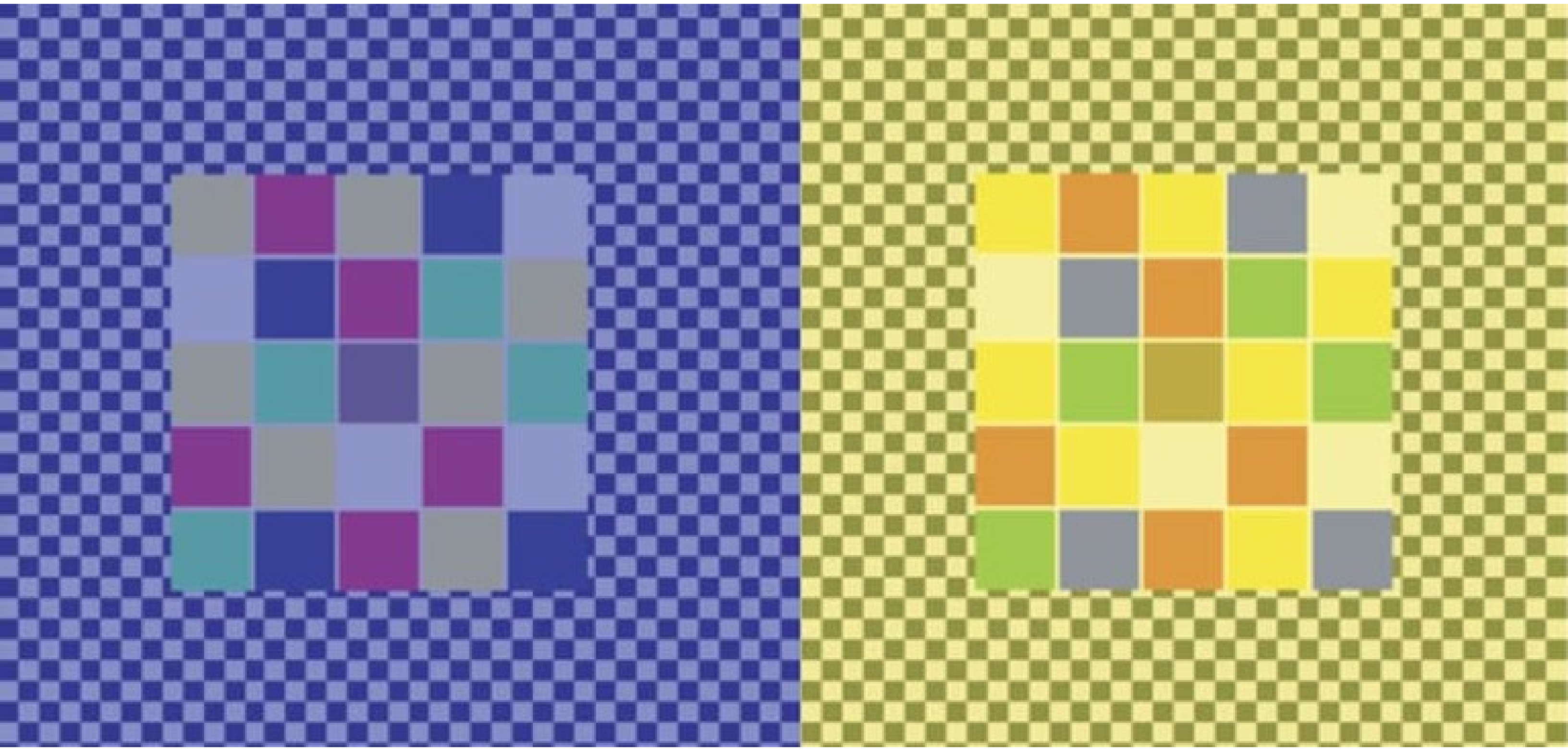


The Blue-Minimizing Robot

The Blue-Minimizing Robot

Imagine a robot with a turret-mounted camera and laser. Each moment, it is programmed to move forward a certain distance and perform a sweep with its camera. As it sweeps, the robot continuously analyzes the average RGB value of the pixels in the camera image; if the blue component passes a certain threshold, the robot stops, fires its laser at the part of the world corresponding to the blue area in the camera image, and then continues on its way.

