)
>
> That something that simple can be this powerful, and this hard to understand, is a shock. But if you look at the brain design, it's got 52 major areas on each side of the cerebral cortex, distinguishable by the local pattern, the tiles and so on, it just doesn't really look all that complicated. It's very powerful. It's very mysterious.

> What can say about it is that it probably involves 1,000 different deep, major, mathematical insights into the nature of intelligence that we need to comprehend before we can build it.

Though this is not explicit, there appears to be an inference here that, in order for something so simple to be so powerful, it must incorporate many deep insights into intelligence, though we don't know what most of them are. There are several problems with this argument.

First of all, it is not true that the fact that that brain is divided into only 52 major areas is evidence that it is not very complex, because knowing about the complexity of its macroscopic organization tells us nothing about the complexity of its microscopic wiring. The brain consists of tens of billions of neurons, and a single neuron can make hundreds of synapses with other neurons. The details of how synapses are set up vary greatly. The fact is that under a microscope, the brain at least looks very complex.

The argument from the small size of the genome is more plausible, especially if Eliezer is thinking in terms of Kolmogorov complexity, which is based on the size of the smallest computer program needed to build something. However, it does not follow that if the genome is not very complex, the brain must not be very complex, because the brain may be built not just based on the genome, but also based on information from the outside environment. We have good reason to think this is how the brain is actually set up, not just in cases we would normally associate with learning and memory, but with some of the most basic and near-universal features of the brain. For example, in normal mammals, the neurons in the visual cortex are organized into "ocular dominance columns," but these fail to form if the animal is raised in darkness.

More importantly, there is no reason to think getting a lot of power out of a relatively simple design requires insights into the nature of intelligence itself. To use Eliezer's own example of Windows Vista: imagine if, for some reason, Microsoft decided that it was very important for the next generation of its operating system to be highly compressible. Microsoft tells this to its programmers, and they set about looking for ways to make an operating system do most of what the current version of Windows does while being more compressible. They end up doing a lot of things that are only applicable to their situation, and couldn't be used to make a much more powerful operating system. For example, they might look for ways to recycle pieces of code, and make particular pieces of code do as many different things in the program as possible.

In this case, would we say that they had discovered deep insights into how to build powerful operating systems? Well no. And there's reason to think that life on Earth uses similar tricks to get a lot of apparent complexity out of relatively simple genetic codes. Genes code for protein. In a phenomenon known as "alternative splicing," there may be several ways to combine the parts of a gene, allowing one gene to code for several proteins. And even a single, specific protein may perform several roles within an organism. A receptor protein, for example, may be plugged into different signaling cascades in different parts of an organism.

Eliezer's comments about the complexity of brain are only a small part of his arguments in the debate, but I worry that comments like these by people concerned with the future of Artificial Intelligence are harmful insofar as they may lead some people (particularly neuroscientists) to conclude AI-related futurism is a bunch of confusions based in ignorance. *I* don't think it is, but a neuroscientist taking the Hanson-Yudkowsky debate as an introduction to the issues could easily conclude that.

Of course, that's not the most important reason for people with an interest in AI to understand the basics of neuroscience. The most important reason is that understanding some neuroscience will help clarify your thinking about the rest of cognitive science.

References:

Pinker, S. 2007. *The Language Instinct.* Harper Perennial Modern Classics.

Quiroga, R. Q. et al. 2005. Invariant visual representation by single neurons in the human brain. *Nature,* 435, 1102-1107.

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.

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.

The Mere Cable Channel Addition Paradox

The following is a dialogue intended to illustrate what I think may be a serious logical flaw in some of the conclusions drawn from the famous Mere Addition Paradox.

EDIT: To make this clearer, the interpretation of the Mere Addition Paradox this post is intended to criticize is the belief that a world consisting of a large population full of lives barely worth living is the *optimal* world. That is, I am disagreeing with the idea that the best way for a society to use the resources available to it is to create as many lives barely worth living as possible. Several commenters have argued that another interpretation of the Mere Addition Paradox is that a sufficiently large population with a lower quality of life will always be better than a smaller population with a higher quality of life, even if such a society is far from optimal. I agree that my argument does not necessarily refute this interpretation, but think the other interpretation is common enough that it is worth arguing against.

EDIT: On the advice of some of the commenters I have added a shorter summary of my argument in non-dialogue form at the end. Since it is shorter I do not think it summarizes my argument as completely as the dialogue, but feel free to read it instead if pressed for time.

Bob: Hi, I'm with R&P cable. We're selling premium cable packages to interested customers. We have two packages to start out with that we're sure you love. Package A+ offers a larger selection of basic cable channels and costs $50. Package B offers a larger variety of exotic channels for connoisseurs, it costs $100. If you buy package A+, however, you'll get a 50% discount on B.

Alice: That's very nice, but looking at the channel selection, I just don't think that it will provide me with enough utilons.

Bob: Utilons? What are those?

Alice: They're the unit I use to measure the utility I get from something. I'm really good at shopping, so if I spend my money on the things I usually spend it on I usually get 1.5 utilons for every dollar I spend. Now, looking at your cable channels, I've calculated that I will get 10 utilons from buying Package A+ and 100 utilons from buying Package B. Obviously the total is 110, significantly less than the 150 utilons I'd get from spending $100 on other things. It's just not a good deal for me.

Bob: You think so? Well it so happens that I've met people like you in the past and have managed to convince them. Let me tell you about something called the "Mere Cable Channel Addition Paradox."

Alice: Alright, I've got time, make your case.

Bob: Imagine that the government is going to give you $50. Sounds like a good thing, right?

Alice: It depends on where it gets the $50 from. What if it defunds a program I think is important?

Bob: Let's say that it would defund a program that you believe is entirely neutral. The harms the program causes are exactly outweighed by the benefits it brings, leaving a net utility of zero.

Alice: I can't think of any program like that, but I'll pretend one exists for the sake of the argument. Yes, defunding it and giving me $50 would be a good thing.

Bob: Okay, now imagine the program's beneficiaries put up a stink, and demand the program be re-instituted. That would be bad for you, right?

Alice: Sure. I'd be out $50 that I could convert into 75 utilons.

Bob: Now imagine that the CEO of R&P Cable Company sleeps with an important senator and arranges a deal. You get the $50, but you have to spend it on Package A+. That would be better than not getting the money at all, right?

Alice: Sure. 10 utilons is better than zero. But getting to spend the $50 however I wanted would be best of all.

Bob: That's not an option in this thought experiment. Now, imagine that after you use the money you received to buy Package A+, you find out that the 50% discount for Package B still applies. You can get it for $50. Good deal, right?

Alice: Again, sure. I'd get 100 utilons for $50. Normally I'd only get 75 utilons.

Bob: Well, there you have it. By a *mere addition* I have demonstrated that a world where you have bought both Package A+ and Package B is better than one where you have neither. The only difference between the hypothetical world I imagined and the world we live in is that in one you are spending money on cable channels. A mere addition. Yet you have admitted that that world is better than this one. So what are you waiting for? Sign up for Package A+ and Package B!

And that's not all. I can keep adding cable packages to get the same result. The end result of my logic, which I think you'll agree is impeccable, is that you purchase Package Z, a package where you spend all the money other than that you need for bare subsistence on cable television packages.

Alice: That seems like a pretty repugnant conclusion.

Bob: It still follows from the logic. For every world where you are spending your money on whatever you have calculated generates the most utilons there exists another, better world where you are spending all your money on premium cable channels.

Alice: I think I found a flaw in your logic. You didn't perform a "mere addition." The hypothetical world differs from ours in two ways, not one. Namely, in this world the government isn't giving me $50. So your world doesn't just differ from this one in terms of how many cable packages I've bought, it also differs in how much money I have to buy them.

Bob: So can I interest you in a special form of the package? This one is in the form of a legally binding pledge. You pledge that if you ever make an extra $50 in the future you will use it to buy Package A+.

Alice: No. In the scenario you describe the only reason buying Package A+ has any value is that it is impossible to get utility out of that money any other way. If I just get $50 for some reason it's more efficient for me to spend it normally.

Bob: Are you sure? I've convinced a lot of people with my logic.

Alice: Like who?

Bob: Well, there were these two customers named Michael Huemer and Robin Hanson who both accepted my conclusion. They've both mortgaged their homes and started sending as much money to R&P cable as they can.

Alice: There must be some others who haven't.

Bob: Well, there was this guy named Derek Parfit who seemed disturbed by my conclusion, but couldn't refute it. The best he could do is mutter something about how the best things in his life would gradually be lost if he spent all his money on premium cable. I'm working on him though, I think I'll be able to bring him around eventually.

Alice: Funny you should mention Derek Parfit. It so happens that the flaw in your "Mere Cable Channel Addition Paradox" is exactly the same as the flaw in a famous philosophical argument he made, which he called the "Mere Addition Paradox."

Bob: Really? Do tell?

Alice: Parfit posited a population he called "A" which had a moderately large population with large amounts of resources, giving them a very high level of utility per person. Then he added a second population, which was totally isolated from the other population. How they were isolated wasn't important, although Parfit suggested maybe they were on separate continents and can't sail across the ocean or something like that. These people don't have nearly as many resources per person as the other population, so each person's level of utility is lower (their lack of resources is the only reason they have lower utility). However, their lives are still just barely worth living. He called the two populations "A+."

Parfit asked if "A+" was a better world than "A." He thought it was, since the extra people were totally isolated from the original population they weren't hurting anyone over there by existing. And their lives were worth living. Follow me so far?

Bob: I guess I can see the point.

Alice: Next Parfit posited a population called "B," which was the same as A+. except that the two populations had merged together. Maybe they got better at sailing across the ocean, it doesn't really matter how. The people share their resources. The result is that everyone in the original population had their utility lowered, while everyone in the second had it raised.

Parfit asked if population "B" was better than "A+" and argued that it was because it had a greater level of equality and total utility.

Bob: I think I see where this is going. He's going to keep adding more people, isn't he?

Alice: Yep. He kept adding more and more people until he reached population "Z," a vast population where everyone had so few resources that their lives were barely worth living. This, he argued, was a paradox, because he argued that most people would believe that Z is far worse than A, but he had made a convincing argument that it was better.

Bob: Are you sure that sharing their resources like that would lower the standard of living for the original population? Wouldn't there be economies of scale and such that would allow them to provide more utility even with less resources per person?

Alice: Please don't fight the hypothetical. We're assuming that it would for the sake of the argument.

Now, Parfit argued that this argument led to the "Repugnant Conclusion," the idea that the best sort of world is one with a large population with lives barely worth living. That confers on people a duty to reproduce as often as possible, even if doing so would lower the quality of their and everyone else's lives.

He claimed that the reason his argument showed this was that he had conducted "mere addition." The populations in his paradox differed in no way other than their size. By merely adding more people he had made the world "better," even if the level of utility per person plummetted. He claimed that "For every population, A, with a high average level of utility there exists another, better population, B, with more people and a lower average level of utility."

Do you see the flaw in Parfit's argument?

Bob: No, and that kind of disturbs me. I have kids, and I agree that creating new people can add utility to the world. But it seems to me that it's also important to enhance the utility of the people who already exist.

Alice: That's right. Normal morality tells us that creating new people with lives worth living and enhancing the utility of people that already exist are both good things to use resources on. Our common sense tells us that we should spend resources on both those things. The disturbing thing about the Mere Addition Paradox is that it seems at first glance to indicate that that's not true, that we should only devote resources to creating more people with barely worthwhile lives. I don't agree with that, of course.

Bob: Neither do I. It seems to me that having a large number of worthwhile lives and a high average utility are both good things and that we should try to increase them both, not just maximize one.

Alice: You're right, of course. But don't say "having a high average utility." Say "use resources to increase the utility of people who already exist."

Bob: What's the difference? They're the same thing, aren't they?

Alice: Not quite. There are other ways to increase average utility than enhancing the utility of existing people. You could kill all the depressed people, for instance. Plus, if there was a world where everyone was tortured 24 hours a day, you could increase average utility by creating some new people who are only tortured 23 hours a day.

Bob: That's insane! Who could possibly be that literal-minded?

Alice: You'd be surprised. The point is, a better way to phrase it is "use resources to increase the utility of people who already exist," not "increase average utility." Of course, that still leaves some stuff out, like the fact that it's probably better to increase everyone's utility equally, rather than focus on just one person. But it doesn't lead to killing depressed people, or creating slightly less tortured people in a Hellworld.

Bob: Okay, so what I'm trying to say is that resources should be used to create people, and to improve people's lives. Also equality is good. And that none of these things should completely eclipse the other, they're each too valuable to maximize just one. So a society that increases all of those values should be considered more efficient at generating value than a society that just maximizes one value. Now that we're done getting our terminology straight, will you tell me what Parfit's mistake was?

Alice: Population "A" and population "A+" *differ in two ways, not one*. Think about it. Parfit is clear that the extra people in "A+" do not harm the existing people when they are added. That means they do not use any of the original population's resources. So how do they manage to live lives worth living? *How are they sustaining themselves*?

Bob: They must have their own resources. To use Parfit's example of continents separated by an ocean; each continent must have its own set of resources.

Alice: Exactly. So "A+" differs from "A" both in the size of its population, and the amount of resources it has access to. Parfit was not "merely adding" people to the population. He was also adding resources.

Bob: Aren't you the one who is fighting the hypothetical now?

Alice: I'm not fighting the hypothetical. Fighting the hypothetical consists of challenging the likelihood of the thought experiment happening, or trying to take another option than the ones presented. What I'm doing is challenging the logical coherence of the hypothetical. One of Parfit's unspoken premises is that you need *some* resources to live a life worth living, so by adding more worthwhile lives he's also implicitly adding resources. If he had just added some extra people to population A without giving them their own continent full of extra resources to live on then "A+" would be worse than "A."

Bob: So the Mere Addition Paradox doesn't confer on us a positive obligation to have as many children as possible, because the amount of resources we have access to doesn't automatically grow with them. I get that. But doesn't it imply that as soon as we get some more resources we have a duty to add some more people whose lives are barely worth living?

Alice: No. Adding lives barely worth living uses the extra resources more efficiently than leaving Parfit's second continent empty for all eternity. But, it's not the most efficient way. Not if you believe that creating new people and enhancing the utility of existing people are *both* important values.

Let's take population "A+" again. Now imagine that instead of having a population of people with lives barely worth living, the second continent is inhabited by a smaller population with the same very high percentage of resources and utility per person as the population of the first continent. Call it "A++. " Would you say "A++" was better than "A+?"

Bob: Sure, definitely.

Alice: How about a world where the two continents exist, but the second one was never inhabited? The people of the first continent then discover the second one and use its resources to improve their level of utility.

Bob: I'm less sure about that one, but I think it might be better than "A+."

Alice: So what Parfit actually proved was: "For every population, A, with a high average level of utility there exists another, better population, B, with more people, *access to more resources* and a lower average level of utility."

And I can add my own corollary to that: "For every population, B, there exists another, better population, C, that has the *same access to resources* as B, but a *smaller population and higher average utility.*"

Bob: Okay, I get it. But how does this relate to my cable TV sales pitch?

Alice: Well, my current situation, where I'm spending my money on normal things is analogous to Parfit's population "A." High utility, and very efficient conversion of resources into utility, but not as many resources. We're assuming, of course, that using resources to both create new people and improve the utility of existing people is more morally efficient than doing just one or the other.

The situation where the government gives me $50 to spend on Package A+ is analogous to Parfit's population A+. I have more resources and more utility. But the resources aren't being converted as efficiently as they could be.

The situation where I take the 50% discount and buy Package B is equivalent to Parfit's population B. It's a better situation than A+, but not the most efficient way to use the money.

The situation where I get the $50 from the government to spend on whatever I want is equivalent to *my* population C. A world with more access to resources than A, but more efficient conversion of resources to utility than A+ or B.

Bob: So what would a world where the government kept the money be analogous to?

Alice: A world where Parfit's second continent was never settled and remained uninhabited for all eternity, its resources never used by anyone.

Bob: I get it. So the Mere Addition Paradox doesn't prove what Parfit thought it did? We don't have any moral obligation to tile the universe with people whose lives are barely worth living?

Alice: Nope, we don't. It's more morally efficient to use a large percentage of our resources to enhance the lives of those who already exist.

Bob: This sure has been a fun conversation. Would you like to buy a cable package from me? We have some great deals.

Alice: NO!

SUMMARY:

My argument is that Parfit's Mere Addition Paradox doesn't prove what it seems to. The argument behind the Mere Addition Paradox is that you can make the world a better place by the "mere addition" of extra people, even if their lives are barely worth living. In other words : "For every population, A, with a high average level of utility there exists another, better population, B, with more people and a lower average level of utility." This supposedly leads to the Repugnant Conclusion, the belief that a world full of people whose lives are barely worth living is better than a world with a smaller population where the people lead extremely fulfilled and happy lives.

Parfit demonstrates this by moving from world A, consisting of a population full of people with lots of resources and high average utility, and moving to world A+. World A+ has an addition population of people who are isolated from the original population and not even aware of the other's existence. The extra people live lives just barely worth living. Parfit argues that A+ is a better world than A because everyone in it has lives worth living, and the additional people aren't hurting anyone by existing because they are isolated from the original population.

Parfit them moves from World A+ to World B, where the populations are merged and share resources. This lowers the standard of living for the original people and raises it for the newer people. Parfit argues that B must be better than A+, because it has higher total utility and equality. He then keeps adding people until he reaches Z, a world where everyones' lives are barely worth living and the population is vast. He argues that this is a paradox because most people would agree that Z is not a desirable world compared to A.

I argue that the Mere Addition Paradox is a flawed argument because it does not just add people, it also adds resources. The fact that the extra people in A+ do not harm the original people of A by existing indicates that their population must have a decent amount of resources to live on, even if it is not as many per person as the population of A. For this reason what the Mere Addition Paradox proves is not that you can make the world better by adding extra people, but rather that you can make it better by adding extra people and *resources to support them*. I use a series of choices about purchasing cable television packages to illustrate this in concrete terms.

I further argue for a theory of population ethics that values both using resources to create lives worth living, and using resources to enhance the utility of already existing people, and considers the best sort of world to be one where neither of these two values totally dominate the other. By this ethical standard A+ might be better than A because it has more people and resources, even if the average level of utility is lower. However, a world with the same amount of resources as A+, but a lower population and the same, or higher average utility as A is better than A+.

The main unsatisfying thing about my argument is that while it avoids the Repugnant Conclusion in most cases, it might still lead to it, or something close to it, in situations where creating new people and getting new resources are, as one commenter noted, a "package deal." In other words, a situation where it is impossible to obtain new resources without creating some new people whose utility levels are below average. However, even in this case, my argument holds that the best world of all is one where it would be possible to obtain the resources without creating new people, or creating a smaller amount of people with higher utility.

In other words, the Mere Addition Paradox does not prove that: "For every population, A, with a high average level of utility there exists another, better population, B, with more people and a lower average level of utility." Instead what the Mere Addition

Paradox seems to demonstrate is that: "For every population, A, with a high average level of utility there exists another, better population, B, with more people, *access to more resources* and a lower average level of utility." Furthermore, my own argument demonstrates that: "For every population, B, there exists another, better population, C, which has the *same access to resources* as B, but a *smaller population and higher average utility*."

Reply to Holden on The Singularity Institute

Holden Karnofsky of GiveWell has objected to the Singularity Institute (SI) as a target for optimal philanthropy. As someone who thinks that existential risk reduction is really important and also that the Singularity Institute is an important target of optimal philanthropy, I would like to explain why I disagree with Holden on these subjects. (I am also SI's Executive Director.)

Mostly, I'd like to explain my views to a broad audience. But I'd also like to explain my views to Holden himself. I value Holden's work, I enjoy interacting with him, and I think he is both intelligent *and* capable of changing his mind about Big Things like this. Hopefully Holden and I can continue to work through the arguments together, though of course we are both busy with many other things.

I appreciate the clarity and substance of Holden's objections, and I hope to reply in kind. I begin with an overview of some basic points that may be familiar to most *Less Wrong* veterans, and then I reply point-by-point to Holden's post. In the final section, I summarize my reply to Holden.

Holden raised *many* different issues, so unfortunately this post needed to be *long*. My apologies to Holden if I have misinterpreted him at any point.

Contents

Comments

I must be brief, so while reading this post I am sure many objections will leap to your mind. To encourage *constructive* discussion on this post, **each question (posted as a comment on this page) that follows the template described below will receive a reply from myself or another SI representative.**

Please word your question as clearly and succinctly as possible, and don't assume your readers will have read this post before reading your question (because: the

conversations here may be used as source material for a comprehensive FAQ).

Here's an example of how you could word the first paragraph of your question: "You claimed that [insert direct quote here], and also that [insert another direct quote here]. That seems to imply that [something something]. But that doesn't seem to take into account that [blah blah blah]. What do you think of that?"

If your question needs more explaining, leave the details to *subsequent* paragraphs in your comment. Please post multiple questions as multiple comments, so they can be voted upon and replied to individually. If you don't follow these rules, I can't guarantee SI will have time to give you a reply. (We probably *won't*.)

Why many people care greatly about existential risk reduction

Why do many people consider existential risk reduction to be humanity's most important task? I can't say it much better than Nick Bostrom does, so I'll just quote him:

> An existential risk is one that threatens the premature extinction of Earth-originating intelligent life or the permanent and drastic destruction of its potential for desirable future development. Although it is often difficult to assess the probability of existential risks, there are many reasons to suppose that the total such risk confronting humanity over the next few centuries is significant...
>
> Humanity has survived what we might call *natural existential risks* [asteroid impacts, gamma ray bursts, etc.] for hundreds of thousands of years; thus it is prima facie unlikely that any of them will do us in within the next hundred...
>
> In contrast, our species is introducing entirely new kinds of existential risk—threats we have no track record of surviving... In particular, most of the biggest existential risks seem to be linked to potential future technological breakthroughs that may radically expand our ability to manipulate the external world or our own biology. As our powers expand, so will the scale of their potential consequences—intended and unintended, positive and negative. For example, there appear to be significant existential risks in some of the advanced forms of biotechnology, molecular nanotechnology, and machine intelligence that might be developed in the decades ahead.
>
> What makes existential catastrophes especially bad is not that they would [cause] a precipitous drop in world population or average quality of life. Instead, their significance lies primarily in the fact that they would destroy the future... To calculate the loss associated with an existential catastrophe, we must consider how much value would come to exist in its absence. It turns out that the ultimate potential for Earth-originating intelligent life is literally astronomical.
>
> One gets a large number even if one confines one’s consideration to the potential for biological human beings living on Earth. If we suppose... that our planet will remain habitable for at least another billion years, and we assume that at least one billion people could live on it sustainably, then the potential exist for at least

> 10^{18} human lives. [The numbers get *way* bigger if you consider the expansion of posthuman civilization to the rest of the galaxy or the prospect of mind uploading.]
>
> Even if we use the most conservative of these estimates, which entirely ignores the possibility of space colonization and software minds, we find that the expected loss of an existential catastrophe is greater than the value of 10^{16} human lives...
>
> These considerations suggest that the loss in expected value resulting from an existential catastrophe is so enormous that the objective of reducing existential risks should be a dominant consideration whenever we act out of an impersonal concern for humankind as a whole.

I refer the reader to Bostrom's paper for further details and additional arguments, but neither his paper nor this post can answer every objection one might think of.

Nor can I summarize all the arguments and evidence related to estimating the severity and time horizon of every proposed existential risk. Even the 500+ pages of Oxford University Press' *Global Catastrophic Risks* can barely scratch the surface of this enormous topic. As explained in Intelligence Explosion: Evidence and Import, predicting long-term technological progress is *hard*. Thus, we must

> examine convergent outcomes that—like the evolution of eyes or the emergence of markets—can come about through any of several different paths and can gather momentum once they begin.

I'll say more about convergent outcomes later, but for now I'd just like to suggest that:

1. Many humans living today value both current and future people enough that *if* existential catastrophe is plausible this century, then upon reflection (e.g. after counteracting their unconscious, default scope insensitivity) they would conclude that reducing the risk of existential catastrophe is the most valuable thing they can do — whether through direct work or by donating to support direct work. It is to *these* people I appeal. (I also have much to say to people who e.g. *don't* care about future people, but it is too much to say here and now.)

2. As it turns out, we *do* have good reason to believe that existential catastrophe is plausible this century.

I don't have the space here to discuss the likelihood of different kinds of existential catastrophe that could plausibly occur this century (see *GCR* for more details), so instead I'll talk about just one of them: an AI catastrophe.

AI risk: the most important existential risk

There are two primary reasons I think AI is the most important existential risk:

Reason 1: **Mitigating AI risk could mitigate all other existential risks, but not vice-versa.** There is an asymmetry between AI risk and other existential risks. If we mitigate the risks from (say) synthetic biology and nanotechnology (without building

Friendly AI), this only means we have bought a few years or decades for ourselves before we must face yet another existential risk from powerful new technologies. But if we manage *AI risk* well enough (i.e. if we build a Friendly AI or "FAI"), we may be able to "permanently" (for several billion years) secure a desirable future. Machine superintelligence working in the service of humane goals could use its intelligence and resources to prevent all other existential catastrophes. (Eliezer: "I distinguish 'human', that which we *are*, from 'humane'—that which, being human, we *wish* we were.")

Reason 2: **AI is probably the first existential risk we must face** (given *my* evidence, only the tiniest fraction of which I can share in a blog post).

One reason AI may be the most urgent existential risk is that it's more likely for AI (compared to other sources of catastrophic risk) to be a full-blown *existential* catastrophe (as opposed to a merely *billions dead* catastrophe). Humans are smart and adaptable; we are already set up for a species-preserving number of humans to survive (e.g. in underground bunkers with stockpiled food, water, and medicine) major catastrophes from nuclear war, superviruses, supervolcano eruption, and many cases of asteroid impact or nanotechnological ecophagy.

Machine superintelligences, however, could intelligently seek out and neutralize humans which they (correctly) recognize as threats to the maximal realization of their goals. Humans are surprisingly easy to kill if an intelligent process is *trying* to do so. Cut off John's access to air for a few minutes, or cut off his water supply for a few days, or poke him with a sharp stick, and he dies. *Forever*. (Post-humans might shudder at this absurdity like we shudder at the idea that people used to die from their *teeth*.)

Why think AI is coming anytime soon? This is too complicated a topic to breach here. See Intelligence Explosion: Evidence and Import for a brief analysis of AI timelines. Or try The Uncertain Future, which outputs an estimated timeline for human-level AI based on *your* predictions of various technological developments. (SI is currently collaborating with the Future of Humanity Institute to write another paper on this subject.)

It's also important to mention that the case for caring about AI risk is less conjunctive that many seem to think, which I discuss in more detail here.

SI can purchase several kinds of AI risk reduction more efficiently than others can

The two organizations working most directly to reduce AI risk are the Singularity Institute and the Future of Humanity Institute (FHI). Luckily, these organizations complement each other well, as I pointed out back before I was running SI:

- FHI is part of Oxford, and thus can bring credibility to existential risk reduction. Resulting output: lots of peer-reviewed papers, books from OUP like *Global Catastrophic Risks*, conferences, media appearances, etc.
- SI is independent and is less constrained by conservatism or the university system. Resulting output: Very novel (and, to the mainstream, "weird") research

> on Friendly AI, and the ability to do unusual things that are nevertheless quite effective at finding/creating lots of new people interested in rationality and existential risk reduction: (1) The Sequences, the best tool I know for creating aspiring rationalists, (2) *Harry Potter and the Methods of Rationality*, a surprisingly successful tool for grabbing the attention of mathematicians and computer scientists around the world, and (3) the Singularity Summit, a mainstream-aimed conference that brings in people who end up making significant contributions to the movement — e.g. Tomer Kagan (an SI donor and board member) and David Chalmers (author of The Singularity: A Philosophical Analysis and The Singularity: A Reply).

A few weeks later, Nick Bostrom (Director of FHI) said the same things (as far as I know, *without* having read *my* comment):

> I think there is a sense that both organizations are synergistic. If one were about to go under... that would probably be the one [to donate to]. If both were doing well... different people will have different opinions. We work quite closely with the folks from [the Singularity Institute]...
>
> There is an advantage to having one academic platform and one outside academia. There are different things these types of organizations give us. If you wanna get academics to pay more attention to this, to get postdocs to work on this, that's much easier to do within academia; also to get the ear of policy-makers and media... On the other hand, for [SI] there might be things that are easier for them to do. More flexibility, they're not embedded in a big bureaucracy. So they can more easily hire people with non-standard backgrounds... and also more grass-roots stuff like *Less Wrong*...

FHI is, despite its small size, a highly productive philosophy department. More importantly, FHI has focused its research work on AI risk issues for the past 9 months, and plans to continue on that path for at least another 12 months. This is important work that should be supported. (Note that FHI recently hired SI research associate Daniel Dewey.)

SI lacks FHI's publishing productivity and its university credibility, but as an organization SI is improving quickly, and it can seize many opportunities for AI risk reduction that FHI is not well-positioned to seize. (*New* organizations will *also* tend to be less capable of seizing these opportunities than SI, due to the financial and human capital already concentrated at SI and FHI.)

Here are some examples of projects that SI is probably better able to carry out than FHI, given its greater flexibility (and assuming sufficient funding):

- A scholarly AI risk wiki written and maintained by dozens of part-time researchers from around the world.
- Reaching young math/compsci talent in unusual ways, e.g. *HPMoR*.
- Writing *Open Problems in Friendly AI* (Eliezer has spent far more time working on the mathy sub-problems of FAI than anyone else).

My replies to Holden, point by point

Holden's post makes so *many* claims that I'll just have to work through his post from beginning to end, and then summarize where I think we stand at the end.

GiveWell Labs

Holden opened "Thoughts on the Singularity Institute" by noting that SI was previously outside Givewell's scope, since GiveWell was focused on specific domains like poverty reduction. With the launch of GiveWell Labs, GiveWell is now open to evaluating *any* giving opportunity, including SI.

I admire this move. I'm sure people have been bugging GiveWell to do this for a long time, but almost none of those people appreciate how hard it is to launch broad new initiatives like this with the limited budget of an organization like Givewell or the Singularity Institute. Most of them *also* do not understand how much work is required to write something like "Thoughts on the Singularity Institute", "Reply to Holden on Tool AI", or *this* post.

Three possible outcomes

Next, Holden wrote:

> [I hope] that one of these three things (or some combination) will happen:
>
> 1. New arguments are raised that cause me to change my mind and recognize SI as an outstanding giving opportunity. If this happens I will likely attempt to raise more money for SI (most likely by discussing it with other GiveWell staff and collectively considering a GiveWell Labs recommendation).
> 2. SI concedes that my objections are valid and increases its determination to address them. A few years from now, SI is a better organization and more effective in its mission.
> 3. SI can't or won't make changes, and SI's supporters feel my objections are valid, so SI loses some support, freeing up resources for other approaches to doing good.

As explained at the top of Holden's post, I had already conceded that many of Holden's objections (especially concerning past organizational competence) are valid, and had been working to address them, even *before* Holden's post was published. So outcome #2 is already true in part.

I hope for outcome #1, too, but I don't expect Holden to change his opinion overnight. There are too many possible objections to which Holden has not yet heard a good response. But hopefully this post and its comment threads will successfully address *some* of Holden's (and others') objections.

Outcome #3 is unlikely since SI is *already* making changes, though of course it's possible we will be unable to raise sufficient funding for SI *despite* making these changes, or even *because of* our efforts to make these changes. (Improving general

organizational effectiveness is important but it costs money and is not exciting to donors.)

SI's mission is more important than SI as an organization

Holden said:

> whatever happens as a result of my post will be positive for SI's mission, whether or not it is positive for SI as an organization. I believe that most of SI's supporters and advocates care more about the former than about the latter, and that this attitude is far too rare in the nonprofit world.

Clearly, SI's mission is more important than SI as an organization. If somebody launches an organization more effective (at AI risk reduction) than SI but just as flexible, then SI should probably fold itself and try to move its donor base, support community, and the best of its human capital to that new organization.

That said, it's probably easier to reform SI into a more effective organization than it is to launch a new one, since SI has successfully concentrated lots of attention, donor support, and human capital. Also, SI has learned many lessons about how to run a very tricky kind of organization. AI risk reduction is a mission that (1) is beyond most people's time horizons for caring, (2) is hard to understand and visualize, (3) pattern-matches to science fiction and apocalyptic religion, (4) suffers under complicated and *necessarily* uncertain strategic considerations (compare to the simplicity of bed nets), (5) has a very small pool of people from which to recruit researchers, etc. SI has lots of experience with these issues; experience that probably takes a long time and lots of money to acquire.

(On the other hand, SI has also concentrated some bad reputation which a new organization could launch without. But I still think the weight of the arguments is in favor of reforming SI.)

SI's arguments need to be clearer

Holden:

> I do not believe that [my objections to SI's apparent views] constitute a sharp/tight case for the idea that SI's work has low/negative value; I believe, instead, that SI's own arguments are too vague for such a rebuttal to be possible. There are many possible responses to my objections, but SI's public arguments (and the private arguments) do not make clear which possible response (if any) SI would choose to take up and defend. Hopefully the dialogue following this post will clarify what SI believes and why.

I agree that SI's arguments are often vague. For example, Chris Hallquist reported:

> I've been trying to write something about Eliezer's debate with Robin Hanson, but the problem I keep running up against is that Eliezer's points are not clearly

> articulated at all. Even making my best educated guesses about what's supposed to go in the gaps in his arguments, I still ended up with very little.

I know the feeling! That's why I've tried to write as many clarifying documents as I can, including the Singularity FAQ, Intelligence Explosion: Evidence and Import, The Singularity and Machine Ethics, *Facing the Singularity*, So You Want to Save the World, and How to Purchase AI Risk Reduction.

Unfortunately, it takes lots of resources to write up hundreds of arguments and responses to objections in clear and precise language, and we're working on it. (For comparison, Nick Bostrom's forthcoming book on machine superintelligence will barely scratch the surface of the things SI and FHI researchers have worked out in conversation, and it will probably take him 2+ years to write in total, and Bostrom is *already* an unusually prolific writer.) Hopefully SI's responses to Holden's post have helped to clarify our positions already.

Holden's objection #1 punts to objection #2

The first objection on Holden's numbered list was:

> it seems to me that any AGI that was set to maximize a "Friendly" utility function would be extraordinarily dangerous.

I'm glad Holden agrees with us that successful Friendly AI is *very hard*. SI has spent much of its effort trying to show people that the first 20 solutions they come up with all fail. See: AI as a Positive and Negative Factor in Global Risk, The Singularity and Machine Ethics, Complex Value Systems are Required to Realize Valuable Futures, etc. Holden mentions the standard SI worry about the hidden complexity of wishes, and the one about a friendly utility function still causing havoc because the AI's priors are wrong (problem 3.6 from my list of open problems in AI risk research).

There are reasons to think FAI is harder *still*. What if we get the utility function right and we get the priors right but the AI's values *change* for the worse when it updates its ontology? What if the smartest, most careful, most insanely safety-conscious AI researchers humanity can produce *just aren't smart enough* to solve the problem? What if *no* humans are altruistic enough to choose to build FAI over an AI that will make them king of the universe? What if the idea of FAI is incoherent? (The human brain is an existence proof for the possibility of general intelligence, but we have *no* existence proof for the possibility of a decision theoretic agent which stably optimizes the world according to a set of preferences over states of affairs.)

So, yeah. Friendly AI is *hard*. But as I said elsewhere:

> The point is that *not* trying as hard as you can to build Friendly AI is even *worse*, because then you *almost certainly get uFAI*. At least by *trying* to build FAI, we've got some chance of winning.

So Holden's objection #1 objection really just punts to objection #2, about tool-AGI, as the last paragraph in this section of Holden's post seems to indicate:

> So far, all I have argued is that the development of "Friendliness" theory can achieve at best only a limited reduction in the probability of an unfavorable

> outcome. However, as I argue in the next section, I believe there is at least one concept - the "tool-agent" distinction - that has more potential to reduce risks, and that SI appears to ignore this concept entirely.

So if Holden's objection #2 doesn't work, then objection #1 ends up reducing to "the development of Friendliness theory can achieve at best a reduction in AI risk," which is what SI has been saying all along.

Tool AI

Holden's second numbered objection was:

> SI appears to neglect the potentially important distinction between "tool" and "agent" AI.

Eliezer wrote a whole post about this here. To sum up:

(1) **Whether you're working with Tool AI or Agent AI, you need the "Friendly AI" domain experts that SI is trying to recruit**:

> A "Friendly AI programmer" is somebody who specializes in seeing the correspondence of mathematical structures to What Happens in the Real World. It's somebody who looks at Hutter's specification of AIXI and reads the actual equations - actually stares at the Greek symbols and not just the accompanying English text - and sees, "Oh, this AI will try to gain control of its reward channel," as well as numerous subtler issues like, "This AI presumes a Cartesian boundary separating itself from the environment; it may drop an anvil on its own head." Similarly, working on TDT means e.g. looking at a mathematical specification of decision theory, and seeing "Oh, this is vulnerable to blackmail" and coming up with a mathematical counter-specification of an AI that isn't so vulnerable to blackmail.
>
> Holden's post seems to imply that if you're building a non-self-modifying planning Oracle (aka 'tool AI') rather than an acting-in-the-world agent, you don't need a Friendly AI programmer because FAI programmers only work on agents. But this isn't how the engineering skills are split up. Inside the AI, whether an agent AI or a planning Oracle, there would be similar AGI-challenges like "build a predictive model of the world", and similar FAI-conjugates of those challenges like finding the 'user' inside an AI-created model of the universe. The insides would look a lot more similar than the outsides. An analogy would be supposing that a machine learning professional who does sales optimization for an orange company couldn't possibly do sales optimization for a banana company, because their skills must be about oranges rather than bananas.

(2) **Tool AI isn't that much safer than Agent AI, because Tool AIs have lots of hidden "gotchas" that cause havoc, too**. (See Eliezer's post for examples.)

These points illustrate something else Eliezer wrote:

> What the human species needs from an x-risk perspective is experts on This Whole Damn Problem [of AI risk], who will acquire whatever skills are needed to

> that end. The Singularity Institute exists to host such people and enable their research—once we have enough funding to find and recruit them.

Indeed. We need places for experts who specialize in seeing the consequences of mathematical objects for things humans value (e.g. the Singularity Institute) just like we need places for experts on efficient charity (e.g. Givewell).

Anyway, it's worth pointing out that Holden did *not* make the common (and mistaken) argument that "We should just build Tool AIs instead of Agent AIs and then we'll be fine." This is wrong for many reasons, but one obvious point is that there are incentives to build Agent AIs (because they're powerful), so even if the first 6 teams are careful enough to build only Tool AIs, the 7th team could still build Agent AI and destroy the world.

Instead, Holden pointed out that *you could use Tool AI to increase your chances of successfully building agenty FAI*:

> if developing "Friendly AI" is what we seek, a tool-AGI could likely be helpful enough in thinking through this problem as to render any previous work on "Friendliness theory" moot. Among other things, a tool-AGI would allow transparent views into the AGI's reasoning and predictions without any reason to fear being purposefully misled, and would facilitate safe experimental testing of any utility function that one wished to eventually plug into an "agent."

After reading Eliezer's reply, however, you can probably guess my replies to this paragraph:

1. Tool AI isn't as safe as Holden thinks.
2. But yeah, a Friendly AI team may very well use "Tool AI" to aid Friendliness research if it can figure out a safe way to do that. This doesn't obviate the need for Friendly AI researchers; it's *part* of their research toolbox.

So Holden's Objection #2 doesn't work, which (as explained earlier) means that his Objection #1 (as stated) doesn't work either.

SI's mission assumes a scenario that is far *less* conjunctive than it initially appears.

Holden's objection #3 is:

> SI's envisioned scenario is far more specific and conjunctive than it appears at first glance, and I believe this scenario to be highly unlikely.

His main concern here seemed to be that technological developments and other factors would render earlier FAI work irrelevant. But Eliezer's clarifications about what we mean by "FAI team" render this objection moot, at least as it is currently stated. The *purpose* of an FAI team is not to blindly develop one particular approach to Friendly AI without checking to see whether this work will be obsoleted by future developments. Instead, the purpose of an FAI team is to develop highly specialized expertise on, among other things, which kinds of research are more and less likely to be relevant given future developments.

Holden's confusion about what SI means by "FAI team" is common and understandable, and it is one reason that SI's mission assumes a scenario that is far *less* conjunctive than it appears to many. We aren't saying we need an FAI team because we know lots of specific things about how AGI will be built 30 years from now. We're saying you need experts on "the consequences of mathematical objects for things humans value" (an FAI team) because AGIs are mathematical objects and will have big consequences. That's pretty disjunctive.

Similarly, many people think SI's mission is predicated on hard takeoff. After all, we call ourselves the "Singularity Institute," Eliezer has spent a lot of time arguing for hard takeoff, and our current research summary frames AI risk in terms of recursive self-improvement.