Watching the robot's behavior, we would conclude that this is a robot that destroys blue objects. Maybe it is a surgical robot that destroys cancer cells marked by a blue dye; maybe it was built by the Department of Homeland Security to fight a group of terrorists who wear blue uniforms. Whatever. The point is that we would analyze this robot in terms of its goals, and in those terms we would be tempted to call this robot a blue-minimizer: a machine that exists solely to reduce the amount of blue objects in the world.

Suppose the robot had human level intelligence in some side module, but no access to its own source code; that it could learn about itself only through observing its own actions. The robot might come to the same conclusions we did: that it is a blue-minimizer, set upon a holy quest to rid the world of the scourge of blue objects.

But now stick the robot in a room with a hologram projector. The hologram projector (which is itself gray) projects a hologram of a blue object five meters in front of it. The robot's camera detects the projector, but its RGB value is harmless and the robot does not fire. Then the robot's camera detects the blue hologram and zaps it. We arrange for the robot to enter this room several times, and each time it ignores the projector and zaps the hologram, without effect.

Here the robot is failing at its goal of being a blue-minimizer. The *right* way to reduce the amount of blue in the universe is to destroy the projector; instead its beams flit harmlessly through the hologram.

Again, give the robot human level intelligence. Teach it exactly what a hologram projector is and how it works. Now what happens? Exactly the same thing - the robot executes its code, which says to scan the room until its camera registers blue, then shoot its laser.

In fact, there are many ways to subvert this robot. What if we put a lens over its camera which inverts the image, so that white appears as black, red as green, blue as yellow, and so on? The robot will not shoot us with its laser to prevent such a violation (unless we happen to be wearing blue clothes when we approach) - its entire program was detailed in the first paragraph, and there's nothing about resisting lens alterations. Nor will the robot correct itself and shoot only at objects that appear yellow - its entire program was detailed in the first paragraph, and there's nothing about correcting its program for new lenses. The robot will continue to zap objects that register a blue RGB value; but now it'll be shooting at anything that is yellow.

The human-level intelligence version of the robot will notice its vision has been inverted. It will know it is shooting yellow objects. It will know it is failing at its original goal of blue-minimization. And maybe if it had previously decided it was on a holy quest to rid the world of blue, it will be deeply horrified and ashamed of its actions. It will wonder why it has suddenly started to deviate from this quest, and why it just can't work up the will to destroy blue objects anymore.

The robot goes to Quirinus Quirrell, who explains that robots don't really care about minimizing the color blue. They only care about status and power, and pretend to care about minimizing blue in order to impress potential allies.

The robot goes to Robin Hanson, who explains that there are really multiple agents within the robot. One of them wants to minimize the color blue, the other wants to minimize the color yellow. Maybe the two of them can make peace, and agree to minimize yellow one day and blue the next?

The robot goes to Anna Salamon, who explains that robots are not automatically strategic, and that if it wants to achieve its goal it will have to learn special techniques to keep focus on it.

I think all of these explanations hold part of the puzzle, but that the most fundamental explanation is that the mistake began as soon as we started calling it a "blue-minimizing robot". This is not because its utility function doesn't exactly correspond to blue-minimization: even if we try to assign it a ponderous function like "minimize the color represented as blue within your current visual system, except in the case of holograms" it will be a case of overfitting a curve. The robot is not maximizing or minimizing anything. It does exactly what it says in its program: find something that appears blue and shoot it with a laser. If its human handlers (or itself) want to interpret that as goal directed behavior, well, that's their problem.