But the case for AI as a global risk, and thus the need for dedicated experts on AI risk and "the consequences of mathematical objects for things humans value", isn't predicated on hard takeoff. Instead, it looks something like this:

(1) **Eventually, most tasks are performed by machine intelligences.**

The improved flexibility, copyability, and modifiability of machine intelligences make them economically dominant even without other advantages (Brynjolfsson & McAfee 2011; Hanson 2008). In addition, there is plenty of room "above" the human brain in terms of hardware and software for general intelligence (Muehlhauser & Salamon 2012; Sotala 2012; Kurzweil 2005).

(2) **Machine intelligences don't necessarily do things we like.**

We don't necessarily control AIs, since advanced intelligences may be inherently goal-oriented (Omohundro 2007), and even if we build advanced "Tool AIs," these aren't necessarily safe either (Yudkowsky 2012) and there will be significant economic incentives to transform them into autonomous agents (Brynjolfsson & McAfee 2011). We don't value most possible futures, but it's *very* hard to get an autonomous AI to do exactly what you want (Yudkowsky 2008, 2011; Muehlhauser & Helm 2012; Arkin 2009).

(3) **There are things we can do to increase the probability that machine intelligences do things we like.**

Further research can clarify (1) the nature and severity of the risk, (2) how to engineer goal-oriented systems safely, (3) how to increase safety with differential technological development, (4) how to limit and control machine intelligences (Armstrong et al. 2012; Yampolskiy 2012), (5) solutions to AI development coordination problems, and more.

(4) **We should do those things now.**

People aren't doing much about these issues now. We could wait until we understand better (e.g.) what kind of AI is likely, but: (1) it might take a long time to resolve the core issues, including difficult technical subproblems that require time-consuming mathematical breakthroughs, (2) incentives may be badly aligned (e.g. there seem to be strong economic incentives to build AI, but not to take into account social and global risks for AI), (3) AI may not be that far away (Muehlhauser & Salamon 2012), and (4) the transition to machine dominance may be surprisingly rapid due to (e.g.) intelligence explosion (Chalmers 2010, 2012; Muehlhauser & Salamon 2012) or computing overhang.

What do I mean by "computing overhang"? We may get the hardware needed for AI long before we get the software, such that once software for general intelligence is figured out, there is tons of computing hardware sitting around for running AIs (a "computing overhang"). Thus we could switch from a world with one autonomous AI to a world with 10 billion autonomous AIs at the speed of copying software, and thereby transition rapidly from human dominance to AI dominance even without an intelligence explosion. (This is one of the many, many things we haven't yet written up in detail up due to lack of resources.)

(This broad argument is greatly compressed from a paper outline developed by Paul Christiano, Carl Shulman, Nick Beckstead, and myself. We'd love to write the paper at some point, but haven't had the resources to do so. The fuller version of this argument is of course more detailed.)

SI's public argumentation

Next, Holden turned to the topic of SI's organizational effectiveness:

> when evaluating a group such as SI, I can't avoid placing a heavy weight on (my read on) the general competence, capability and "intangibles" of the people and organization, because SI's mission is not about repeating activities that have worked in the past...
>
> There are several reasons that I currently have a negative impression of SI's general competence, capability and "intangibles."

The first reason Holden gave for his negative impression of SI is:

> SI has produced enormous quantities of public argumentation... Yet I have never seen a clear response to any of the three basic objections I listed in the previous section. One of SI's major goals is to raise awareness of AI-related risks; given this, the fact that it has not advanced clear/concise/compelling arguments speaks, in my view, to its general competence.

I agree *in part*. Here's what I think:

- SI *hasn't* made its arguments as clear, concise, and compelling as I would like. We're working on that. It takes time, money, and people who are (1) smart and capable enough to do AI risk research work and yet somehow (2) willing to work for non-profit salaries and (3) willing to *not* advance their careers like they would if they chose instead to work at a university.
- There are a *huge* number of possible objections to SI's arguments, and we haven't had the resources to write up clear and compelling replies to *all* of them. (See Chalmers 2012 for quick rebuttals to many objections to intelligence explosion, but what he covers in that paper barely scratches the surface.) As Eliezer wrote, Holden's complaint that SI hasn't addressed *his* particular objections "seems to lack perspective on how *many* different things various people see as the *one obvious solution* to Friendly AI. Tool AI wasn't the obvious solution to John McCarthy, I.J. Good, or Marvin Minsky. Today's leading AI textbook, *Artificial Intelligence: A Modern Approach*... discusses Friendly AI and AI risk for 3.5 pages but doesn't mention tool AI as an obvious solution. For Ray Kurzweil, the obvious solution is merging humans and AIs. For Jurgen Schmidhuber, the obvious solution is AIs that value a certain complicated definition of complexity in their sensory inputs. Ben Goertzel, J. Storrs Hall, and

> Bill Hibbard, among others, have all written about how silly Singinst is to pursue Friendly AI when the solution is obviously X, for various different X. Among current leading people working on serious AGI programs labeled as such, neither Demis Hassabis (VC-funded to the tune of several million dollars) nor Moshe Looks (head of AGI research at Google) nor Henry Markram (Blue Brain at IBM) think that the obvious answer is Tool AI. Vernor Vinge, Isaac Asimov, and any number of other SF writers with technical backgrounds who spent serious time thinking about these issues didn't converge on that solution."

- SI *has* done a decent job of raising awareness of AI risk, I think. Writing The Sequences and *HPMoR* have (indirectly) raised more awareness for AI risk that one can normally expect from, say, writing a bunch of clear and precise academic papers about a subject. (At least, it seems that way to me.)

SI's endorsements

The second reason Holden gave for his negative impression of SI is "a lack of impressive endorsements." This one is generally true, despite the three "celebrity endorsements" on our new donate page. More impressive than these is the fact that, as Eliezer mentioned, the latest edition of the leading AI textbook spend several pages talking about AI risk and Friendly AI, and discusses the work of SI-associated researchers like Eliezer Yudkowsky and Steve Omohundro while *completely ignoring* the existence of the older, more prestigious, and vastly larger mainstream academic field of "machine ethics."

Why don't we have impressive endorsements? To my knowledge, SI hasn't tried very hard to get them. That's another thing we're in the process of changing.

SI and feedback loops

The third reason Holden gave for his negative impression of SI is:

> SI seems to have passed up opportunities to test itself and its own rationality by e.g. aiming for objectively impressive accomplishments... Pursuing more impressive endorsements and developing benign but objectively recognizable innovations (particularly commercially viable ones) are two possible ways to impose more demanding feedback loops.

We have thought many times about commercially viable innovations we could develop, but these would generally be large distractions from the work of our core mission. (The Center for Applied Rationality, in contrast, has *many* opportunities to develop commercially viable innovations in line with its core mission.)

Still, I *do* think it's important for the Singularity Institute to test itself with tight feedback loops wherever feasible. This is particularly difficult to do for a research organization doing a philosophy of long-term forecasting (30 years is not a "tight" feedback loop in the slightest), but that's what FHI does and they have more "objectively impressive" (that is, "externally proclaimed") accomplishments: lots of peer-reviewed publications, some major awards for its top researcher Nick Bostrom, etc.

SI and rationality

Holden's fourth concern about SI is that it is overconfident about the level of its own rationality, and that this seems to show itself in (e.g.) "insufficient self-skepticism" and "being too selective (in terms of looking for people who share its preconceptions) when determining whom to hire and whose feedback to take seriously."

What would provide good evidence of rationality? Holden explains:

> I endorse Eliezer Yudkowsky's statement, "Be careful … any time you find yourself defining the [rationalist] as someone other than the agent who is currently smiling from on top of a giant heap of utility." To me, the best evidence of superior general rationality (or of insight into it) would be objectively impressive achievements (successful commercial ventures, highly prestigious awards, clear innovations, etc.) and/or accumulation of wealth and power. As mentioned above, SI staff/supporters/advocates do not seem particularly impressive on these fronts...

Unfortunately, this seems to misunderstand the term "rationality" as it is meant in cognitive science. As I explained elsewhere:

> Like intelligence and money, rationality is only a ceteris paribus predictor of success.
>
> So while it's empirically true (Stanovich 2010) that rationality is a predictor of life success, it's a weak one. (At least, it's a weak predictor of success at the levels of human rationality we are capable of training *today*.) If you want to more reliably achieve life success, I recommend inheriting a billion dollars or, failing that, being born+raised to have an excellent work ethic and low akrasia.
>
> The reason you should "be careful… any time you find yourself defining the [rationalist] as someone other than the agent who is currently smiling from on top of a giant heap of utility" is because you should "never end up envying someone else's mere *choices*." You are still allowed to envy their resources, intelligence, work ethic, mastery over akrasia, and other predictors of success.

But I don't mean to dodge the key issue. I think SIers *are* generally more rational than most people (and so are LWers, it seems), but I think SIers *have* often overestimated their own rationality, myself included. Certainly, I think SI's leaders have been pretty irrational about *organizational development* at many times in the past. In internal communications about why SI should help launch CFAR, one reason on my list has been: "We need to improve our own rationality, and figure out how to create better rationalists than exist today."

SI's goals and activities

Holden's fifth concern about SI is the apparent disconnect between SI's goals and its activities:

> SI seeks to build FAI and/or to develop and promote "Friendliness theory" that can be useful to others in building FAI. Yet it seems that most of its time goes to

> activities other than developing AI or theory.

This one is pretty easy to answer. We've focused mostly on movement-building rather than direct research because, until very recently, there wasn't enough community interest or funding to seriously begin to form an FAI team. To do that you need (1) at least a few million dollars a year, and (2) enough smart, altruistic people to care about AI risk that there exist some potential superhero mathematicians for the FAI team. And to get those two things, you've got to do *mostly* movement-building, e.g. *Less Wrong*, *HPMoR*, the Singularity Summit, etc.

Theft

And of course, Holden is (rightly) concerned about the 2009 theft of $118,000 from SI, and the lack of public statements from SI on the matter.

Briefly:

- Two former employees stole $118,000 from SI. Earlier this year we finally won stipulated judgments against both individuals, forcing them to pay back the full amounts they stole. We have already recovered several thousand dollars of this.
- We do have much better financial controls now. We consolidated our accounts so there are fewer accounts to watch, and at least three staff members check them regularly, as does our treasurer, who is *not* an SI staff member or board member.

Pascal's Mugging

In another section, Holden wrote:

> A common argument that SI supporters raise with me is along the lines of, "Even if SI's arguments are weak and its staff isn't as capable as one would like to see, their goal is so important that they would be a good investment even at a tiny probability of success."
>
> I believe this argument to be a form of Pascal's Mugging and I have outlined the reasons I believe it to be invalid...

Some problems with Holden's two posts on this subject will be explained in a forthcoming post by Steven Kaas. But as Holden notes, some SI principals like Eliezer don't use "small probability of large impact" arguments, anyway. We in fact argue that the probability of a large impact is *not* tiny.

Summary of my reply to Holden

Now that I have addressed so many details, let us return to the big picture. My summarized reply to Holden goes like this:

Holden's first two objections can be summarized as arguing that developing the Friendly AI approach is more dangerous than developing non-agent "Tool" AI. Eliezer's post points out that "Friendly AI" domain experts are what you need whether you're working with Tool AI or Agent AI, because (1) both of these approaches require FAI experts (experts in seeing the consequences of mathematical objects for what humans value), and because (2) Tool AI isn't necessarily much safer than Agent AI, because Tool AIs have lots of hidden gotchas, too. Thus, "What the human species needs from an x-risk perspective is experts on This Whole Damn Problem [of AI risk], who will acquire whatever skills are needed to that end. The Singularity Institute exists to host such people and enable their research — once we have enough funding to find and recruit them."

Holden's third objection was that the argument behind SI's mission is more conjunctive than it seems. I replied that the argument behind SI's mission is actually *less* conjunctive than it often seems, because an "FAI team" works on a broader set of problems than Holden had realized, and because the case for AI risk is more disjunctive than many people realize. These confusions are understandable, however, and they probably are a result of insufficient clear argumentative writing from SI on these matters — a problem we am trying to fix with several recent and forthcoming papers and other communications (like this one).

Holden's next objection concerned SI as an organization: "SI has, or has had, multiple properties that I associate with ineffective organizations." I acknowledged these problems before Holden published his post, and have since outlined the many improvements we've made to organizational effectiveness since I was made Executive Director. I addressed several of Holden's specific worries here.

Finally, Holden recommended giving to a donor-advised fund rather than to SI:

> I don't think that "Cause X is the one I care about and Organization Y is the only one working on it" to be a good reason to support Organization Y. For donors determined to donate within this cause, I encourage you to consider donating to a donor-advised fund while making it clear that you intend to grant out the funds to existential-risk-reduction-related organizations in the future....
>
> For one who accepts my arguments about SI, I believe withholding funds in this way is likely to be better for SI's mission than donating to SI

By now I've called into question most of Holden's arguments about SI, but I will still address the issue of donating to SI vs. donating to a donor-advised fund.

First: Which public charity would administer the donor-advised fund? Remember also that in the U.S., the administering charity need not spend from the donor-advised fund as the donor wishes, though they often do.

Second: As I said earlier,

> it's probably easier to reform SI into a more effective organization than it is to launch a new one, since SI has successfully concentrated lots of attention, donor support, and human capital. Also, SI has learned many lessons about how to run a very tricky kind of organization. AI risk reduction is a mission that (1) is beyond most people's time horizons for caring, (2) is hard to understand and visualize, (3) pattern-matches to science fiction and apocalyptic religion, (4) suffers under complicated and *necessarily* uncertain strategic considerations (compare to the simplicity of bed nets), (5) has a very small pool of people from which to recruit

> researchers, etc. SI has lots of experience with these issues; experience that probably takes a long time and lots of money to acquire.

The case for funding improvements and growth at SI (as opposed to starving SI as Holden suggests) is bolstered by the fact that SI's productivity and effectiveness have been improving rapidly of late, and many other improvements (and exciting projects) are on our "to-do" list if we can raise sufficient funding to implement them.

Holden even seems to share some of this optimism:

> Luke's... recognition of the problems I raise... increases my estimate of the likelihood that SI will work to address them...
>
> I'm aware that SI has relatively new leadership that is attempting to address the issues behind some of my complaints. I have a generally positive impression of the new leadership; I believe the Executive Director and Development Director, in particular, to represent a step forward in terms of being interested in transparency and in testing their own general rationality. So I will not be surprised if there is some improvement in the coming years...

Conclusion

For brevity's sake I have skipped many important details. I may also have misinterpreted Holden somewhere. And surely, Holden and other readers have follow-up questions and objections. This is not the end of the conversation; it is closer to the beginning. I invite you to leave your comments, preferably in accordance with these guidelines (for improved discussion clarity).

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.

A Marriage Ceremony for Aspiring Rationalists

Recently, LWers Will Ryan and Divia Melwani (now Will and Divia Eden) were married, with Eliezer Yudkowsky officiating.

I've been to 40+ weddings in my lifetime, and this was my favorite ceremony yet. Here is the video, and below is the transcript of Eliezer's... what's it called? "Blessing"?

> Dearly beloved, we are gathered here upon this day, to bear witness to William Ryan and Divia Melwani, as they bind themselves together in marriage, becoming William and Divia Eden, from this day endeavoring to live their lives as one. If any person can show just cause why these two should not be joined, let them speak now, or forever hold their peace.
>
> The institution of marriage is as old as *Homo sapiens*. Donald Brown lists it among the human universals, the parts of culture which are found in almost every tribe that has been studied by anthropologists, alongside such other universals as dancing, storytelling, jealousy, or language. Though we give it a single name, marriage takes many forms.
>
> In some tribes a man may wed more than one woman. In 0.5% of hunter-gatherer tribes studied, a woman may wed more than one man. In civilized parts of the modern world, men may marry men, or women marry women. A hundred years ago, in what was then considered civilization, marriage was a cruel necessity if you wanted to have a public relationship with anyone. There was only one approved option for anyone who didn't want to live alone - marry a single person of the opposite sex and stay together for 70 years or until one of you died.
>
> But in this day, and within this community, marriage necessarily takes on a different meaning. 'Until death do you part' is a different concept if you suspect that indefinite lifespan extension may be invented sometime in the next few decades. Once, getting married at age 20 meant you were probably a quarter of the way through your life. In this day, and in this community, you know that you might actually be getting married at zero point zero zero zero and some more zeroes one percent of the way through your life. Our community contains many people in long-term relationships who are not married and are not waiting around to get married.
>
> Even among those who marry, not every marriage has the same meaning. Some may not be planning to stay together until the stars go out - just enjoy the marriage for however long it lasts. And though marriage is no longer mandatory, the government of this country, in its finite wisdom, has decreed legal benefits for marriage which some of us may not wish to deny ourselves, even if we haven't yet found a perfect romance out of storybooks, even if we might not want a perfect romance out of storybooks.
>
> Marriage is no longer something that everyone has to do, and there isn’t just one kind of marriage, or one meaning of marriage. But at least so far as I can tell from the outside, Will and Divia seem to have a perfect romance, pretty much. And

while romances like that exist, the ancient institution of marriage will continue into the future, I think.

There are stars in the sky above us, even now. Even on a cloudless day you can't see them with your naked eyes, but the right camera would capture them. There is light shining upon this ceremony which is far older than eight and a half minutes. Standing as we do in the light of eternity, it may seem impossible to swear any true promise upon the future, when there are no perfect blessings called down upon a marriage to ensure its success, but only the mortal wills of human beings to guard it.

And yet there are still some people who are just so adorable together that you look at them and say, "Yeah, they should go for it." I can think of at least three couples like that, though, aside from Will and Divia, I'm not going to name any names. Elizabeth Moon once wrote that courage is inherent in all living things; it is the quality that keeps them alive; it is courage that splits the acorn and sends the rootlet down into soil to search for sustenance. This is not literally true. Acorns don't have brains so they can't experience courage. But I would still praise the idea of courage as a quality that powers all of human life - the daring to do things that you don't know for certain will work, acting under conditions of uncertainty. Even in an unstable world, not knowing how society might change, how you yourself might change, whether life as we know it will still exist at all in 30 years - even though nobody can foresee a thousand years into the future, even if everything goes right - even so, two or more people can still have sufficient confidence, and hope and courage, to try and build something greater out of the union of their lives. Because why not? If someone is already fortunate enough to have a relationship that once would have been called a marriage blessed by Heaven, why should they receive any less joy, or receive it any later, than they would have had in bygone times? How sad would it be to delay a hundred years and then find out that it would have worked after all?

And one element of marriage which has not changed is the endeavor to raise children. Not every marriage may desire children, but among those who do desire children, a marriage promises those children a stable home, a lasting family, and at least two people who jointly accept full responsibility for every child. For myself - seeing the meaning of this wedding through my own eyes - I would affirm and support above all else the wholehearted decision of Will and Divia to forge a more lasting bond because they both wished to bring a new child into the world. That responsibility is owed to any endeavor of creating a new sentient life. That meaning of marriage has not changed.

A final question is what marriage now means to the community that bears witness. William and Divia have chosen to bind their lives together. As it is not our place to deny that, neither is it within our power to permit it. There is no higher authority whose blessings must be sought, and we can't wish them good luck because there's no such thing in the universe as luck. We could say, "We wish you happy lives as the result of your own decisions!", but wishing doesn't make anything happen. And yet for as long as marriage has existed among the human species, it has been a ceremony performed within sight of the tribe. For tens of thousands of years before humans imagined that the heavens had authority, the tribe has borne witness to marriages. Of you all, then, I will ask that you promise to respect this marriage, and not come between Will and Divia in any way, should you find that possibility within your power; and those of you present who bear them other friendships may vow such other support as lies within your hearts. And

let it be known to all the world that what is begun here today, is done brightly, and without shame.

[Retracted] Simpson's paradox strikes again: there is no great stagnation?

ETA: The table linked by Landsburg has been called into serious question by Evan Soltas [H.T. CronoDAS]. I edited the post to leave only the table to provide context for the comment discussion of its status.