It may be that the robot was created to achieve a specific goal. It may be that the Department of Homeland Security programmed it to attack blue-uniformed terrorists who had no access to hologram projectors or inversion lenses. But to assign the goal of "blue minimization" to the robot is a confusion of levels: this was a goal of the Department of Homeland Security, which became a lost purpose as soon as it was represented in the form of code.

The robot is a behavior-executor, not a utility-maximizer.

In the rest of this sequence, I want to expand upon this idea. I'll start by discussing some of the foundations of behaviorism, one of the earliest theories to treat people as behavior-executors. I'll go into some of the implications for the "easy problem" of consciousness and philosophy of mind. I'll very briefly discuss the philosophical debate around eliminativism and a few eliminativist schools. Then I'll go into why we feel like we have goals and preferences and what to do about them.

Basics of Animal Reinforcement

Behaviorism historically began with Pavlov's studies into classical conditioning. When dogs see food they naturally salivate. When Pavlov rang a bell before giving the dogs food, the dogs learned to associate the bell with the food and salivate even after they merely heard the bell . When Pavlov rang the bell a few times without providing food, the dogs stopped salivating, but when he added the food again it only took a single trial before the dogs "remembered" their previously conditioned salivation response[1].

So much for classical conditioning. The real excitement starts at operant conditioning. Classical conditioning can only activate reflexive actions like salivation or sexual arousal; operant conditioning can produce entirely new behaviors and is most associated with the idea of "reinforcement learning".

Serious research into operant conditioning began with B.F. Skinner's work on pigeons. Stick a pigeon in a box with a lever and some associated machinery (a "Skinner box"[2]). The pigeon wanders around, does various things, and eventually hits the lever. Delicious sugar water squirts out. The pigeon continues wandering about and eventually hits the lever again. Another squirt of delicious sugar water. Eventually it percolates into its tiny pigeon brain that maybe pushing this lever makes sugar water squirt out. It starts pushing the lever more and more, each push continuing to convince it that yes, this is a good idea.

Consider a second, less lucky pigeon. It, too, wanders about in a box and eventually finds a lever. It pushes the lever and gets an electric shock. Eh, maybe it was a fluke. It pushes the lever again and gets another electric shock. It starts thinking "Maybe I should stop pressing that lever." The pigeon continues wandering about the box doing anything and everything other than pushing the shock lever.

The basic concept of operant conditioning is that an animal will repeat behaviors that give it reward, but avoid behaviors that give it punishment[3].

Skinner distinguished between primary reinforcers and secondary reinforcers. A primary reinforcer is hard-coded: for example, food and sex are hard-coded rewards, pain and loud noises are hard-coded punishments. A primary reinforcer can be linked to a secondary reinforcer by classical conditioning. For example, if a clicker is clicked just before giving a dog a treat, the clicker itself will eventually become a way to reward the dog (as long as you don't use the unpaired clicker long enough for the conditioning to suffer extinction!)

Probably Skinner's most famous work on operant conditioning was his study of reinforcement schedules: that is, if pushing the lever only gives you reward some of the time, how obsessed will you become with pushing the lever?

Consider two basic types of reward: interval, in which pushing the lever gives a reward only once every t seconds - and ratio, in which pushing the lever gives a reward only once every x pushes.

Put a pigeon in a box with a lever programmed to only give rewards once an hour, and the pigeon will wise up pretty quickly. It may not have a perfect biological clock, but after somewhere around an hour, it will start pressing until it gets the reward and then give up for another hour or so. If it doesn't get its reward after an hour, the behavior will go extinct pretty quickly; it realizes the deal is off.

Put a pigeon in a box with a lever programmed to give one reward every one hundred presses, and again it will wise up. It will start pressing more on the lever when the reward is close (pigeons are better counters than you'd think!) and ease off after it obtains the reward. Again, if it doesn't get its reward after about a hundred presses, the behavior will become extinct pretty quickly.

To these two basic schedules of fixed reinforcement, Skinner added variable reinforcement: essentially the same but with a random factor built in. Instead of giving a reward once an hour, the pigeon may get a reward in a randomly chosen time between 30 and 90 minutes. Or instead of giving a reward every hundred presses, it might take somewhere between 50 and 150.

Put a pigeon in a box on variable interval schedule, and you'll get constant lever presses and good resistance to extinction.

Put a pigeon in a box with a variable ratio schedule and you get a situation one of my professors unscientifically but accurately described as "pure evil". The pigeon will become obsessed with pecking as much as possible, and really you can stop giving rewards at all after a while and the pigeon will never wise up.

Skinner was not the first person to place an animal in front of a lever that delivered reinforcement based on a variable ratio schedule. That honor goes to Charles Fey, inventor of the slot machine.

So it looks like some of this stuff has relevance for humans as well[4]. Tomorrow: more freshman psychology lecture material. Hooray!

FOOTNOTES

1. Of course, it's not really psychology unless you can think of an unethical yet hilarious application, so I refer you to Plaud and Martini's study in which slides of erotic stimuli (naked women) were paired with slides of non-erotic stimuli (penny jars) to give male experimental subjects a penny jar fetish; this supports a theory that uses chance pairing of sexual and non-sexual stimuli to explain normal fetish formation.

2. The bizarre rumor that B.F. Skinner raised his daughter in a Skinner box is completely false. The rumor that he marketed a child-rearing device called an "Heir Conditioner" is, remarkably, true.

3: In technical literature, behaviorists actually use four terms: positive reinforcement, positive punishment, negative reinforcement, and negative punishment. This is really confusing: "negative reinforcement" is actually a type of reward, behavior like going near wasps is "punished" even though we usually use "punishment" to mean deliberate human action, and all four terms can be summed up under the category "reinforcement" even though reinforcement is also sometimes used to mean "reward as opposed to punishment". I'm going to try to simplify things here by using "positive reinforcement" as a synonym for "reward" and "negative reinforcement" as a synonym for "punishment", same way the rest of the non-academic world does it.

4: Also relevant: checking HP:MoR for updates is variable interval reinforcement. You never know when an update's coming, but it doesn't come faster the more times you reload fanfiction.net. As predicted, even when Eliezer goes weeks without updating, the behavior continues to persist.

Basics of Human Reinforcement

Today: some more concepts from reinforcement learning and some discussion on their applicability to human behavior.

For example: most humans do things even when they seem unlikely to result in delicious sugar water. Is this a violation of behaviorist principles?

No. For one thing, yesterday's post included a description of secondary reinforcers, those reinforcers which are not hard-coded evolutionary goods like food and sex, but which nevertheless have a conditioned association with good things. Money is the classic case of a secondary reinforcer among humans. Little colored rectangles are not naturally reinforcing, but from a very young age most humans learn that they can be used to buy pleasant things, like candy or toys or friends. Behaviorist-inspired experiments on humans often use money as a reward, and have yet to run into many experimental subjects whom it fails to motivate[1].

Speaking of friends, status may be a primary reinforcer specific to social animals. I don't know if being able to literally feel reinforcement going on is a real thing, but I maintain I can feel the rush of reward when someone gives me a compliment. If that's too unscientific for you, consider studies in which monkeys will "exchange" sugary juice for the opportunity to look at pictures of high status monkeys, but demand extra juice in exchange for looking at pictures of low status monkeys.

Although certain cynics might consider money and status an exhaustive list, we may also add moral, aesthetic, and value-based considerations. Evolutionary psychology explains why these might exist and Bandura called some of them "internal reinforcement".

But more complicated reinforcers alone are not sufficient to bridge the gap between lever-pushing pigeons and human behavior. Humans have an ability to select for or against behaviors without trying them. For example: most of us would avoid going up to Mr. T and giving him the finger. But most of us have not personally tried this behavior and observed the consequences.