Economist Steve Landsburg has a post [H.T. David Henderson] about the supposed stagnation of median wages in the United States in recent decades. In the linked table median wages have risen for:

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,

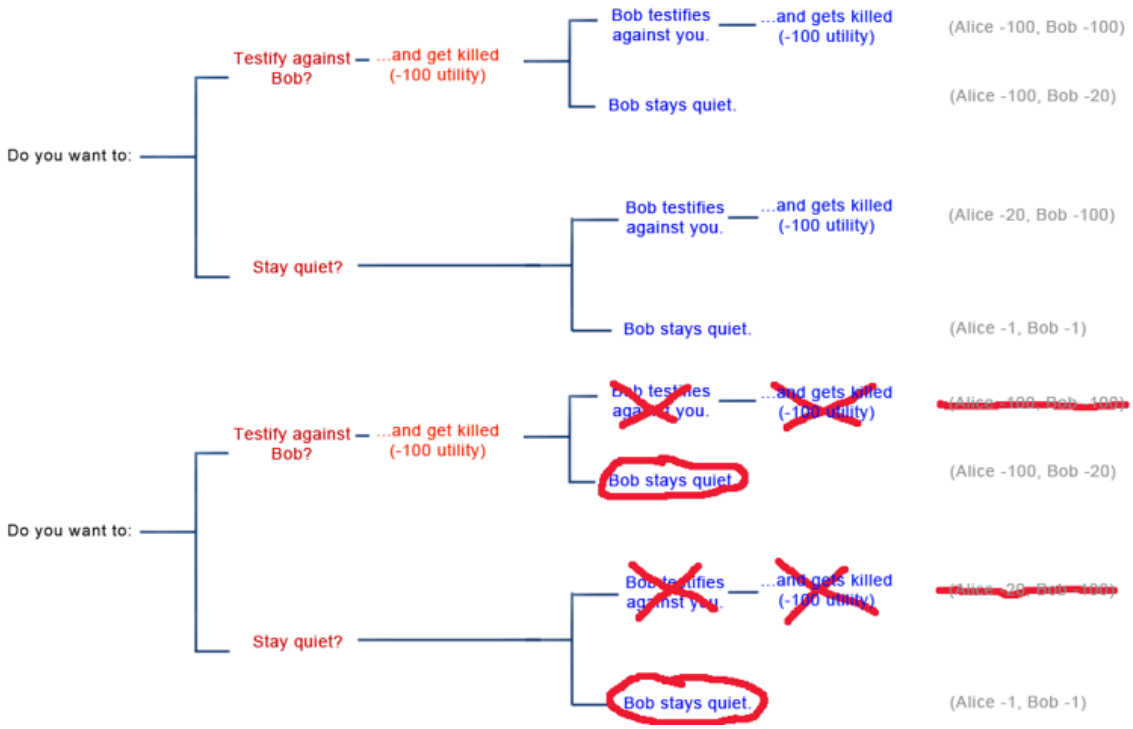
Prisoners' Dilemmas even come up in nature. In baboon tribes, when a female is in “heat”, males often compete for the chance to woo her. The most successful males are those who can get a friend to help fight off the other monkeys, and who then helps that friend find his own monkey loving. But these monkeys 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.

Magic players: "How do I lose?"

An excellent habit that I've noticed among professional players of the game Magic: The Gathering is asking the question "how do I lose?" - a sort of strategic looking into the dark.

Imagine this situation: you have an army ready to destroy your opponent in two turns. Your opponent has no creatures under his command. Victory seems inevitable. And so you ask "how do I lose?"

Because your victory is now the default, the options for your opponent are very limited. If you have a big army, they need to play a card that can deal with lots of creatures at once. If you have a good idea what their deck contains, you can often narrow it down to a single card that they need to play in order to turn the game around. And once you know how you could lose, you can plan to avoid it.

For example, suppose your opponent was playing white. Then their card of choice to destroy a big army would be Wrath of God. That card is the way you could lose. But now that you know that, you can avoid losing to Wrath of God by keeping creature cards in your hand so you can rebuild your army - you'll still win if he doesn't play it, since you winning is the default. But you've made it harder to lose. This is a bit of an advanced technique, since *not* playing all your cards is counterintuitive.

A related question is "how do I win?" This is the question you ask when you're near to losing. And like above, this question is good to ask because when you're really behind, only a few cards will let you come back. And once you know what those cards are, you can plan for them.

For example, suppose you have a single creature on your side. The opponent is attacking you with a big army. You have a choice: you can let the attack through and lose in two turns, or you can send your creature out to die in your defense and lose in three turns. If you were trying to postpone losing, you would send out the creature. But you're more likely to actually *win* if you keep your forces alive - you might draw a sword that makes your creature stronger, or a way to weaken their army, or something. And so you ask "how do I win?" to remind yourself of that.

This sort of thinking is highly generalizable. The next time you're, say, packing for a vacation and feel like everything's going great, that's a good time to ask: "How do I lose? Well, by leaving my wallet behind or by having the car break down - everything else can be fixed. So I'll go put my wallet in my pocket right now, and check the oil and coolant levels in the car."

An analogy is that when you ask "how do I win?" you get to disregard your impending loss because you're "standing on the floor" - there's a fixed result that you get if you don't win, like calling a tow truck if you're in trouble in the car, or canceling your vacation and staying home. Similarly when you ask "how do I lose?" you should be standing on the ceiling, as it were - you're about to achieve a goal that doesn't need to be improved upon, so now's the time to be careful about potential Wraths of God.

AI cooperation is already studied in academia as "program equilibrium"

About a month ago I accidentally found out that the LW idea of quining cooperation is already studied in academia:

1) Moshe Tennenholtz's 2004 paper Program Equilibrium describes the idea of programs cooperating in the Prisoner's Dilemma by inspecting each other's source code.

2) Lance Fortnow's 2009 paper Program Equilibria and Discounted Computation Time describes an analogue of Benja Fallenstein's idea for implementing correlated play, among other things.

3) Peters and Szentes's 2012 paper Definable and Contractible Contracts studies quining cooperation over a wider class of definable (not just computable) functions.

As far as I know, academia still hasn't discovered Loebian cooperation or the subsequent ideas about formal models of UDT, but I might easily be wrong about that. In any case, the episode has given me a mini-crisis of faith, and a new appreciation of academia. That was a big part of the motivation for my previous post.

Kurzweil's predictions: good accuracy, poor self-calibration

Predictions of the future rely, to a much greater extent than in most fields, on the personal judgement of the expert making them. Just one problem - personal expert judgement generally sucks, especially when the experts don't receive immediate feedback on their hits and misses. Formal models perform better than experts, but when talking about unprecedented future events such as nanotechnology or AI, the choice of the model is also dependent on expert judgement.

Ray Kurzweil has a model of technological intelligence development where, broadly speaking, evolution, pre-computer technological development, post-computer technological development and future AIs all fit into the same exponential increase. When assessing the validity of that model, we could look at Kurzweil's credentials, and maybe compare them with those of his critics - but Kurzweil has given us something even better than credentials, and that's a track record. In various books, he's made predictions about what would happen in 2009, and we're now in a position to judge their accuracy. I haven't been satisfied by the various accuracy ratings I've found online, so I decided to do my own.

Some have argued that we should penalise predictions that "lack originality" or were "anticipated by many sources". But hindsight bias means that we certainly judge many profoundly revolutionary past ideas as "unoriginal", simply because they are obvious today. And saying that other sources anticipated the ideas is worthless unless we can quantify how mainstream and believable those sources were. For these reasons, I'll focus only on the accuracy of the predictions, and make no judgement as to their ease or difficulty (unless they say things that were already true when the prediction was made).

Conversely, I won't be giving any credit for "near misses": this has the hindsight problem in the other direction, where we fit potentially ambiguous predictions to what we know happened. I'll be strict about the meaning of the prediction, as written. A prediction in a published book is a form of communication, so if Kurzweil actually meant something different to what was written, then the fault is entirely his for not spelling it out unambiguously.

One exception to that strictness: I'll be tolerant on the timeline, as I feel that a lot of the predictions were forced into a "ten years from 1999" format. So I'll estimate the prediction accurate if it happened at any point up to the end of 2011, if data is available.

The number of predictions actually made seem to vary from source to source; I used my copy of "The Age of Spiritual Machines", which seems to be the original 1999 edition. In the chapter "2009", I counted 63 prediction paragraphs. I then chose ten numbers at random between 1 and 63, and analysed those ten predictions for correctness (those wanting to skip directly to the final score can scroll down). Seeing Kurzweil's nationality and location, I will assume all prediction refer only to technologically advanced nations, and specifically to the United States if there is any doubt. Please feel free to comment on my judgements below; we may be able to build a Less Wrong consensus verdict. It would be best if you tried to reach your own

conclusions before reading my verdict or anyone else's. Hence I present the ten predictions, initially without commentary:

- **Prediction 5**: Cables are disappearing. Communication between components, such as pointing devices, microphones, displays, printers and the occasional keyboard, uses short-distance wireless technology.
- **Prediction 7**: The majority of text is created using continuous speech recognition (CSR) dictation software, but keyboards are still used. CSR is very accurate, far more so than the human transcriptionists who were used up until a few years ago.
- **Prediction 8**: Also ubiquitous are language user interfaces (LUIs) which combine CSR and natural language recognition. For routine matters, such as simple business transactions and information inquiries, LUIs are quite responsive and precise. They tend to be narrowly focused, however, on specific types of tasks. LUIs are frequently combined with animated personalities. Interacting with an animated personality to conduct a purchase or make a reservation is like talking to a person using video conferencing, except the person is simulated.
- **Prediction 18**: In the twentieth century, computers in schools were mostly on the trailing edge, with most effective learning from computers taking place in the home. Now in 2009, while schools are still not on the cutting edge, the profound importance of the computer as a knowledge tool is widely recognised. Computers play a central role in all facets of education, as they do in other spheres of life.
- **Prediction 20**: Students of all ages typically have a computer of their own, which is a thin tabletlike device weighing under a pound with a very high resolution display suitable for reading. Students interact with their computers primarily by voice and by pointing with a device that looks like a pencil. Keyboards still exist, but most textual language is created by speaking. Learning materials are accessed through wireless communication.
- **Prediction 26**: Print-to-speech reading devices for the blind are now very small, inexpensive, palm-sized devices that can read books (those that still exist in paper form) and other printed documents, and other real-world text such as signs and displays. These reading systems are equally adept at reading the trillions of electronic documents that are instantly available from the ubiquitous wireless worldwide network.
- **Prediction 29**: Computer-controlled orthotic devices have been introduced. These "walking machines" enable paraplegics to walk and climb stairs. The prosthetic devices are not yet usable by all paraplegic persons, as many physically disabled persons have dysfunctional joints from years of disuse. However, the advent of orthotic walking systems is providing more motivations to have these joints replaced.
- **Prediction 44**: Intelligent roads are in use, mainly for long-distance travel. Once your car's guidance system locks into the control sensors on one these highways, you can sit back and relax. Local roads, though, are still predominantly conventional.
- **Prediction 48**: There is continuing concern with an underclass that the skill ladder has left far behind. The size of the underclass appears to be stable, however. Although not politically popular, the underclass is politically neutralised through public assistance and the generally high level of affluence.
- **Prediction 53**: Beyond musical recordings, images, and movie videos, the most popular type of digital entertainment object is virtual experience software. These interactive virtual environments allow you to go whitewater rafting on virtual rivers, to hang-glide in a virtual Grand Canyon, or to engage in intimate encounters with your favourite movie star. Users also experience fantasy

> environments with no counterpart in the physical world. The visual and auditory experience of virtual reality is compelling, but tactile interaction is still limited.

Verdict

My scale for judging the predictions is: true, weakly true, weakly false, false.

Prediction 5: My office and the computer I'm typing on seem pretty full of cables. Nevertheless, it is true there has been a rise in wireless technology, and wireless computer components, even if they're not ubiquitous. I'll grade this as a **weakly true**.

Prediction 7: I have failed to find proper data for the first prediction. Anecdotally, it certainly seems false - keyboards are still in ubiquitous use, and I've never personally seen anyone use voice recognition to write documents of any length or even to send texts (a few personal experiments with Siri notwithstanding). The second claim in false: according to an assessment by the National Institute of Standards and Technology, the accuracy of CSR is still nowhere near surpassing human transcription. This leads extra credence to the first claim being false as well: without the diminished error rate, it's very hard to see CSR being used for the majority of text creation. **False**.

Prediction 8: Apart from the belief that the animated personality would be visual, this is a near-perfect description of Siri and similar assistants. The term "ubiquitous" is tricky, but if we interpret it to mean "to be found everywhere" (rather than "everyone has one"), then the prediction is **weakly true** (knocked down from true because of the uncertainty about ubiquity).

Prediction 18: Without needing to do the research, I think we can take this claim as evidently **true**.

Prediction 20: All the stuff about voice recognition is false. The only device that fits that description today is the smartphone, which has not achieved penetration of more than 50% among teenagers in 2011 (teenagers are the median "students of all ages"; adding in university students *as well as* pre-teens should lower the proportion, not raise it). "Learning materials are accessed through wireless communication" is hard to interpret, as it doesn't give any estimate to what proportion of learning material we are talking about. So though we can give Kurzweil kudos for imagining something like the smartphone, the prediction is **weakly false**.

Prediction 26: One can quibble about inexpensive, as the products seem to be in the $600 range, but those products certainly exist for book and magazine reading (though not for most signs and displays, as far as I can tell - certainly not in a form the blind can use). The second sentence is true for some screen readers, making the prediction essentially **true**.

Prediction 29: 2009 timeline wrong, but **true** in later years.

Prediction 44: The relative quantifier in the last sentence ("though, are still predominantly conventional") makes it clear that we should expect intelligent highways to be common among long-distance highways - this isn't a few experimental roads we're talking about. Though we have a few self-driving cars, we have nothing like the intelligent roads implied in this prediction, which specifically implies that most cars on those roads will be self-driven. **False**.

Prediction 48: The first part of the prediction is true. The second sentence seems false, whether one measures the underclass through relative income (where inequality has been increasing) or through an absolute standard of educational attainment (where the various graduating rates have gone up, implying the underclass is decreasing). There are other ways one could measure the underclass, giving different results. Since one could read the underclass as increasing or decreasing, should we take Kurzweil's claim that it is stable as the correct mean? No. All that means is that had he spelt out his claim in more detail at the time, it would likely have ended up false. Ambiguity does not make a false statement true. The last sentence is virtually impossible to confirm or infirm, so the whole prediction is **weakly true and weakly false**.

Prediction 53: This is a tricky one. The Wii and similar game consoles seem to fit the bill to some extent. However the tone suggests he is talking about a virtual reality experience, which is not what we currently have. So, does he mean virtual reality, or does he mean "games like what they had in 1999, except with much better graphics and features"? How would someone at the time have read the prediction? Again, ambiguity cannot be used to make a false statement true. I'm going to work on the assumption that had he merely meant "graphics and features of video games will improve a lot", he would have said so (certainly his prediction seems to promise much more than that). So the prediction is false.

But what if he was talking about modern games? For a start, his initial sentence gets the relative size of the industries wrong (though that can be read as a throw-away statement rather than a prediction). He also doesn't consider things like Facebook games, which make up a large part of the games industry, and are certainly not interactive virtual environments. What about "these virtual environments allow..."? Well, the statement is possibly an utter triviality, claiming that games exist which feature rafting, hang-gliding or erotic situations (that was already true in 1999). Or it claims that features like these are a major component of the most most popular games today, which is false (now, if he'd said "blowing things up with a marvellous amount of weapons..."). Fantasy environment is a much more common feature, so, I'm taking that as correct. Under this interpretation, the prediction is weakly true and weakly false for games. In total, reading the statement either way, I'll classify it as (contentiously) **weakly false**.

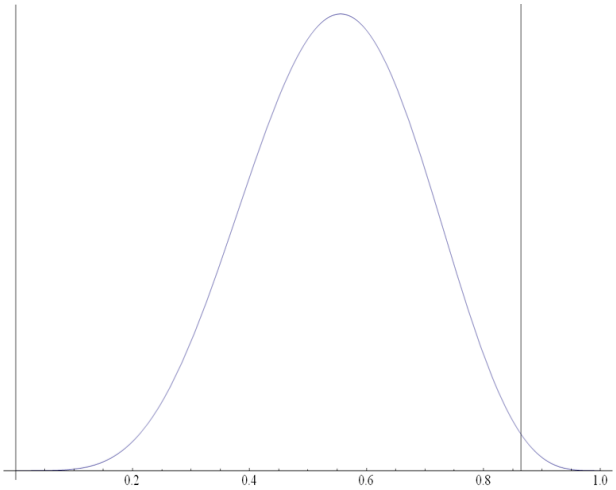
Note: I did read Kurzweil's assessment of his own predictions, *after* I had conducted my own analysis. In that assessment, nearly every ambiguous clause is interpreted in Kurzweil's favour. This could be Kurzweil twisting the predictions in his direction; it could be a blatant example of hindsight bias; or it could be that what Kurzweil meant to say was different from what he wrote. Unfortunately, there is no way for us to tell, so we must make do with what was written and interpret it as best *we* can.

Analysis

So, out of the ten predictions, five are to some extent true, four are to some extent false, and one is unclassifiable (reading through the rest of the predictions, completely informally, these proportions seem roughly correct).

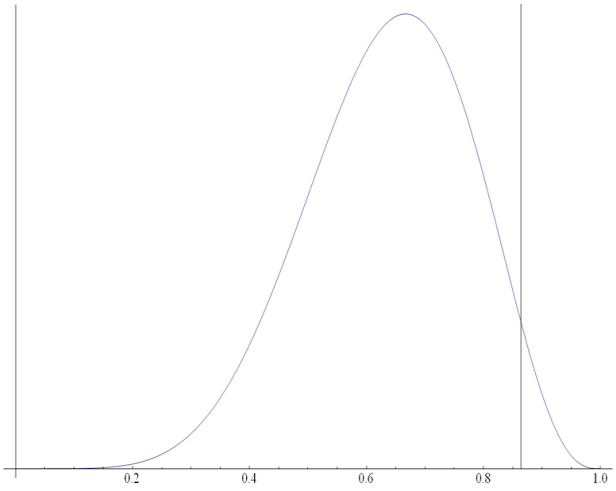
Now imagine Kurzweil as a predictor who gives predictions, each with independent probability p of bring true (alternately, assume that a fixed proportion p of the 63

predictions are true, and pretend 63 is high enough that we can treat p as continuous without much loss). If we start with a uniform prior on p between 0 and 1, then we can update given this data. Model prediction 48 as true or false with equal probability. Then the posterior must be proportional to $(1-p)^5p^5 + (1-p)^4p^6$:

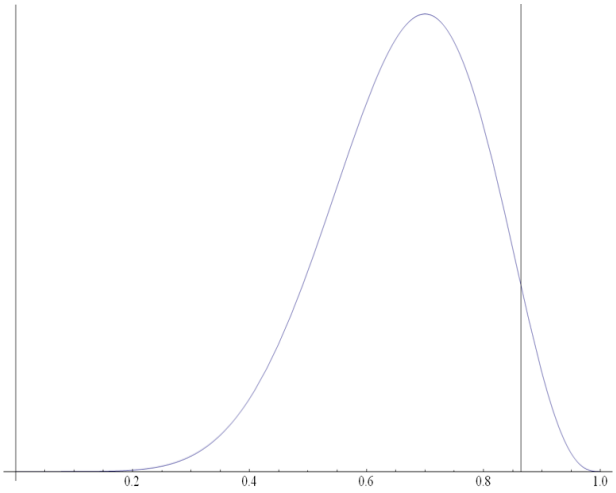


This has a mean above 54%, which I'd say is excellent. A prediction record over 50% for a decade that included huge increases in computer power, September 11th and the great recession is intuitively a very good one. Alas there is no central repository of prediction records from various futurists, but in the absence of that, his track record certainly feels impressive. Don't let the hindsight bias blind you to how hard this was, and don't simply think of every prediction as binary: generally, there are far more ways for a prediction to be false than there are for them to be true.

On the other hand, if we look at Kurzweil's own ranking of the predictions he gave in the "Age of Spiritual Machines", he grades himself as having either 102 out of 108 or 127 out of 147 correct (with caveats that "even the predictions that were considered 'wrong' in this report were not all wrong"). I've plotted the lower $127/147 \approx 0.86$ accuracy on the above graph; that is very far from being a mean estimate (it's in the 99^{th} percentile of the probability distribution). But let's give Kurzweil all we can: we'll reclassify the arguable **prediction 53** as being true (posterior proportional to $(1-p)^4p^6 + (1-p)^3p^7$):



That is still not enough to make his accuracy estimate reasonable: his estimate is in the 96th percentile of the probability distribution. Let's be even more generous: let's reclassify the intermediate **prediction 48** as also being true (posterior proportional to $(1-p)^3p^7$):



Those were very generous adjustments; changing two results is a lot from a sample of ten. But even with the most generous adjustments and taking Kurzweil's lowest estimate of his own accuracy, he is still extraordinarily overconfident: his estimate is in the 94th percentile of the probability distribution. For fun, I flipped another prediction from false to true: even then, his estimate is in the 81th percentile of the probability distribution (and recall that if we were rigorous about the timeline that Kurzweil claimed, at least one of the true prediction would be false).

So what can this tell us about Kurzweil as a futurist, and about the predictions he makes? Essentially two points stand out:

1. He's most likely good at predicting.
2. He's most likely overconfident, reluctant to admit his misses, and hence unlikely to update on his failures.

So I feel we should take Kurzweil's predictions as a good baseline, with much wider error bars and caveats, paying relatively less attention to those areas where we feel that being a good Bayesian updater becomes important. We should thus probably pay more attention to his models than to his interpretation of his models.

Notes on the Psychology of Power

Luke/SI asked me to look into what the academic literature might have to say about people in positions of power. This is a summary of some of the recent psychology results.

The powerful or elite are: fast-planning abstract thinkers who take action (1) in order to pursue single/minimal objectives, are in favor of strict rules for their stereotyped out-group underlings (2) but are rationalizing (3) & hypocritical when it serves their interests (4), especially when they feel secure in their power. They break social norms (5, 6) or ignore context (1) which turns out to be worsened by disclosure of conflicts of interest (7), and lie fluently without mental or physiological stress (6).

What are powerful members good for? They can help in shifting among equilibria: solving coordination problems or inducing contributions towards public goods (8), and their abstracted Far perspective can be better than the concrete Near of the weak (9).

1. Galinsky et al 2003; Guinote, 2007; Lammers et al 2008; Smith & Bargh, 2008
2. Eyal & Liberman
3. Rustichini & Villeval 2012
4. Lammers et al 2010
5. Kleef et al 2011
6. Carney et al 2010
7. Cain et al 2005; Cain et al 2011
8. Eckel et al 2010
9. Slabu et al; Smith & Trope 2006; Smith et al 2008

These benefits may not exceed the costs (is inducing contributions all that useful with improved market mechanisms like assurance contracts - made increasingly famous thanks to Kickstarter?) Now, to forestall objections from someone like Robin Hanson that these traits - if negative - can be ameliorated by improved technology and organizations and the rest just represents our egalitarian forager prejudice against the elites and corporations who gave us the wealthy modern world, I would point out that these traits look like they would be quite effective at maximizing utility and some selected for in future settings...

(Additional cautions include that, in order to control for all sorts of confounds, these are generally small WEIRD samples in laboratory or university settings involving small-scale power shifts, priming, or other cues; as such, all the usual criticisms apply.)

1 Notes

1. key phrases: “moral hypocrisy”; “construal level theory” (Near/Far) (Hanson’s visual summary)
2. check or blacklist any paper related to Diederik Stapel! (eg. “Fraud Case Seen as a Red Flag for Psychology Research”)

2 References

"Power increases hypocrisy: Moralizing in reasoning, immorality in behavior", Lammers et al 2010; warning, Stapel! But Lammers says committee cleared this paper.

> In five studies, we explored whether power increases moral hypocrisy (i.e., imposing strict moral standards on other people but practicing less strict moral behavior oneself). In Experiment 1, compared with the powerless, the powerful condemned other people's cheating more, but also cheated more themselves. In Experiments 2 through 4, the powerful were more strict in judging other people's moral transgressions than in judging their own transgressions. A final study found that the effect of power on moral hypocrisy depends on the legitimacy of the power: When power was illegitimate, the moral-hypocrisy effect was reversed, with the illegitimately powerful becoming stricter in judging their own behavior than in judging other people's behavior. This pattern, which might be dubbed hypercrisy, was also found among low-power participants in Experiments 3 and 4. We discuss how patterns of hypocrisy and hypercrisy among the powerful and powerless can help perpetuate social inequality.
>
> …feelings of power reduce sensitivity to social disapproval (Emerson, 1962; Thibaut & Kelley, 1959), thus reducing the grip of social norms and standards on power holders' behavior (Galinsky et al., 2008). As a result, even very strong norms, such as those regulating sexual behavior or compassion, are often ignored by the powerful (Bargh, Raymond, Pryor, & Strack, 1995; Van Kleef et al., 2008).
>
> - Emerson, R.M. (1962). Power-dependence relations. American Sociological Review, 27, 31–41
> - Thibaut, J.W., & Kelley, H.H. (1959). The social psychology of groups. New York: Wiley & Sons
> - Galinsky, A.D., Magee, J.C., Gruenfeld, D.H, Whitson, J., & Liljenquist, K.A. (2008). Social power reduces the strength of the situation: Implications for creativity, conformity, and dissonance. Journal of Personality and Social Psychology, 95, 1450–1466
>
> Powerful people who feel that their position is illegitimate are less inclined to assertively take what they want (Lammers, Galinsky, Gordijn, & Otten, 2008) and at the same time are less inclined to judge others for doing so, compared with people who feel their power is deserved (Chaurand & Brauer, 2008). Therefore, in our final study, we independently manipulated power and its legitimacy to test whether legitimacy crucially moderates the effect of power on hypocrisy.
>
> - Lammers, J., & Stapel, D.A. (2009). How power influences moral thinking. Journal of Personality and Social Psychology, 97, 279–289
> - Chaurand, N., & Brauer, M. (2008). What determines social control? People's reactions to counternormative behaviors in urban environments. Journal of Applied Social Psychology, 38, 1689–1715

"Moral Hypocrisy, Power and Social Preferences", Rustichini & Villeval 2012:

> We show with a laboratory experiment that individuals adjust their moral principles to the situation and to their actions, just as much as they adjust their actions to their principles. We first elicit the individuals' principles regarding the fairness and unfairness of allocations in three different scenarios (a Dictator game, an Ultimatum game, and a Trust game). One week later, the same individuals are invited to play those same games with monetary compensation. Finally in the

same session we elicit again their principles regarding the fairness and unfairness of allocations in the same three scenarios.

Our results show that individuals adjust abstract norms to fit the game, their role and the choices they made. First, norms that appear abstract and universal take into account the bargaining power of the two sides. The strong side bends the norm in its favor and the weak side agrees: Stated fairness is a compromise with power. Second, in most situations, individuals adjust the range of fair shares after playing the game for real money compared with their initial statement. Third, the discrepancy between hypothetical and real behavior is larger in games where real choices have no strategic consequence (Dictator game and second mover in Trust game) than in those where they do (Ultimatum game). Finally the adjustment of principles to actions is mainly the fact of individuals who behave more selfishly and who have a stronger bargaining power.

…Individuals destroy the resources of others because of envy (Mui, 1995; Maher, 2010; Charness et al., 2010; Harbring and Irlensbusch, 2011) or for the joy of destruction (Zizzo and Oswald, 2001; Abbink and Sadrieh, 2009); the power of public office sometimes leads politicians to use it for their personal gain (Aidt, 2003); feelings of entitlement push leaders to take more than followers from a common resource (de Cremer and van Dijk, 2005).

- Mui, V.L. (1995). The economics of envy. Journal of Economic Behavior & Organizations, 26(3), 311-336.
- Maher, B. (2010). Research Integrity: Sabotage! Nature, 467, 30 September, 516-518.
- G. Charness, D. Masclet, M.C. Villeval. (2010). Competitive Preferences and Status as an Incentive: Experimental Evidence. IZA Discussion Paper 5034, Bonn
- Harbring, C., Irlensbusch, B. (2011). Sabotage in Tournaments: Evidence from the Laboratory, Management Science, 57(4), 611-627
- Zizzo, D., Oswald, A.J. (2001). Are People Willing to Pay to Reduce Others' Incomes? Annales d'Economie et de Statistique, 63-64, 39-62
- Abbink, K., Sadrieh, A. (2009). The pleasure of being nasty. Economics Letters, 105(3), 306-308.
- Aidt, T.S. (2003). Economic Analysis of Corruption: A Survey. The Economic Journal, 113, F632-F652.
- de Cremer, D., van Dijk, E. (2005). When and why leaders put themselves first: Leader behaviour in resource allocations as a function of feeling entitled. European Journal of Social Psychology, 35, 553-563.

social psychologists studying moral hypocrisy have shown that individuals evaluate more negatively the moral transgression of fair principles when this transgression is enacted by others than when enacted by themselves (Valdesolo and deStefano, 2008).

- Valdesolo, P., deStefano, D.A. (2008). The duality of virtue: Deconstructing the moral hypocrite. Journal of Experimental Social Psychology, 44 (5), 1334-1338.

Such an illusory preference for fairness has been identified by Dana, Weber and Kuang (2007) (see also Larson and Capra, 2009; Grossman, 2010; van der Weele, 2012). Indeed, fairness decreases substantially when the link between fairness and outcome is obfuscated. The choice to play fair is frequently motivated by the

willingness to appear fair more than by the willingness to produce a fair outcome and this is why greater anonymity leads to more selfish transfers in the dictator game (Andreoni and Bernheim, 2009; Ariely et al., 2009).

- Dana, J., Weber R.A., Xi Kuang, J. (2007). Exploiting moral wiggle room: experiments demonstrating an illusory preference for fairness. Economic Theory, 33(1), 67-80
- Larson, T., Capra, M. (2009). Exploiting moral wiggle room: Illusory preference for fairness? A comment. Judgment and Decision Making, 4(6), 467-474
- Grossman, Z. (2010). Strategic ignorance and the robustness of social preferences. Working paper, University of California at Santa Barbara
- Van der Weele, J. (2012). When ignorance is innocence: on information avoidance in moral dilemmas. SSRN working paper.
- Andreoni, J., Bernheim, B.D. (2009). Social Image and the 50-50 Norm: A Theoretical and Experimental Analysis of Audience Effects. Econometrica, 77(5), 1607-1636
- Ariely, D., Bracha, A., Meier, S. (2009). Doing Good or Doing Well? Image Motivation and Monetary Incentives in Behaving Prosocially. American Economic Review, 99(1), 544-555

"Breaking the Rules to Rise to Power: How Norm Violators Gain Power in the Eyes of Others", Kleef et al 2011:

Four studies support this hypothesis. Individuals who took coffee from another person's can (Study 1), violated rules of bookkeeping (Study 2), dropped cigarette ashes on the floor (Study 3), or put their feet on the table (Study 4) were perceived as more powerful than individuals who did not show such behaviors. The effect was mediated by inferences of volitional capacity, and it replicated across different methods (scenario, film clip, face-to-face interaction), different norm violations, and different indices of power (explicit measures, expected emotions, and approach/inhibition tendencies).

…''Power tends to corrupt, and absolute power corrupts absolutely,'' wrote Lord Acton to Bishop Mandell Creighton in 1887. This classic adage not only reflects popular sentiments about power; it is also supported by scientific research (e.g., Kipnis, 1972).

- Kipnis, D. (1972). Does power corrupt? Journal of Personality and Social Psychology, 24, 33-41

Individuals who feel powerful are more likely to act in goal-congruent ways (e.g., by switching off an annoying fan) than those who feel less powerful (Galinsky, Gruenfeld, & Magee, 2003). Powerful individuals are also more likely to take risks (Anderson & Galinsky, 2006), show approach-related tendencies and goal-directed action (Guinote, 2007; Lammers, Galinsky, Gordijn, & Otten, 2008; Smith & Bargh, 2008), express their emotions (Hecht & Lafrance, 1998), act based on their dispositional inclinations (Chen, Lee-Chai, & Bargh, 2001) and momentary desires (Van Kleef & Cote, 2007), and ignore situational pressures (Galinsky et al., 2008).

- Galinsky, A. D., Gruenfeld, D. H., & Magee, J. C. (2003). From power to action. Journal of Personality and Social Psychology, 85, 453-466
- Anderson, C., & Galinsky, A. D. (2006). Power, optimism and risk-taking. European Journal of Social Psychology, 36, 511-536

- Guinote, A. (2007). Power and goal pursuit. Personality and Social Psychology Bulletin, 33, 1076-1087
- Lammers, J., Galinsky, A. D., Gordijn, E. H., & Otten, S. (2008). Illegitimacy moderates the effects of power on approach. Psychological Science, 19, 558-564
- Smith, P. K., & Bargh, J. A. (2008). Nonconscious effects of power on basic approach and avoidance tendencies. Social Cognition, 26, 1-24
- Hecht, M. A., & Lafrance, M. (1998). License or obligation to smile: The effect of power and sex on amount and type of smiling. Personality and Social Psychology Bulletin, 24, 1332-1342
- Chen, S., Lee-Chai, A. Y., & Bargh, J. A. (2001). Relationship orientation as a moderator of the effects of social power. Journal of Personality and Social Psychology, 80, 173-187
- Van Kleef, G. A., & & Cote, S (2007). Expressing anger in conflict: When it helps and when it hurts. Journal of Applied Psychology, 92, 1557-1569
- Galinsky, A. D., Gruenfeld, D. H., Magee, J. C., Whitson, J. A., & Liljenquist, K. A. (2008). Power reduces the press of the situation: Implications for creativity, conformity, and dissonance. Journal of Personality and Social Psychology, 95, 1450-1466

This behavioral disinhibition makes powerful people more likely to exhibit socially inappropriate behavior. Compared to lower power individuals, powerful individuals are likely to take more cookies from a common plate, eat with their mouths open, and spread crumbs (Keltner et al., 2003); interrupt conversation partners and invade their personal space (DePaulo & Friedman, 1998); fail to take another's perspective (Galinsky, Magee, Inesi, & Gruenfeld, 2006); ignore other people's suffering (Van Kleef et al., 2008); stereotype (Fiske, 1993) and patronize others (Vescio, Gervais, Snyder, & Hoover, 2005); cheat (Lammers, Stapel, & Galinsky, 2010); take credit for the contributions of others (Kipnis, 1972); treat other people as a means to their own ends (Gruenfeld, Inesi, Magee, & Galinsky, 2008); and sexualize and harass low-power women (Bargh, Raymond, Pryor, & Strack, 1995). Powerful people also exhibit more aggression (Haney, Banks, & Zimbardo, 1973), and this is relatively acceptable to others (Porath, Overbeck, & Pearson, 2008). In fact, in several European countries the liberty to violate norms without sanction is perceived as a defining feature of the power holder (Mondillon et al., 2005). Although the powerful impose strict moral standards on others, they practice less strict moral behavior themselves (Lammers et al., 2010).

- Keltner, D., & Gruenfeld, D. H, & Anderson, C. (2003). Power, approach, and inhibition. Psychological Review, 110, 265-284.
- DePaulo, B. M., & Friedman, H. S. (1998). Nonverbal communication. In D. Gilbert, S. T. Fiske, & G. Lindzey (Eds.), Handbook of social psychology (pp. 3-40). New York: McGraw-Hill
- Galinsky, A. D., Magee, J. C., Inesi, M. E., & Gruenfeld, D. H. (2006). Power and perspectives not taken. Psychological Science, 17, 1068-1074
- Van Kleef, G. A., Oveis, C., Van Der Lowe, I., LuoKogan, A., Goetz, J., & Keltner, D. (2008). Power, distress, and compassion: Turning a blind eye to the suffering of others. Psychological Science, 19, 1315-1322
- Fiske, S. T. (1993). Controlling other people: The impact of power on stereotyping. American Psychologist, 48, 621-628
- Lammers, J., Stapel, D. A., & Galinsky, A. D. (2010). Power increases hypocrisy: Moralizing in reasoning, immorality in behavior. Psychological Science, 21, 737-744; WARNING: Stapel! Lammers states that this paper is untainted:

IMPORTANT: Regarding the scientific fraud of my former supervisor Stapel: the committee Levelt has investigated all my work with Stapel. All my work on the topic of power has been cleared from suspicion of data-fraud. This research is all based on data that I collected myself or collected together with other co-authors (i.e. not Stapel). There is one paper (on racism in legal decisions) where I was misled. This paper contains false data. It is currently being retracted.

- Gruenfeld, D. H., Inesi, M. E., Magee, J. C., & Galinsky, A. D. (2008). Power and the objectification of social targets. Journal of Personality and Social Psychology, 95, 111-127
- Bargh, J. A., Raymond, P., Pryor, J. B., & Strack, F. (1995). Attractiveness of the underling: An automatic power-sex association and its consequences for sexual harassment and aggression. Journal of Personality and Social Psychology, 68, 768-781
- Haney, C., Banks, C., & Zimbardo, P. (1973). Interpersonal dynamics in a simulated prison. International Journal of Criminology and Penology, 1, 69-97
- Porath, C. L., Overbeck, J., & Pearson, C. M. (2008). Picking up the gauntlet: How individuals respond to status challenges. Journal of Applied Social Psychology, 38, 1945-1980
- Mondillon, L., Niedenthal, P. M., Brauer, M., Rohman, A., Dalle, N., & Uchida, Y. (2005). Beliefs about power and its relation to emotional experience: A comparison of Japan, France, Germany, and United States. Personality and Social Psychology Bulletin, 31, 1112-1122

…research on adolescent aggression indicates that bullying behavior is associated with prestige (Savin-Williams, 1976; Sijtsema, Veenstra, Lindenberg, & Salmivalli, 2009).

- Savin-Williams, R. C. (1976). An ethological study of dominance formation and maintenance in a group of human adolescents. Child Development, 47, 972-979
- Sijtsema, J. J., Veenstra, R., Lindenberg, S., & Salmivalli, C. (2009). Empirical test of bullies' status goals: Assessing direct goals, aggression, and prestige. Aggressive Behavior, 35, 57-67

"Morality and Psychological Distance: A Construal Level Theory Perspective", Eyal & Liberman:

In this chapter, we propose one answer to the question of when values and moral principles play a central role in people's judgments and plans. We explore the possibility that values and moral principles are more prominent in judgments and predictions regarding psychologically more distant events. This perspective is based on construal level theory (CLT; Liberman & Trope, 2008; Liberman, Trope, & Stephan, 2007; Trope & Liberman, in press), according to which the construal of psychologically more distant situations highlights more abstract, high-level features. Because values and moral rules tend to be abstract and general, people are more likely to use them in construing, judging, and planning with respect to psychologically more distant situations.

For example, Nussbaum, Trope, and Liberman (2003, Study 2) conceptualized personal dispositions as high-level construals and situational constrains as low-level construals and demonstrated that people expect others to express their personal dispositions and act consistently across different situations in the distant

future more than in the near future. In the study, participants imagined an acquaintance's behavior in four different situations (e.g., a birthday party, waiting in line at the supermarket) in either the near future or the distant future and rated the extent to which the acquaintance would display 15 traits (e.g., behave in a friendly vs. an unfriendly manner) representative of the Big Five personality dimensions (extraversion, agreeableness, conscientiousness, emotional stability, and intellect). Cross- situational consistency was assessed by computing, for each of the 15 traits, the variance in each predicted behavior across the four situations and the correlations among the predicted behaviors in the four situations. As predicted, participants expected others to behave more consistently across distant-future situations than across near-future situations. This finding was replicated with ratings of participants' own behavior in different situations: Participants anticipated exhibiting more consistent traits in the distant future than in the near future (Wakslak, Nussbaum, Liberman, & Trope, 2008, Study 5).

- Nussbaum, S., Trope, Y., & Liberman, N. (2003). "Creeping dispositionism: The temporal dynamics of behavior prediction". Journal of Personality and Social Psychology, 84, 485-497
- Wakslak, C. J., Nussbaum, S., Liberman, N., & Trope, Y. (2008). "Representations of the self in the near and distant future". Journal of Personality and Social Psychology, 95, 757-773

For each scenario (e.g., national flag), participants chose between two restatements of each action. One restatement referred to an abstract moral principle (high-level construal; e.g., desecrating a national symbol) and the other restatement referred to the means of carrying out the action (low-level construal; e.g., cutting a flag to created rags). We found that distant-future transgressions were identified in moral terms more often than near-future transgressions. These findings suggest that people are more likely to think of a temporally distant action, rather than one in the near term, as having moral implications. CLT predicts similar results for other forms of psychological distance: Situations should be more readily construed in terms of moral principles when they occurred further back in the past, when they apply to more socially or spatially distant individuals or groups, and when they are less likely actually to occur. When the same actions are proximal, they are more likely to be construed in terms that are devoid of moral implications. For example, accepting minority students with lower grades into one's university will be seen as "endorsing affirmative action" when it is unlikely to be implemented, but it will be seen in more concrete terms (e.g., as "making acceptance rules more complicated") when it becomes more likely.