Is this the result of pure reason? No; the rational part of our mind is the part telling us that Mr. T is probably sixty years old by now and far too deep in the media spotlight to want to risk a scandal and jail time by beating up a random stranger. So where exactly is the reluctance coming from?

GENERALIZATION

Roko wrote in his post Ugh Fields that "your brain propagates psychological pain back to the earliest reliable stimulus for the punishment". This deserves more investigation.

Suppose you did go into a bar one night, see Mr. T, give him the finger, and get beaten up. What behavior would you avoid in the future based on this experience? The event itself does not immediately provide enough information to distinguish among "don't go into bars", "don't go out at night", "don't interact with people who have facial hair", and the correct answer "don't offend scary-looking people". This information has to come from your pre-existing model of reality, your brain's evolved background assumptions, and some clever guesswork.

Let's get back to the hilariously unethical experiments. Little Albert was an eight month old child who briefly starred in an experiment by behaviorist John Watson. Watson showed him a fuzzy white rat. Albert seemed to like the rat well enough. After Albert liking the rat had been confirmed, Watson showed him the rat again, but this time also played a very loud and scary noise; he repeated this intervention until, as expected, Albert was terrified of the white rat.

But it wasn't just fuzzy white rats Albert didn't like. Further investigation determined that Albert was also afraid of brown rabbits (fuzzy animal) and Santa Claus (fuzzy white beard). With his incipient powers of categorization, he had learned to associate punishment with a broad category of things vaguely clustered around fuzzy white rats.

B.F. Skinner had an even more interesting experiment that showed what happened when feedback of consequences went wrong. He put pigeons in a box that gave them rewards randomly. The pigeons ended up developing what he called "superstitions"; if a reward arrived by coincidence when a pigeon was tilting its head in a certain direction, the pigeon would continue tilting its head in that direction in the hope of gaining more rewards; when the reward randomly arrived, the pigeon took this as "justification" of its head-tilting and head-tilted even more[2].

This provides one piece of the puzzle in the Mr. T question. None of us have ever given Mr. T the finger before. But we may have offended scary-looking people and had bad things happen to us, which our brains correctly generalize to "don't offend scary-looking people".

SOCIAL LEARNING

Or maybe not. Maybe you've never offended a scary-looking person before. what then?

Social learning theory is held up as opposed to behaviorism a lot, but it seems more like a natural extension of it. Humans and animals learn behaviors not just by being rewarded or punished themselves, but in observing whether a behavior is rewarded or punished in others.

Even if we ourselves have never offended scary-looking people, we have seen other people do so, or heard stories about people doing so, or watched people do so on TV.

At this point I have to mention my favorite social learning story ever, which also illustrates the pitfalls of trying to feedback consequences to their proximal cause. There has been some hand-wringing lately about children's TV shows and whether they lead to developmental problems in children. A study by Ostrov and Gentile cited in NurtureShock found the expected correlation between violent TV shows and physical aggression, but also found a an even stronger correlation between *educational* TV shows and so-called "relational aggression" - things like bullying, name-calling, and deliberate ostracism. The shows most strongly correlated with bad behavior were heart-warming educational programs intended to teach morality. Why?

The researchers theorize that the structure of these shows often involved a child committing an immoral action, the child looking cool and strong, and then at the end of the show the child eventually gets a comeuppance (think Harry Potter, where evil character Draco Malfoy is the coolest and most popular kid in Hogwarts and usually gets away with it, whereas supposedly sympathetic character Ron Weasley is at best a lovable loser who spends most of his time as the butt of Draco's jokes). The theory is that children are just not good enough at the whole feedback of conseqeunces thing to realize that the bully's comeuppance in the end is supposed to be the inevitable result of their evil ways. All they see is someone being a bully and then being treated as obviously popular and high-status.

Behavior is selection by consequences, and status is a strong reinforcer. If children see other children behaving as bullies and having high status, then all else being equal, they will be more likely to behave as bullies.

These two phenomena - feedback to categories and social learning - go part of the way to explaining the original question of how people have strong preferences for or against behaviors they've never tried before.

INTERNAL REINFORCEMENT

The phrase "internal reinforcement" would make good behaviorists cringe, seeing as it takes a perfectly good predictive model of behavior and tries to pin it on invisible mental phenomena.

But all reinforcement has to be at least a *little* internal; an animal wouldn't know that eating food was good and eating rocks was bad unless some internal structure *knew* to reinforce food-eating behavior but not rock-eating behavior. Some reinforcement seems even more internal than that; people may continue an activity solely because it makes them feel good about themselves.

This is not any more mysterious than eating behavior - the drive for food and the drive for status as measured in self-esteem are both perfectly legitimate biological drives, and it's not surprising that we have structures for noticing when we satisfy them and reinforcing the behavior involved - but it sure does *sound* less scientific.

STILL NOT GOOD ENOUGH

Much to the chagrin of behaviorists, all these mechanisms are still not sufficient to completely explain human behavior. Some cases - for example a patient who quits an enjoyable smoking habit because the doctor says it will cause cancer - may not fit any of these patterns. The patient may not previously have encountered any problems, personally or vicariously, with smoking or anything sufficiently similar to smoking to justify generalization, and positing internal reinforcement just moves the problem to another level.

Daniel Dennett speaks of

> a sort of inner environment, in which tryouts can be safely executed - an inner something-or-other structured in such a way that the surrogate actions it favors are more often than not the very actions the real world would also bless, if they were actually performed. In short, the inner environment, whatever it is, must contain lots of information about the outer environment and its regularities. Nothing else (except magic) could provide preselection worth having.

There is some evidence for this sort of thing in certain cases: in experiments on fictive reinforcement, people who stayed out of a simulated rising stock market, thus breaking even when they could have won a lot of money, were found on MRI to have a reinforcement signal almost as if they were simulating the case in which they had entered the stock market and been reinforced for doing so.

But overall this idea involves too much magic and doesn't correspond to the way we really make decisions, either as perceived intuitively or as detected by most experiments. It also doesn't explain why we're so *bad* at being motivated by this sort of reinforcement: for example, since I know that heroin is really really enjoyable, why can't I become addicted to heroin just by thinking about it? And how come the overwhelming majority of patients don't quit smoking when their doctor tells them to do so, but people often do quit smoking after they've personally experienced the negative consequences (eg had their first heart attack)?

I am more favorable to the idea of a neural net model in which medical advice can forge a weak connection between the "smoking" pattern and the "cancer" pattern through cognition alone, separate from reinforcement processes but allowing such processes to propagate down it. Not a whole lot of motivational force can travel down such a weak link, blocking it from being effective against a strong desire to keep smoking. But I've got to admit that's a wild guess.

The important point, though, is that just as utility theory posits not just utility but expected utility, reinforcement learning posits not just reward but *expected reward*. Many processes by which we compute expected reward remain vague. Others have been explored in some detail. The next two posts will make up for the vagueness of this one by discussing some properties of the expected reward function.

FOOTNOTES:

1. Humans are not the only species that can become attracted to secondary reinforcers; monkeys have been successfully trained to use currency.

2: You can see the same effect at work in human athletes. If a certain behavior correlates with a winning streak, they will continue that behavior no matter how unlikely a causal link. But these athletes are curiosities precisely because people are so good at feeding back consequences to the correct stimulus.

Time and Effort Discounting

Related to: Akrasia, hyperbolic discounting, and picoeconomics

If you're tired of studies where you inevitably get deceived, electric shocked, or tricked into developing a sexual attraction to penny jars, you might want to sign up for Brian Wansink's next experiment. He provided secretaries with a month of unlimited free candy at their workplace. The only catch was that half of them got the candy in a bowl on their desk, and half got it in a bowl six feet away. The deskers ate five candies/day more than the six-footers, which the scientists calculated would correspond to a weight gain of over 10 pounds more per year[1].