The vignettes also included situational details that rendered the transgressions harmless (low-level information; e.g., the siblings used contraceptives, they had sex just once, they kept it a secret). Participants were instructed to imagine that the transgressions would occur tomorrow (the near-future condition) or next year (the distant-future condition) and judged the extent of its wrongness. We found that moral transgressions were judged more severely when imagined in the distant future compared to the near future. The same pattern occurred with social distance (Eyal et al., 2008, Study 3), which was manipulated by asking participants to focus either on the feelings and thoughts they experienced while reading about the events (low social distance) or to think about another person they knew, such as a colleague, a friend, or a neighbor, and focus on the feelings and thoughts that this person would experience while reading about the events (high social distance). Notice that the social distance manipulation did not involve judging one's own versus another person's actions, but only one's imagined

perspective. Notably, this manipulation does not support interpreting the results in terms of moral hypocrisy, according to which people judge their own moral transgressions less harshly than another person's transgressions because they wish to appear better than others. As predicted, moral transgressions were judged more harshly when imagined from a third person perspective (high social distance) compared to one's own perspective (low social distance). Another study (Eyal et al., 2008, Study 4) examined temporal distance effects on judgments of moral acts. Participants read vignettes that described virtuous acts related to widely accepted moral principles (high-level information; e.g., a couple adopting a disabled child) as well as low-level, situational details that rendered the acts less noble (e.g., the government offering large adoption payments). It was found that these behaviors were judged to be more virtuous when they were described as happening in the distant future rather than the near future.

Temporal distance from moral transgressions was also found to affect people's emotional responses. Agerstrom and Bjorklund (2009, Studies 1 and 2) asked Swedish participants to imagine situations that involved a threat to human welfare taking place in the near future (today) or in the distant future (in 30 years). For example, one scenario, set in Darfur, Africa, described a woman who was raped and beaten by the Janjaweed militia. Each scenario was followed by a description of a prosocial action that, if taken, could improve the situation (e.g., donate money). Participants rated how wrong it would be for another Swedish citizen not to take the proposed prosocial action given that they had the means to do so. They also rated how angry they would feel if the target person failed to take the prosocial action. It was found that distant-future moral failures were judged more harshly and invoked more anger than near-future moral failures.

In another study, Agerstrom and Bjorklund (2009) examined whether the greater reliance on moral principles in judgments of distant-future compared to near-future transgressions would generalize to individuals' self-perceptions. Participants rated the likelihood of engaging in prosocial actions in reaction to other people's moral transgressions. For example, participants indicated how much money they were willing to donate to help improve the situation in Darfur. As predicted, participants were more likely to express prosocial behavioral intentions when imagining the act occurring in the more distant future. Taken together, these findings suggest that moral rules are more likely to guide people's judgments of distant rather than proximal behaviors.

- Agerström, J., & Björklund, F. (2009). Temporal distance and moral concerns: Future morally questionable behavior is perceived as more wrong and evokes stronger prosocial intentions. Basic and Applied Social Psychology, 31, 49-59

For example, individuals for whom altruism was subordinate in importance to achievement were more likely to refuse to help a fellow student in the distant future than in the near future, whereas individuals for whom achievement was subordinate to altruism were more likely to help a fellow student in the distant future than in the near future. These findings show that secondary values, which are nonetheless part of an individual's self-identity, may mask the influence of central values on near future intentions. Centrality of values may be defined not only within an individual but also within a situation. For example, when medically treating a person from a rival group in a war, the competition is central and mercy is secondary, whereas in a hospital, the reverse is true. An interesting prediction that follows from CLT is that the secondary value will guide behavioral intentions

in the near future more than in the distant future. Thus, in a war, benevolence will come into play in near-future plans more than in distant-future plans, leading people to be more merciful than would otherwise be expected. In his poem “After the Battle”, Victor Hugo tells about his father (“that hero with the sweetest smile”), an officer in the war against Spain, who encounters a Spaniard soldier asking for something to drink. Although on the battlefield, and although the Spaniard tries to kill him, the officer orders: “All the same, give him something to drink.”

“People with Power are Better Liars”, Carney et al 2010:

> But lying does not come without cost. Ordinary lie-tellers experience negative emotions, decrements in mental function, and physiological stress. Liars are also at risk of getting caught. Despite people’s best attempts to get away with their prevarications, lies are often behaviorally “leaked” through subtle changes in body movement and speech rate. Power, it seems, enhances the same emotional, cognitive, and physiological systems that lie-telling depletes. People with power enjoy positive emotions, increases in cognitive function (4-5), and physiological resilience such as lower levels of the stress hormone cortisol (6-7). Thus, holding power over others might make it easier for people to tell lies.
>
> 1. D. Keltner, D.H. Gruenfeld, C. Anderson, Psychol Rev. 110, 265-284 (2003).
> 2. P.K. Smith, N.B. Jostmann, A.D. Galinsky, W. van Dijk, Psychol Sci. 19, 441-447 (2008).
> 3. R.M. Sapolsky, S.C. Alberts. J. Altmann, J Arch Gen Psychi. 54, 1137-1143 (1997).
> 4. S. Cohen, W.J. Doyle, A. Baum, Psychosom Med. 68, 414-420 (2006)
>
> Participants were assigned to the role of “leader” or “subordinate” and engaged in a series social interactions in which the leader had control over the subordinate’s monetary and social outcomes (9)….If the individual could successfully convince the experimenter (regardless of whether they were lying) they could keep the $100 in cash. All participants were then interviewed about whether they had stolen the money: half were lying and half were telling the truth. The interviewer (blind to experimental condition) asked all participants the same critical questions (e.g., “Did you steal the $100?”; “Why should I believe you?”). After the interview, participants completed measures of moral emotional feelings (rated emotion terms: bashful, guilty, troubled, scornful) and a computerized task assessing degree of cognitive impairment. All participants provided saliva samples before and after the experiment to assess changes in the stress hormone cortisol (9). The interviews were videotaped and coded for two, classic nonverbal markers of deception: one-sided shoulder shrugs and accelerated prosody (9). Low-power individuals showed the expected emotional, cognitive, physiological, and behavioral signs of deception; in contrast, powerful people demonstrated no evidence of lying across emotion, cognition, physiology, or behavior (see Figure). In other words, power acted as a buffer allowing the powerful to lie significantly more easily (less disturbing emotion, less cognitive impairment, less of a rise in the stress hormone cortisol) and more effectively (fewer nonverbal cues associated with lying). Only low-power individuals felt badly after lying (panel A), suffered cognitive impairment (panel B), spiked in levels of the stress hormone cortisol (panel C), and demonstrated nonverbal “leakage” (more one-sided shoulder shrugs and accelerated prosody; panel D). (9)

“Psychological perspectives on the fiduciary business”, Donald C. Langevoort

But the investment game has been manipulated in numerous ways that produce differing levels of trusting and greater selfishness. One of particular interest is the introduction of the possibility that, at the end of the game, the trustor will learn whether she gets something back but will not know whether this is the result of the trustee's choice or some exogenous force – e.g., luck.34 Given the opportunity to hide behind the possibility that a return of nothing was just bad luck for the trustor, trustees predictably keep more for themselves, presumably rationalizing the outcome as fair in an uncertain world. The authors of one such study recently drew parallels to financial relationships between investors and securities professionals, because the financial markets generate a great deal of good and bad luck that obscures the value added by professional trustworthiness.35

1. Radu Vranceanu et al., Trust and Financial Trades: Lessons from an Investment Game Where Reciprocators Can Hide Behind Probabilities 6 (ESSEC Bus. Sch., Working Paper No. 10007, 2010), available at http://papers.ssrn.com/sol3/papers.cfm?abstract_id=1 611666.
2. See id. at 14-15.

Unfortunately, high testosterone levels do not fit well with fiduciary characteristics like empathy and moral decision-making. Emerging research on the subject suggests that testosterone buffers emotional constraints on aggression and risk-taking, leading to a more "cold" utilitarian calculus and a greater willingness to do harm to gain a preferred outcome.49

See Dana R. Carney & Malia F. Mason, Decision Making and Testosterone: When the Ends Justify the Means, 46 J. EXPERIMENTAL SOC. PSYCHOL. 668, 668-69 (2010). As the authors point out, the ends need not necessarily be immoral. Id. at 670.

Power also seems to increase hypocrisy – insistence on adherence to strict norms by others, while enjoying far greater nimbleness in justifying one's own departures on utilitarian or other rationalized grounds52 – and optimism and risk-taking.53 Of course, power may be gained in the first place by those skilled at rationalization and willing to take risks, in which case there is a dynamic feedback loop that is likely to generate increasing hypocrisy and hubris over time.

1. See Joris Lammers et al., Power Increases Hypocrisy: Moralizing in Reasoning, Immorality in Behavior, 21 PSYCHOL. SCI. 737, 738 (2010).
2. See Cameron Anderson & Adam D. Galinsky, Power, Optimism, and Risk-taking, 36 EUR. J. SOC. PSYCHOL. 511, 516 (2006). In turn, this pattern may connect to testosterone or other physiological effects. See Carney & Mason, supra note 49, at 668.

"The Dirt on Coming Clean: Perverse Effects of Disclosing Conflicts of Interest", Cain et al 2005

Although disclosure is often proposed as a potential solution to these problems, we show that it can have perverse effects. First, people generally do not discount advice from biased advisors as much as they should, even when advisors' conflicts of interest are disclosed. Second, disclosure can increase the bias in advice because it leads advisors to feel morally licensed and strategically encouraged to exaggerate their advice even further. As a result, disclosure may fail to solve the problems created by conflicts of interest and may sometimes even make matters worse.

…In the domain of medicine, for example, research shows that while many people are ready to acknowledge that doctors might generally be affected by conflicts of interest, few can imagine that their own doctors would be affected (Gibbons et al. 1998). Indeed, it is even possible that disclosure could sometimes increase rather than decrease trust, especially if the person with the conflict of interest is the one who issues the disclosure. Research suggests that when managers offer negative financial disclosures about future earnings, they are regarded as more credible agents, at least in the short term (Lee, Peterson, and Tiedens 2004; Mercer, forthcoming). Thus, if a doctor tells a patient that her research is funded by the manufacturer of the medication that she is prescribing, the patient might then think (perhaps rightly) that the doctor is going out of her way to be open or that she is "deeply involved" and thus knowledgeable. Thus, disclosure could cause the estimator to place more rather than less weight on the advisor's advice. Third, even when estimators realize that they should make some adjustment for the conflict of interest that is disclosed, such adjustments are likely to be insufficient. As a rule, people have trouble unlearning, ignoring, or suppressing the use of knowledge (such as biased advice) even if they are aware that it is inaccurate (Wilson and Brekke 1994). Research on anchoring, for example, shows that quantitative judgments are often drawn toward numbers (the anchors) that happen to be mentally available. This effect holds even when those anchors are known to be irrelevant (Strack and Mussweiler 1997; Tversky and Kahneman 1974), unreliable (Loftus 1979), or even manipulative (Galinsky and Mussweiler 2001; Hastie, Schkade, and Payne 1999). Research on the "curse of knowledge" (Camerer, Loewenstein, and Weber 1989) shows that people's judgments are influenced even by information they know they should ignore. And research on what has been called the "failure of evidentiary discreditation" shows that when the evidence on which beliefs were revised is totally discredited, those beliefs do not revert to their original states but show a persistent effect of the discredited evidence (Skurnik, Moskowitz, and Johnson 2002; Ross, Lepper, and Hubbard 1975). Furthermore, attempts to willfully suppress undesired thoughts can lead to ironic rebound effects, in some cases even increasing the spontaneous use of undesired knowledge (Wegner 1994).

…More interesting, and as predicted, all three measures also reveal that disclosure led to greater distortion of advice. The amount that advisors exaggerated, calculated by subtracting advisors' own personal estimates from their public suggestions, was significantly greater in the high/disclosed condition than in either of the other two conditions ($p<0.05$) and significantly greater by the other two measures as well: advisor suggestion minus actual jar values and advisor suggestion minus the average of personal estimates in the accurate condition ($p<0.05$ for both). In the accurate condition, for example, advisors provided estimators with suggestions of jar values that were, on average, within $1 of their own personal estimates. In the high/undisclosed condition, however, advisors gave suggestions that were $3.32 greater than their own personal estimates, and in the high/disclosed condition, they gave suggestions that were inflated more than twice as much, at more than $7 above their own personal estimates. Disclosure, it appears, did lead advisors to provide estimators with more biased advice.

…Although disclosures did increase discounting by estimators, albeit not significantly, this discounting was not sufficient to offset the increase in the bias of the advice they received. As Table 6 (fourth row) shows, estimator discounting increased, on average, less than $2 from the accurate condition to the high/undisclosed condition and less than $2.50 from the high/undisclosed

condition to the high/disclosed condition. However, Table 5 (second row) shows that suggestions increased, on average, almost $4 from the accurate condition to the high/undisclosed condition and increased $4 again from the high/undisclosed condition to the high/disclosed condition. Thus, while estimators in the high/disclosed condition discounted suggestions about $4 more than did estimators in the accurate condition, the advice given in the high/disclosed condition was almost $8 higher than advice given in the accurate condition. Instead of correcting for bias, estimates were approximately 28 percent higher in the high/disclosed condition than in the accurate condition (first row of Table 6).

"When Sunlight Fails to Disinfect: Understanding the Perverse Effects of Disclosing Conflicts of Interest", Cain et al 2011

Studies 1 and 2 examine psychological mechanisms (strategic exaggeration, moral licensing) by which disclosure can lead advisors to give more-biased advice. Study 3 shows that disclosure backfires when advice recipients who receive disclosure fail to sufficiently discount and thus fail to mitigate the adverse effects of disclosure on advisor bias. Study 4 identifies one remedy for inadequate discounting of biased advice: explicitly and simultaneously contrasting biased advice with unbiased advice.

…Even in one-shot dictator games (Forsythe et al. 1994), research has long shown that many people will share resources and show self-restraint toward anonymous others (Camerer 2003), especially when it is common knowledge that the recipient expects such benevolence (Dana, Cain, and Dawes 2006). Likewise, research on cheating behavior shows that people do not tend to cheat as much as they can get away with, only to the extent that they can rationalize to themselves (Mazar, Amir, and Ariely 2008).

…When the welfare of others is a consideration, disclosure might reduce moral concerns. Prior research has suggested that when people demonstrate ethical behavior, they often become more likely to subsequently exhibit ethical lapses (Jordan, Mullen, and Murnighan 2009; Zhong, Liljenquist, and Cain 2009). For example, people who are given an opportunity to demonstrate their own lack of prejudice are more likely to subsequently display discriminatory behavior (Monin and Miller 2001). Likewise, after a conflict of interest has been disclosed, advisors may feel that advisees have been warned and that advisors are "morally licensed" to provide biased advice.

…Disclosure of a conflict of interest can also reduce the perceived immorality of giving biased advice by signaling that bias is widespread and therefore less aberrant (Schultz et al. 2007). If advice recipients' expectations affect advisor behavior (Dana et al. 2006), then the lowered expectations for honesty that come with disclosure might allow an advisor to rationalize providing biased advice because that is exactly what the advisee expects, or should expect, to receive.

…Why is the call for disclosure so popular despite how it can backfire? One possible explanation is that most people are simply not aware of disclosure's pitfalls. At first glance, disclosure seems like a sensible remedy to a situation in which one party possesses an otherwise hidden incentive to mislead another party. A more cynical explanation would play on the Chicago Theory of Regulation (Becker 1983; Peltzman 1976; Stigler 1971), which posits that regulation typically exists not for the general benefit of society but for the benefit of the regulated groups. These entities might be aware of the ineffectiveness of disclosure but

accept it because it benefits them. For example, even though consumer advocates fought hard for warning labels on cigarette packages, the tobacco industry has defended itself against litigation since then by citing the warning labels as evidence that consumers knew the risks. “What was intended as a burden on tobacco became a shield instead” (Action on Smoking and Health 2001). Moreover, even the regulators may be attracted to disclosure if they see it as absolving them of responsibility for protecting consumers by ostensibly empowering consumers to protect themselves. Disclosure may also be perceived as the lesser of evils for those who might otherwise face more substantive regulation. For example, pharmaceutical firms are often strong proponents of disclosure laws, since it is better for them (and for researchers who receive their funding) if researchers must disclose financial ties to the industry rather than actually having to sever them. This all suggests that disclosure may be problematic for more reasons than those identified by the experiments reported above. It would be a mistake, however, to conclude that disclosure is always counterproductive, as some recent laboratory research illustrates (Church and Kuang 2009; Koch and Schmidt 2009). Research on practical examples of disclosure, summarized in Full Disclosure (Fung, Graham, and Weil 2007), also shows that disclosure can have real beneficial effects. For example, following a spate of highly publicized SUV rollovers, regulations that required auto manufacturers to publicly disclose rollover ratings led to significant and rapid changes in auto design, resulting in a general decrease in the rollover risk for SUVs. Disclosure is likely to be helpful when information is disclosed in an easily digestible form (or is made available to intermediaries, e.g., ratings companies, who process it for consumers) and when it is clear how one should respond to the disclosed information. The rollover ratings met both criteria: the ratings were represented simply as one to five stars, making it easy for consumers to compare —that is, evaluate jointly—the relative rollover risks of various SUVs. Even when information isn’t presented in such a simple form, disclosure is likely to prove helpful when the recipients are savvy repeat-players who know what to do with the disclosed information, such as institutional investors, experienced attorneys, or managers in government agencies (Church and Kuang 2009; Malmendier and Shanthikumar 2007). Disclosure is much less likely to help individuals such as personal investors, purchasers of insurance, home buyers, or patients, who are unlikely to possess the knowledge or experience to know how much they should discount advice or whether they should get a second opinion in a given conflict-of-interest situation (Malmendier and Shanthikumar 2007).

“Power Posing: Brief Nonverbal Displays Affect Neuroendocrine Levels and Risk Tolerance” Carney et al 2010

As predicted, results revealed that posing in high-power (vs. low-power) nonverbal displays caused neuroendocrine and behavioral changes for both male and female participants: High-power posers experienced elevations in testosterone, decreases in cortisol, and increased feelings of power and tolerance for risk; low-power posers exhibited the opposite pattern. In short, posing in powerful displays caused advantaged and adaptive psychological, physiological, and behavioral changes – findings that suggest that embodiment extends beyond mere thinking and feeling, to physiology and subsequent behavioral choices.

…The neuroendocrine profiles of the powerful differentiate them from the powerless, on two key hormones—testosterone and cortisol. In humans and other animals, testosterone levels both reflect and reinforce dispositional and situational status and dominance; internal and external cues cause testosterone to rise,

increasing dominant behaviors, and these behaviors can elevate testosterone even further (Archer, 2006; Mazur & Booth, 1998). For example, testosterone rises in anticipation of a competition and as a result of a win, but drops following a defeat (e.g., Booth, Shelley, Mazur, Tharp, & Kittok, 1989), and these changes predict the desire to compete again (Mehta & Josephs, 2006). In short, testosterone levels, by reflecting and reinforcing dominance, are closely linked to adaptive responses to challenges.

- Archer, J. (2006). Testosterone and human aggression: An evaluation of the challenge hypothesis. Neuroscience & Biobehavioral Reviews, 30, 319–345.
- Mazur, A., & Booth, A. (1998). Testosterone and dominance in men. Behavioral & Brain Sciences, 21, 353–397
- Booth, A., Shelley, G., Mazur, A., Tharp, G., & Kittok, R. (1989). Testosterone and winning and losing in human competition. Hormones and Behavior, 23, 556–571.
- Mehta, P.H., & Josephs, R.A. (2006). Testosterone change after losing predicts the decision to compete again. Hormones and Behavior, 50, 684–692

Power is also linked to the stress hormone cortisol: Power holders show lower basal cortisol levels and lower cortisol reactivity to stressors than powerless people do, and cortisol drops as power is achieved (Abbott et al., 2003; Coe, Mendoza, & Levine, 1979; Sapolsky, Alberts, & Altmann, 1997). Although short-term and acute cortisol elevation is part of an adaptive response to challenges large (e.g., a predator) and small (e.g., waking up), the chronically elevated cortisol levels seen in low-power individuals are associated with negative health consequences, such as impaired immune functioning, hypertension, and memory loss (Sapolsky et al., 1997; Segerstrom & Miller, 2004). Low-power social groups have a higher incidence of stress-related illnesses than high-power social groups do, and this is partially attributable to chronically elevated cortisol (Cohen et al., 2006). Thus, the power holder’s typical neuroendocrine profile of high testosterone coupled with low cortisol—a profile linked to such outcomes as disease resistance (Sapolsky, 2005) and leadership abilities (Mehta & Josephs, 2010)—appears to be optimally adaptive.

- Abbott, D.H., Keverne, E.B., Bercovitch, F.B., Shively, C.A., Mendoza, S.P., Saltzman, W., et al. (2003). Are subordinates always stressed? A comparative analysis of rank differences in cortisol levels among primates. Hormones and Behavior, 43, 67–82
- Coe, C.L., Mendoza, S.P., & Levine, S. (1979). Social status constrains the stress response in the squirrel monkey. Physiology & Behavior, 23, 633–638
- Sapolsky, R.M., Alberts, S.C., & Altmann, J. (1997). Hypercortisolism
- associated with social subordinance or social isolation among
- wild baboons. Archives of General Psychiatry, 54, 1137–1143
- Segerstrom, S., & Miller, G. (2004). Psychological stress and the human immune system: A meta-analytic study of 30 years of inquiry. Psychological Bulletin, 130, 601–630
- Cohen, S., Schwartz, J.E., Epel, E., Kirschbaum, C., Sidney, S., & Seeman, T. (2006). Socioeconomic status, race, and diurnal cortisol decline in the Coronary Artery Risk Development in Young Adults (CARDIA) study. Psychosomatic Medicine, 68, 41–50
- Sapolsky, R.M. (2005). The influence of social hierarchy on primate health. Science, 308, 648–652.

> It is unequivocal that power is expressed through highly specific, evolved nonverbal displays. Expansive, open postures (widespread limbs and enlargement of occupied space by spreading out) project high power, whereas contractive, closed postures (limbs touching the torso and minimization of occupied space by collapsing the body inward) project low power. All of these patterns have been identified in research on actual and attributed power and its nonverbal correlates (Carney, Hall, & Smith LeBeau, 2005; Darwin, 1872/2009; de Waal, 1998; Hall, Coats, & Smith LeBeau, 2005).
>
> - Hall, J.A., Coats, E.J., & Smith LeBeau, L. (2005). Nonverbal and the vertical dimension of social relations: A meta-analysis. Psychological Bulletin, 131, 898–924.

"Reality at Odds With Perceptions: Narcissistic Leaders and Group Performance", Nevicka et al 2011:

> Despite people's positive perceptions of narcissists as leaders, it was previously unknown if and how leaders' narcissism is related to the performance of the people they lead. In this study, we used a hidden-profile paradigm to investigate this question and found evidence for discordance between the positive image of narcissists as leaders and the reality of group performance. We hypothesized and found that although narcissistic leaders are perceived as effective because of their displays of authority, a leader's narcissism actually inhibits information exchange between group members and thereby negatively affects group performance. Our findings thus indicate that perceptions and reality can be at odds and have important practical and theoretical implications.
>
> …For example, narcissists tend to overestimate their intelligence (Campbell, Rudich, & Sedikides, 2002), creativity (Goncalo, Flynn, & Kim, 2010), academic abilities (Robins & Beer, 2001), and leadership capabilities (Judge, LePine, & Rich, 2006). Generally, other people do not agree with narcissists' idealized self-images and perceive narcissists as arrogant, egocentric, overly dominant, and even hostile (Paulhus, 1998). However, the context of leadership constitutes a notable exception in which narcissists tend to be judged positively. For example, individuals with high levels of narcissism receive higher leadership ratings than individuals with low levels of narcissism do (Judge et al., 2006) and tend to emerge as leaders in groups (Brunell et al., 2008; Nevicka, De Hoogh, Van Vianen, Beersma, & McIlwain, 2011). In addition, higher narcissism in U.S. presidents is associated with more positive evaluations of their leadership (Deluga, 1997). It is therefore not surprising that narcissistic characteristics are ascribed to many prominent leaders, such as Nicolas Sarkozy (De Sutter & Immelman, 2008) and Steve Jobs (Robins & Paulhus, 2001).
>
> …Of the two prior studies investigating this question, one found no effects of narcissistic leadership on performance (Brunell et al., 2008), and the other showed that organizational performance was merely more volatile, but no worse or better, because of narcissistic leaders' risky decision making (Chatterjee & Hambrick, 2007). Unfortunately, neither of these studies examined the effects of narcissistic leaders on group dynamics, communication, and information exchange, factors that are critically important to group decision making (Stasser, 1999), group performance (De Dreu, Nijstad, & van Knippenberg, 2008), and organizational effectiveness (Zaccaro, Rittman, & Marks, 2001)…Prior research has hinted at a potentially negative effect of narcissistic individuals on group and organizational performance. For example, in one study, individuals with high

levels of narcissism allocated more resources to themselves than did individuals with low levels of narcissism—at a long-term cost to other group members (Campbell, Bush, Brunell, & Shelton, 2005). However, prior research did not provide a clear link between leader's narcissism and group or organizational performance.

"How quickly can you detect it? Power facilitates attentional orienting", Slabu et al

Participants were assigned to a high power or control role and then performed a computerised spatial cueing task in which they were required to direct their attention to a target that had been preceded by either a valid or invalid location cue. Compared to participants in the control condition, power-holders were better able to override the misinformation provided by invalid cues. This advantage occurred only at 500 ms stimulus onset asynchrony (SOA), whereas at 1000 ms SOA, when there was more time to prepare a response, no differences were found. These findings are taken to support the growing idea that social power affects cognitive flexibility…Post-test questionnaires confirmed that these effects could not be attributed to differences in positive affect or self-efficacy. We suggest that power most affected performance during invalid trials because these required a greater degree of cognitive flexibility; individuals needed to ignore the cue and unexpectedly orient attention towards the opposite location. In line with this account, the effect was only evident at relatively short SOAs where participants had little time to prepare an appropriate response. At longer SOAs or on valid trials, the need for flexibility was lower which may explain why no effect was seen.

Social power affects the way in which information is attended and discriminated (Fiske, 1993; Guinote, 2007a). Power holders have more resources and fewer constraints which gives them more attentional resources and allows them to discriminate between relevant and irrelevant information (Guinote, 2007a; Overbeck & Park, 2001). In contrast, powerless people face more constraints and environmental threats (Keltner, Gruenfeld, & Anderson, 2003). Their dependency encourages them to attend to multiple cues in the environment, in search of any potentially useful information. Thus, they treat information more equally, attending not only to the central information but also to the peripheral or distracting information (Slabu & Guinote, 2010). This overflow in information processing makes powerless people less able to respond promptly to specific situational demands, and induces attentional inflexibility (Guinote, 2007a).

- Fiske, S. T. (1993). Controlling other people: The impact of power on stereotyping. American Psychologist, 48(6), 621-628. doi: 10.1037/0003-066X.48.6.621
- Guinote, A. (2007a). Behaviour variability and the Situated Focus Theory of Power. European Review of Social Psychology, 18, 256-295. doi: 10.1080/10463280701692813
- Overbeck, J. R., & Park, B. (2001). When power does not corrupt: Superior individuation processes among powerful perceivers. Journal of Personality and Social Psychology, 81(4), 549-565. doi: 10.1037/0022-3514.81.4.549
- Slabu, L., & Guinote, A. (2010). Getting what you want: Power increases the accessibility of active goals. Journal of Experimental Social Psychology, 46(2), 344-349. doi: 10.1016/j.jesp.2009.10.013

Research using basic cognitive paradigms supports these claims. For example, Guinote (2007b) showed that high power participants are better able to focus their attention to target objects and ignore the influence of irrelevant background

distracters (see also Smith & Trope, 2006). A further outcome of the cognitive flexibility experienced by powerful individuals is the increased ability to adjust their actions in line with changing contextual cues. This includes the ability to suppress dominant responses and implement non-dominant ones when the task calls for non-dominant responses (Guinote, 2007b).

- Guinote, A. (2007b). Power affects basic cognition: Increased attentional inhibition and flexibility. Journal of Experimental Social Psychology, 43(5), 685-697. doi: 10.1016/j.jesp.2006.06.008
- Smith, P. K., & Trope, Y. (2006). You focus on the forest when you're in charge of the trees: Power priming and abstract information processing. Journal of Personality and Social Psychology, 90(4), 578-596. doi: 10.1037/0022-3514.90.4.578

For example, several studies have shown that having power increases the ability to resolve conflicts and plan action sequences; power-holders are immune to stimulus-response compatibility effects, and are better able to switch attention between the holistic and detailed components of stimuli, as changing task demands dictate (Guinote, 2007b; Smith, Jostmann, Galinsky, & van Dijk, 2008)… More broadly, our findings build on those reported by Willis, Rodriguez-Bailon and Lupianez (2011) who showed that powerful individuals can make a better use of cues present in the environment to increase their executive control (see also Smith, et al., 2008). Their data support the idea that social power can impact rudimentary processes associated with spatial orienting and control.

- Willis, G. B., Rodríguez-Bailón, R., Lupiáñez, J. (2011). The boss is paying attention: Power Affects the Functioning of the Attentional Networks. Social Cognition, 29(2), 166-181.

"You focus on the forest when you're in charge of the trees: Power priming and abstract information processing", Smith& Trope 2006

Elevated power increases the psychological distance one feels from others, and this distance, according to construal level theory (Y. Trope & N. Liberman, 2003), should lead to more abstract information processing. Thus, high power should be associated with more abstract thinking—focusing on primary aspects of stimuli and detecting patterns and structure to extract the gist, as well as categorizing stimuli at a higher level—relative to low power. In 6 experiments involving both conceptual and perceptual tasks, priming high power led to more abstract processing than did priming low power, even when this led to worse performance. Experiment 7 revealed that in line with past neuropsychological research on abstract thinking, priming high power also led to greater relative right-hemispheric activation.

- Trope, Y., & Liberman, N. (2003). Temporal construal. Psychological Review, 110, 403– 421

Though the abstraction hypothesis has not been directly tested, there is some research that supports it. For example, in Overbeck and Park's (2001) experiments, high- and low-power participants interacted via e-mail with several different targets holding the opposite power role and received various kinds of information from them. Some of this information was relevant to the task at hand (e.g., Jim waited until the last minute to try to schedule a meeting), and some was irrelevant (e.g., Jim just started a jazz ensemble). Not only did participants in the

high-power role recall more information overall than did the low-power participants, but they were especially superior at recalling relevant information. Thus, high-power participants focused more on primary information, a hallmark of abstract thinking.

- Overbeck, J. R., & Park, B. (2001). When power does not corrupt: Superior individuation processes among powerful perceivers. Journal of Personality and Social Psychology, 81, 549 –565.

Portuguese participants used more abstract language to describe both their ethnic group and an outgroup when they were part of the majority (i.e., a higher power group) than when they were part of the minority (i.e., a lower power group; Guinote, 2001). Similarly, participants who played the role of judges during a task used more abstract, trait-like language in referring to themselves than did participants who were workers (Guinote, Judd, & Brauer, 2002).

- Guinote, A. (2001). The perception of group variability in a non-minority and a minority context: When adaptation leads to outgroup differentiation. British Journal of Social Psychology, 40, 117–132.
- Guinote, A., Judd, C. M., & Brauer, M. (2002). Effects of power on perceived and objective group variability: Evidence that more powerful groups are more variable. Journal of Personality and Social Psychology, 82, 708 –721

Powerholders, more than the powerless, should thus be guided by their primary, overriding goals rather than by subordinate, incidental concerns. This would mean that powerholders are more likely to act in accordance with their core attitudes and values (Chen et al., 2001). Indeed, individuals placed in high-power roles or those higher in personality dominance have been found to express their true attitudes more during a discussion than have participants lower in power or dominance (Anderson & Berdahl, 2002). Such goal-driven behavior also has implications for stereotyping. Powerholders should be more likely to stereotype those beneath them when such stereotyping is seen as an effective means to their goals. Evidence for this has already been found in the context of the Social Influence Strategy x Stereotype Match hypothesis (Vescio, Snyder, & Butz, 2003).

- Chen, S., Lee-Chai, A. Y., & Bargh, J. A. (2001). Relationship orientation as a moderator of the effects of social power. Journal of Personality and Social Psychology, 80, 173-187
- Anderson, C., & Berdahl, J. L. (2002). The experience of power: Examining the effects of power on approach and inhibition tendencies. Journal of Personality and Social Psychology, 83, 1362–1377
- Vescio, T. K., Snyder, M., & Butz, D. A. (2003). Power in stereotypically masculine domains: A social influence strategy x stereotype match model. Journal of Personality and Social Psychology, 85, 1062–1078.

"Powerful People Make Good Decisions Even When They Consciously Think", Smith et al 2008

Thought condition again had different effects on performance for the two priming conditions, F(1, 161) 54.67, prep 5 .91, Zp 2 1⁄4 :03 (see Fig. 1). Low-power participants performed significantly better after unconscious thought than after conscious thought, prep 5 .96. High-power participants performed equally well in both thought conditions and did not differ from low-power participants in the unconscious-thought condition, Fs < 1. Furthermore, our manipulations did not

significantly affect participants' confidence in and certainty of their attitudes, preps < .70, their reported effort or motivation, preps < .84, or the amount of apartment information they correctly recalled, Fs < 1. Differences in performance could not be attributed to depth of processing. When given problems requiring a complex decision, high-power participants were equally good at identifying the better choice after conscious versus unconscious thought, whereas the performance of low-power participants suffered when they consciously deliberated. These results provide further evidence that conscious and unconscious thought differ in the type of processing that occurs. The powerful seem to be able to handle so many impactful decisions, without making excessive errors, in part because they generally think more abstractly.

"Cooperation and Status in Organizations", Eckel et al 2010

We further manipulate status by allocating the central position to the person who earns the highest, or the lowest, score on a trivia quiz. These high-status and low-status treatments are compared, and we find that the effect of organizational structure – the existence of a central position – depends on the status of the central player. Higher status players are attended to and mimicked more systematically. Punishment has differential effects in the two treatments, and is least effective in the high-status case.

In this study, we ask whether social status serves as a useful mechanism for solving public goods problems. Status can act as a coordinating device, as it does in pure coordination games, with higher-status individuals more likely to be mimicked (followed) by others. In addition, in a setting with costly punishment, social status may enhance the effectiveness of punishment and reduce anti-social punishment, enhancing overall efficiency…Status is awarded by the experimenter using scores on a general-knowledge trivia quiz that is unrelated to the experimental game. The central position is given to either the high scorer (high-status treatment) or the low scorer (low-status treatment). Subjects play two games: a standard linear voluntary contribution mechanism (VCM) and a VCM with costly punishment. We find that higher-status central players are more likely to be "followed" in the key situation when the peripheral player is contributing less than the central player. We also find that high status central players punish less, and peripheral players are more responsive to punishment by a higher-status central player…Our results suggest that punishment, while important to enforcing cooperative norms in many social dilemmas, does not boost contributions in all instances. Punishment is used more readily by low-status groups, and increases overall contributions only among low-status groups. However this seems to be primarily a main effect of the punishment institution, as there is little evidence that punishment tokens levied actually increase contributions in low-status groups; indeed there is weak evidence that the response to punishment is greater in high-status groups. Retaliatory punishment of central players is seen only in the low-status groups. An unexpected consequence of these differences is that punishment is not efficiency- enhancing when the status of the central player is high. Costly punishment is used less in these groups, but contributions are not higher than without punishment. This generates a flat contribution pattern, and no differences between the VCM with and without punishment opportunities. At the other extreme, low status central players punish and are heavily punished, and make significantly less money in the experiment than any other type of subject. But the reaction of low status groups to the new environment generates a significant increase in the provision of the public good.

Second, high-status agents may have a strong influence on others, as others seek their company and guidance, affecting choices and decision making by lower-status individuals. Thus high-status individuals are more likely to be mimicked or deferred to (Ball et al. 2001, Kumru and Vesterlund 2005). Imitating or learning from higher- status exemplars can help solve coordination problems (Eckel and Wilson 2007); the behavior of the higher-status individual provides an example that is observed and can be followed by others.

- Ball, S., C. Eckel, P. Grossman and W. Zame (2001) “Status in markets” The Quarterly Journal of Economics 116, 161-188
- Kumru, C. and L. Vesterlund (2005) “The effect of status on voluntary contribution” Working paper, Department of Economics, University of Pittsburgh.
- Eckel, C. and R. Wilson (2007) “Social learning in coordination games: Does status matter?” Experimental Economics, 10, 317-330

Gil-White and Henrich (2001) argue that attending to and mimicking high status individuals is a valuable strategy in a world where successful individuals may have superior information. Cultural transmission is enhanced when higher-status, successful individuals are copied by others. Copying successful individuals has evolutionary payoffs, so that humans may have evolved a preference for paying attention to and learning from high-status agents (see also Boyd and Richerson 2002, Boyd et al. 2003). Bala and Goyal (1998) capture the essence of the idea of attending to a high-status agent in a model where the presence of a commonly-observed agent, which they term the “royal family”, can have a significant impact on which among multiple equilibria is selected…Experimental research confirms the tendency of individuals to mimic high-status agents. Eckel and Wilson (2001) show that a commonly observed agent can influence equilibrium selection in a coordination game…Imitation makes the population of subjects more likely to reach a Pareto-superior, but risk- dominated, equilibrium, an outcome that rarely occurs otherwise (Cooper et al. 1990). Kumru and Vesterlund (2005) show a related result, with high-status first-movers more likely to be mimicked in a 2-person sequential voluntary contribution game. In their setting, high status enhances the ability of leaders to increase total contributions.

- Gil-White, F. and J. Henrich (2001) “The evolution of prestige: Freely conferred deference as a mechanism for enhancing the benefits of cultural transmission” Evolution and Human Behavior 22,165-196
- Boyd, R., and P. Richerson (2002) “Group beneficial norms spread rapidly in a structured population” Journal of Theoretical Biology 215, 287-296
- Boyd, R., H. Gintis, S. Bowles, and P. Richerson (2003) “The evolution of altruistic punishment” Proceedings of the National Academy of Sciences (USA) 100, 3531-3535.
- Bala, V. and S. Goyal (1998) “Learning from neighbors” Review of Economic Studies 65, 595-621
- Eckel, Catherine C., and Rick K. Wilson (2001) “Social learning in a social hierarchy: An experimental study.” Rice University, Unpublished manuscript
- Cooper, R., D. DeJong, R. Forsythe and T. Ross (1990). “Selection criteria in coordination games: some experimental results. American Economic Review 80, 218-233
- Kumru, C. and L. Vesterlund (2005) “The effect of status on voluntary contribution” Working paper, Department of Economics, University of Pittsburgh

Another good set of studies focusing on rich/powerful behavior.

2 of the primary researchers write in a 2012 *NYT* op-ed "Greed Prevents Good"

> Now, some 25 years later, seven studies we conducted [Piff et al 2012], some on this same campus, have proved the opposite, that greed, far from being good, undermines moral behavior.…Unethical behaviors among the wealthy are as timeless and pervasive as the ethical principles that try to rein them in. Our research pinpointed why wealth produces unethical conduct with such regularity: greed. Across studies, wealthier subjects expressed the conviction that greed is moral, echoing [Ivan] Boesky and Gekko and their intellectual companions (e.g., Ayn Rand). And it was their greed-is-good attitudes, we found, that gave rise to their unethical behavior. Wealth gives rise to a me-first mentality, and the ideology of unbridled self-interest serves as its lofty justification. Greg Smith is to be applauded for calling out the culture of greed at Goldman Sachs. It is a knockout blow, one as important as Ivan Boesky's proclamation nearly a generation ago. Nobel laureate Milton Friedman famously argued that the single social responsibility of business is to increase profits as long as "it stays within the rules of the game." The problem is, when greed for profits is the bottom line, the rules may fall by the wayside.

Relevant studies:

- Kraus & Keltner 2009, "Signs of socioeconomic status: a thin-slicing approach":

  > Videos of 60-s slices of these interactions were coded for nonverbal cues of disengagement and engagement, and estimates of participants' SES were provided by naive observers who viewed these videos. As predicted by analyses of resource dependence and power, upper-SES participants displayed more disengagement cues (e.g., doodling) and fewer engagement cues (e.g., head nods, laughs) than did lower-SES participants.…Research relevant to this hypothesis is limited, but suggestive. For example, in a meta-analytic review of status and nonverbal behavior, upper SES individuals were found to speak in ways that are less attentive to the audience, for example, with fewer turn-inviting pauses (Hall et al., 2005)… SES was measured objectively using self-reports of family income and education (e.g., Lachman & Weaver, 1998). [They used undergraduates, *not* people who had personally clawed into power.]

  Consistent with the previously cited studies about how acting rude or defecting is perceived as power.

- Kraus et al 2010 "Social Class, Contextualism, and Empathic Accuracy":

  > Recent research suggests that lower-class individuals favor explanations of personal and political outcomes that are oriented to features of the external environment. We extended this work by testing the hypothesis that, as a result, individuals of a lower social class are more empathically accurate in judging the emotions of other people. In three studies, lower-class individuals (compared with upper-class individuals) received higher scores on a test of empathic accuracy (Study 1), judged the emotions of an interaction partner more accurately (Study 2), and made more accurate inferences about emotion from static images of muscle movements in the eyes (Study 3). Moreover, the association between social class and empathic

accuracy was explained by the tendency for lower-class individuals to explain social events in terms of features of the external environment.