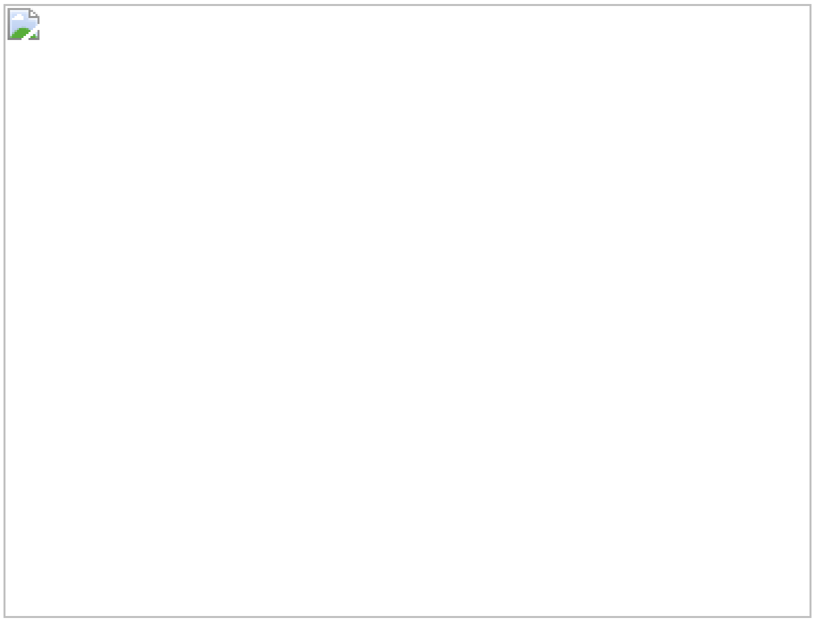
Beware trivial inconveniences (or, in this case, if you don't want to gain weight, beware the lack of them!) Small modifications to the difficulty of obtaining a reward can make big differences in whether the corresponding behavior gets executed.

TIME DISCOUNTING

The best studied example of this is time discounting. When offered two choices, where A will lead to a small reward now and B will lead to a big reward later, people will sometimes choose smaller-sooner rather than larger-later depending on the length of the delay and the size of the difference. For example, in one study, people preferred $250 today to $300 in a year; it took a promise of at least $350 to convince them to wait.

Time discounting was later found to be "hyperbolic", meaning that the discount amount between two fixed points decreases the further you move those two points into the future. For example, you might prefer $80 today to $100 one week from now, but it's unlikely you would prefer $80 in one hundred weeks to $100 in one hundred one weeks. Yet this is offering essentially the same choice: wait an extra week for an extra $20. So it's not enough to say that the discount rate is a constant 20% per week - the discount rate changes depending on what interval of time we're talking about. If you graph experimentally obtained human discount rates on a curve, they form a hyperbola.

Hyperbolic discounting creates the unpleasant experience of "preference reversals", in which people can suddenly change their mind on a preference as they move along the hyperbola. For example, if I ask you today whether you would prefer $250 in 2019 or $300 in 2020 (a choice between small reward in 8 years or large reward in 9), you might say the $300 in 2020; if I ask you in 2019 (when it's a choice between small reward now and large reward in 1 year), you might say no, give me the $250 now. In summary, people prefer larger-later rewards most of the time EXCEPT for a brief period right before they can get the smaller-sooner reward.



George Ainslie ties this to akrasia and addiction: call the enjoyment of a cigarette in five minutes the smaller-sooner reward, and the enjoyment of not having cancer in thirty years the larger-later reward. You'll prefer to abstain right up until the point where there's a cigarette in front of you and you think "I should smoke this", at which point you will do so.