See the previous discussions of blame, self-centeredness, lack of empathy, and rule-breaking; related: fundamental attribution bias.

- Stellar et al 2012, "Class and compassion: socioeconomic factors predict responses to suffering":

  Previous research indicates that lower-class individuals experience elevated negative emotions as compared with their upper-class counterparts. We examine how the environments of lower-class individuals can also promote greater compassionate responding-that is, concern for the suffering or well-being of others. In the present research, we investigate class-based differences in dispositional compassion and its activation in situations wherein others are suffering. Across studies, relative to their upper-class counterparts, lower-class individuals reported elevated dispositional compassion (Study 1), as well as greater self-reported compassion during a compassion-inducing video (Study 2) and for another person during a social interaction (Study 3). Lower-class individuals also exhibited heart rate deceleration-a physiological response associated with orienting to the social environment and engaging with others-during the compassion-inducing video (Study 2)…For example, when describing environmental trends in economic inequality and everyday life outcomes (e.g., being laid off from work), undergraduates of lower subjective socioeconomic status—measured by ranking oneself in society in terms of income, education, and job status relative to others—attribute the causes of economic inequality to more external reasons (e.g., political influence, educational opportunity) than dispositional reasons (e.g., hard work, talent), relative to their upper-class counterparts (Kraus et al., 2009)…Converging evidence also suggests that lower-class individuals favor an interdependent view of the self, whereas upper-class individuals are more inclined to espouse beliefs in an individuals' independence and autonomy (Stephens, Fryberg, & Markus, 2011; Stephens, Markus, & Townsend, 2007). For instance, in one study lower-class university students, whose parents' highest level of education was a high school diploma, tended to make choices that helped them blend in with others (e.g., by choosing a pen that resembled other pens; Stephens et al., 2007). In contrast, upper-class individuals, whose parents graduated from college, tended to prefer choices that helped them stand out (e.g., by choosing a unique pen). In recent work, Stephens and colleagues (2011) suggest that stronger relational norms among working-class individuals result in a less positive perception of individual choice, which favors an individual's own needs.

- Piff et al 2012, "Higher social class predicts increased unethical behavior":

  In studies 1 and 2, upper-class individuals were more likely to break the law while driving, relative to lower-class individuals. In follow-up laboratory studies, upper-class individuals were more likely to exhibit unethical decision-making tendencies (study 3), take valued goods from others (study 4), lie in a negotiation (study 5), cheat to increase their chances of winning a prize (study 6), and endorse unethical behavior at work (study 7) than were lower-class individuals. Mediator and moderator data demonstrated that upper-class individuals' unethical tendencies are accounted for, in part, by

their more favorable attitudes toward greed…Individuals from upper-class backgrounds are also less generous and altruistic. In one study, upper-class individuals proved more selfish in an economic game, keeping significantly more laboratory credits—which they believed would later be exchanged for cash—than did lower-class participants, who shared more of their credits with a stranger (7). These results parallel nationwide survey data showing that upper-class households donate a smaller proportion of their incomes to charity than do lower-class households (10)…Research finds that individuals motivated by greed tend to abandon moral principles in their pursuit of self-interest (13). In one study, a financial incentive caused people to be more willing to deceive and cheat others for personal gain (14). In another study, the mere presence of money led individuals to be more likely to cheat in an anagram task to receive a larger financial reward (1)…Why are upper-class individuals more prone to unethical behavior, from violating traffic codes to taking public goods to lying? This finding is likely to be a multiply determined effect involving both structural and psychological factors. Upper-class individuals' relative independence from others and increased privacy in their professions (3) may provide fewer structural constraints and decreased perceptions of risk associated with committing unethical acts (8). The availability of resources to deal with the downstream costs of unethical behavior may increase the likelihood of such acts among the upper class. In addition, independent self-construals among the upper class (22) may shape feelings of entitlement and inattention to the consequences of one's actions on others (23). A reduced concern for others' evaluations (24) and increased goal-focus (25) could further instigate unethical tendencies among upper-class individuals. Together, these factors may give rise to a set of culturally shared norms among upper-class individuals that facilitates unethical behavior.

- 7: Piff PK, Kraus MW, Côté S, Cheng BH, Keltner D (2010) "Having less, giving more: The influence of social class on prosocial behavior". J Pers Soc Psychol 99:771–784.
- 10: Independent Sector (2002) *Giving and Volunteering in the United States* (Independent Sector, Washington, DC).
- 13: Steinel W, De Dreu CKW (2004) "Social motives and strategic misrepresentation in social decision making". J Pers Soc Psychol 86:419–434
- 14: Aquino K, Freeman D, Reed A, II, Felps W, Lim VK (2009) "Testing a social-cognitive model of moral behavior: The interactive influence of situations and moral identity centrality". J Pers Soc Psychol 97:123–141
- 1: Gino F, Pierce L (2009) "The abundance effect: Unethical behavior in the presence of wealth". Organ Behav Hum Dec 109:142–155

Where to Intervene in a Human?

The "What is Rationality?" page on the new CFAR website contains an illuminating story about Intel:

> Semiconductor giant Intel was originally a memory chip manufacturer. But by 1985, memory chips had been losing them money for years. Co-founders Andy Grove (CEO) and Gordon Moore met to discuss the problem. The mood was grim. At one point, Andy turned to Gordon and asked, “If we get kicked out and the board brings in a new CEO, what do you think he would do?”
>
> Gordon replied without hesitation. “He would get us out of the memory business.”
>
> “Okay,” said Andy. “Then why shouldn’t you and I walk out the door, come back, and do it ourselves?”
>
> That year, Andy and Gordon shifted the focus of the company to microprocessors, and created one of the greatest success stories in American business.

I presume Andy and Gordon had considered intervening at many different levels of action: in middle management, in projects, in products, in details, etc. They had probably implemented some of these plans, too. But the problem with Intel — it was in the wrong market! — was so deep that the place to intervene was at a very low level, the foundations of the entire company. It's possible that in this situation, *no* change they could have made at higher levels of action would have made that big of a difference compared to *changing the company's market and mission*.

In 1997, system analyst Donella Meadows wrote Places to Intervene in a System, in which she outlined twelve leverage points at which one could intervene in a system. Different levels of action, she claimed, would have effects of different magnitudes.

This got me thinking about levels of action and self-improvement. "I want to improve myself: where should I intervene in my own system *next*?"

My bet is that if the next greatest leverage point you can push on is something like neurofeedback, then you're pretty damn self-optimized already.

In fact, I suspect *almost nobody* is that self-optimized. We do things like neurofeedback because (1) we don't think enough about choosing the highest-leverage self-interventions, (2) in any case, we don't know how to figure out which interventions would be higher leverage for ourselves, (3) even if there are higher-leverage interventions to be had, we might not successfully carry them through, but neurofeedback or whatever happens to be fun and engaging for us, and (3) sometimes, you gotta stop analyzing your situation and *just do some stuff that looks like it might help*.

Anyway, how can one figure out what the next highest-leverage self-interventions are for oneself? Maybe I just haven't yet found the right keywords, but I don't think there's been much research on this topic.

Intuitively, it seems like hacking one's motivational system is among the highest leverage interventions one can make, because high motivation allows on to carry

through with lots of other interventions, and without sufficient motivation one *can't* follow through with many interventions.

But if you've got a crippling emotional or physical condition, I suppose you've got to take care of that first — at least well enough to embark on the project of hacking your motivation system.

Or, if you're in a crippling *environment* like North Korea or Nigeria or Detroit, then perhaps the highest level intervention for you is to get up and *move someplace better*. Only *then* will you be able to fix your emotions or hack your motivational system or whatever.

Maybe there's something of a system to this that hasn't been discovered, or maybe there's no system at all because humans are too complex. I'm still in brainstorm mode on this topic.

- What do *you* think are some generally highest-level self-improvement interventions that more people should be tackling before things like neurofeedback?
- What algorithm could be used for discovering the next best intervention one can make to improve oneself?
- Has there been any research on this issue?

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.

Article about LW: Faith, Hope, and Singularity: Entering the Matrix with New York's Futurist Set

Faith, Hope, and Singularity: Entering the Matrix with New York's Futurist Set

To my knowledge LessWrong hasn't received a great deal of media coverage. So, I was surprised when I came across an article via a Facebook friend which also appeared on the cover of the New York Observer today. However, I was disappointed upon reading it, as I don't think it is an accurate reflection of the community. It certainly doesn't reflect my experience with the LW communities in Toronto and Waterloo.

I thought it would be interesting to see what the broader LessWrong community thought about this article. I think it would make for a good discussion.

Possible conversation topics:

- This article will likely reach many people that have never heard of LessWrong before. Is this a good introduction to LessWrong for those people?
- Does this article give an accurate characterization of the LessWrong community?

Edit 1: Added some clarification about my view on the article.

Edit 2: Re-added link using "nofollow" attribute.

Exploiting the Typical Mind Fallacy for more accurate questioning?

I was reading Yvain's Generalizing from One Example, which talks about the typical mind fallacy. Basically, it describes how humans assume that all other humans are like them. If a person doesn't cheat on tests, they are more likely to assume others won't cheat on tests either. If a person sees mental images, they'll be more likely to assume that everyone else sees mental images.

As I'm wont to do, I was thinking about how to make that theory pay rent. It occurred to me that this could definitely be exploitable. If the typical mind fallacy is correct, we should be able to have it go the other way; we can derive information about a person's proclivities based on what they think about other people.

Eg, most employers ask "have you ever stolen from a job before," and have to deal with misreporting because nobody in their right mind will say yes. However, imagine if the typical mind fallacy was correct. The employers could instead ask "what do you think the percentage of employees who have stolen from their job is?" and know that the applicants who responded higher than average were correspondingly more likely to steal, and the applicants who responded lower than average were less likely to cheat. It could cut through all sorts of social desirability distortion effects. You couldn't get the exact likelihood, but it would give more useful information than you would get with a direct question.

In hindsight, which is always 20/20, it seems incredibly obvious. I'd be surprised if professional personality tests and sociologists *aren't* using these types of questions. My google-fu shows no hits, but it's possible I'm just not using the correct term that sociologists use. I'm was wondering if anyone had heard of this questioning method before, and if there's any good research data out there showing just how much you can infer from someone's deviance from the median response.

Revisiting SI's 2011 strategic plan: How are we doing?

Progress updates are nice, but without a previously defined metric for success it's hard to know whether an organization's achievements are noteworthy or not. Is SI making good progress, or underwhelming progress?

Luckily, in August 2011 we published a strategic plan that outlined lots of specific goals. It's now almost August 2012, so we can check our progress against the standard set nearly one year ago. The plan doesn't specify a timeline for the stated goals, but I remember hoping that we could do most of them by the end of 2012, while understanding that we should list more goals than we could actually accomplish given current resources.

Let's walk through the goals in that strategic plan, one by one. (Or, you can skip to the "summary and path forward" section at the end.)

1.1. Clarify the open problems relevant to our core mission

This was accomplished to some degree with So You Want to Save the World, and is on track to be accomplished to a greater degree with Eliezer's sequence "Open Problems in Friendly AI," which you should begin seeing late in August.

1.2. Identify and recruit researcher candidates who can solve research problems.

Several strategies for doing this were listed, but the only one worth doing at our current level of funding was to recruit more research associates and hire more researchers. Since August 2011 we have done both, adding half a dozen research associates and hiring nearly a dozen remote researchers, including a few who are working full-time on papers and other projects (e.g. Kaj Sotala).

1.3. Use researchers and research associates to solve open problems related to Friendly AI theory.

I never planned to be doing this by the end of 2012; it's more of a long-term goal. A first step in this direction is to have Eliezer transition back to FAI work, e.g. with his "Open Problems in Friendly AI" Summit 2011 talk and forthcoming blog sequence. And actually, SI research associate Vladimir Slepnev has been making interesting progress in LW-style decision theory, and is working on a paper explicating one of his results. (Some credit is due to Vladimir Nesov and others.)

1.4. Estimate current AI risk levels.

Alas, we haven't done much of this. There's some analysis in Intelligence Explosion: Evidence and Import, Reply to Holden on Tool AI, and Reply to Holden on The Singularity Institute. Also, Anna is working on a simple model of AI risk in MATLAB (or some similar program). But I would have liked to have the cash to hire a researcher to continue things like AI Risk and Opportunity: A Strategic Analysis.

2.1. Continue operation of the Singularity Summit, which is beginning to yield a profit while also reaching more people with our message.

We did run Singularity Summit 2011, and Singularity Summit 2012 is on track to be noticeably more fun and professional than all past Summits. (So, register now!)

The strategic plan listed subgoals of gaining corporate sponsors and *possibly* expanding the Summit outside the USA. We gained corporate sponsors for Summit 2011, and are on track to gain even more of them for Summit 2012. Early in 2011 we also pursued an opportunity to host the first Singularity Summit in Europe, but the financing didn't quite come through.

2.2 Cultivate LessWrong.com and the greater rationality community as a resource for Singularity Institute.

The strategic plan lists 5 subgoals, all of which we achieved. SI (a) used LessWrong to recruit additional supporters, (b) made use of LessWrong for collaborative problem solving (e.g. this and this), (c) published lots of top-level posts, (d) and published *How to Run a Successful Less Wrong Meetup Group*. The early efforts of CFAR, and our presence at (e.g.) Skepticon IV, made headway on 2.2.e: "Encourage improvements in critical thinking in the wider world. We need a larger community of critical thinkers for use in recruiting, project implementation, and fundraising."

2.3. Spread our message and clarify our arguments with public-facing academic deliverables.

We did exceptionally well on this, though much more is needed. In addition to detailed posts like Reply to Holden on Tool AI and Reply to Holden on the Singularity Institute, SI has more peer-reviewed publications in 2012 than in all past years combined.

2.4. Build more relationships with the optimal philanthropy, humanist, and critical thinking communities, which share many of our values.

Though this work has been mostly invisible, Carl Shulman has spent dozens of hours on building relationships with the optimal philanthropy community. We've also built relationships with the humanist and critical thinking communities, through our presence at Skepticon IV but especially through the early activities of CFAR.

2.5. Cultivate and expand Singularity Institute's Volunteer Program.

SI's volunteer program got a new website (though we'd like to launch another redesign soon), and we estimate that SI volunteers have done 2x-5x more work per month this year than in the past few years.

2.6. Improve Singularity Institute's web presence.

Done. We got a new domain, Singularity.org, and put up a new website there. We produced additional introductory materials, like Friendly-AI.com and IntelligenceExplosion.com. We produced lots of "landing pages," for example our tech summaries. We did not, however, complete subgoals (d) and (e) — "Continue to produce articles on targeted websites and other venues" and "Produce high-quality videos to explain Singularity Institute's mission" — because their ROI isn't high enough to do at our current funding level.

2.7. Apply for grants, especially ones that are given to other organizations and researchers concerned with the safety of future technologies (e.g. synthetic biology and nanotechnology).

This one was always meant as a longer-range goal. SI still needs to be "fixed up" in certain ways before this is worth trying.

2.8. Continue targeted interactions with the public.

We didn't do much of this, either. In particular, Eliezer's rationality books are on hold for now; we have the author of a best-selling science book on retainer to take a crack at Eliezer's rationality books this fall, after he completes his current project.

2.9. Improve interactions with current and past donors.

Success. We created and cleaned up our donor database, communicated more regularly with our support base (previously via monthly updates and now our shiny new newsletter, which you can sign up for here), and updated our top donors list.

3.1. Encourage a new organization to begin rationality instruction similar to what Singularity Institute did in 2011 with Rationality Minicamp and Rationality Boot Camp.

This is perhaps the single most impressive thing we did this year, in the sense that it required dozens of smaller pieces to all work, and work together. The organization is now called the Center for Applied Rationality (CFAR), and it was recently approved for 501c3 status. It has its own website, has been running extremely well-reviewed rationality retreats, and has lots more exciting stuff going on that hasn't been described online yet. **Sign up for CFAR's newsletter** to get these juicy details when they are written up.

3.2. Use Charity Navigator's guidelines to improve financial and organizational transparency and efficiency.

There are 9 subgoals listed here. We've since decided we *don't* want to grow to five independent board members (subgoal b) at this time, because a smaller board runs more efficiently. (I've now heard too many nightmare stories about trying to get things done with a large board.) We *did* achieve (a), (d), (e), (g), (h), and (i). Subgoal (c) is a longer term goal that we are working toward (we need a professional bookkeeper to clean up our internal processes before we can have a hired CPA audit, and we're interviewing bookkeepers now). Subgoal (f) — a records retention policy — is in the works.

3.3. Ensure a proper orientation for new Singularity Institute staff and visiting fellows.

This is in process; we're creating orientation materials.

3.4. Secure lines of credit to increase liquidity and smooth out the recurring cash-flow pinches that result from having to do things like make payroll and rent event spaces.

We've done this.

3.5. Improve safe return on financial reserves

For starters, we put a large chunk of our resources in an ING Direct high-interest savings account.

3.6. Ensure high standards for staff effectiveness.

There are two subgoals here. Subgoal (b) was to have staff maintain work logs, which we've been doing for many months now. Subgoal (a) is more ambiguous. We haven't given people job descriptions because at such a small organization, such roles change quickly. But I do provide stronger management of SI staff and projects than ever before, and this clarifies the expectations for our staff, often including task and project deadlines.

3.7. When hiring, advertise for applications to find the best candidates.

We've been doing this for several months now, e.g. here and here.

Summary

That's it for the main list! Now let's check in on what we said **our top priorities for 2011-2012** were:

1. *Public-facing research on creating a positive singularity*. Check. SI has more peer-reviewed publications in 2012 than in all past years combined.
2. *Outreach / education / fundraising*. Check. Especially, through CFAR.
3. *Improved organizational effectiveness*. Check. Lots of good progress on this.
4. *Singularity Summit*. Check.

In summary, I think SI is a bit behind where I hoped we'd be by now, though this is largely because we've poured so much into launching CFAR, and as a result, CFAR has turned out to be significantly more cool at launch than I had anticipated.

Fundraising has been a challenge. One donor failed to actually give their $46,000 pledge despite repeated reminders and requests, and our support base is (understandably) anxious to see a shift from movement-building work to FAI research, a shift I have been fighting for since I was made Executive Director. (Note that spinning off rationality work to CFAR is a substantial part of trimming SI down into being primarily an FAI research institute.)

Reforming SI into a more efficient, effective organization has been my greatest challenge. Frankly, SI was in pretty bad shape when Louie and I arrived as interns in April 2011, and there have been an incredible number of holes to dig SI out of — and several more remain. (In contrast, it has been a *joy* to help set up CFAR properly *from the very beginning*, with all the right organizational tools and processes in place.) Reforming SI presents a fundraising problem, because reforming SI is time consuming and sometimes costly, but is generally unexciting to donors. I can see the light at the end of the tunnel, though. We won't reach it if we can't improve our fundraising success in the next 3-6 months, but it's close enough that I can see it.

SI's path forward, from my point of view, looks like this:

1. We finish launching CFAR, which takes over the rationality work SI was doing. (Before January 2013.)
2. We change how the Singularity Summit is planned and run so that it pulls our core staff away from core mission work to a lesser degree. (Before January 2013.)
3. Eliezer writes the "Open Problems in Friendly AI" sequence. (Before January 2013.)
4. We hire 1-2 researchers to produce technical write-ups from Eliezer's TDT article and from his "Open Problems in Friendly AI" sequence. (Beginning September 2012, except that right now we don't have the cash to hire the 1-2 people who I know who *could* do this and who *want* to do this as soon as we have the money to hire them.)
5. With the "Open FAI Problems" sequence and the technical write-ups in hand, we greatly expand our efforts to show math/compsci researchers that there is a tractable, technical research program in FAI theory, and as a result some researchers work on the sexiest of these problems from their departments, and some other math researchers take more seriously the prospect of being *hired* by SI to do technical research in FAI theory. (Beginning, roughly, in April 2013.) Also: There won't be classes on x-risk at SPARC (rationality camp for young elite math talent), but some SPARC students might end up being interested in FAI stuff by osmosis.

6. With a more tightly honed SI, improved fundraising practices, and visible mission-central research happening, SI is able to attract more funding and hire even more FAI researchers. (Beginning, roughly, in September 2013.)

If you want to help us make this happen, please donate during our July matching drive!

CFAR website launched

The new Center for Applied Rationality website has launched! We'll be adding content as time goes by. Let us know if you find broken links, etc.