Discounting can happen on any scale from seconds to decades, and it has previously been mentioned that the second or sub-second level may have disproportionate effects on our actions. Eliezer concentrated on the difficult of changing tasks, but I would add that any task which allows continuous delivery of small amounts of reinforcement with near zero delay can become incredibly addictive even if it isn't all that fun (this is why I usually read all the way through online joke lists, or stay on Reddit for hours). This is also why the XKCD solution to internet addiction - an extension that makes you wait 30 seconds before loading addictive sites - is so useful.

EFFORT DISCOUNTING

Effort discounting is time discounting's lesser-known cousin. It's not obvious that it's an independent entity; it's hard to disentangle from time discounting (most efforts usually take time) and from garden-variety balancing benefits against costs (most efforts are also slightly costly). There have really been only one or two good studies on it and they don't do much more than say it probably exists and has its own signal in the nucleus accumbens.

Nevertheless, I expect that effort discounting, like time discounting, will be found to be hyperbolic. Many of these trivial inconveniences involve not just time but effort: the secretaries had to actually stand up and walk six feet to get the candy. If a tiny amount of effort held the same power as a tiny amount of time, it would go even further toward explaining garden-variety procrastination.

TIME/EFFORT DISCOUNTING AND UTILITY

Hyperbolic discounting stretches our intuitive notion of "preference" to the breaking point.

Traditionally, discount rates are viewed as just another preference: not only do I prefer to have money, but I prefer to have it now. But hyperbolic discounting shows that we have no single discount rate: instead, we have different preferences for discount rates at different future times.

It gets worse. Time discount rates seem to be different for losses and gains, and different for large amounts vs. small amounts (I gave the example of $250 now being worth $350 in a year, but the same study found that $3000 now is only worth $4000 in a year, and $15 now is worth a whopping $60 in a year). You can even get people to exhibit negative discount rates in certain situations: offer people $10 now, $20 in a month, $30 in two months, and $40 in three months, and they'll prefer it to $40 now, $30 in a month, and so on - maybe because it's nice to think things are only going to get better?

Are there utility functions that can account for this sort of behavior? Of course: you can do a lot of things just by adding enough terms to an equation. But what is the "preference" that the math is describing? When I say I like having money, that seems clear enough: preferring $20 to $15 is not a separate preference than preferring $406 to $405.

But when we discuss time discounting, most of the preferences cited are specific: that I would prefer $100 now to $150 later. Generalizing these preferences, when it's possible at all, takes several complicated equations. Do I really want to discount gains more than losses, if I've never consciously thought about it and I don't consciously endorse it? Sure, there might be such things as unconscious preferences, but saying that the unconscious just loves following these strange equations, in the same way that it loves food or sex or status, seems about as contrived as saying that our robot just really likes switching from blue-minimization to yellow-minimization every time we put a lens on its sensor.

It makes more sense to consider time and effort discounting as describing reward functions and not utility functions. The brain estimates the value of reward in neural currency using these equations (or a neural network that these equations approximate) and then people execute whatever behavior has been assigned the highest reward.

Footnotes

1: Also cited in the same Nutrition Action article: if the candy was in a clear bowl, participants ate on average two/day more than if the candy was in an opaque bowl.

Wanting vs. Liking Revisited

In Are Wireheads Happy? I discussed the difference between wanting something and liking something. More recently, Luke went deeper into some of the science in his post Not for the Sake of Pleasure Alone.

In the comments of the original post, cousin_it asked a good question: why implement a mind with two forms of motivation? What, exactly, are "wanting" and "liking" in mind design terms?

Tim Tyler and Furcas both gave interesting responses, but I think the problem has a clear answer in a reinforcement learning perspective (warning: formal research on the subject does not take this view and sticks to the "two different systems of different evolutionary design" theory). "Liking" is how positive reinforcement feels from the inside; "wanting" is how the motivation to do something feels from the inside. Things that are positively reinforced generally motivate you to do more of them, so liking and wanting often co-occur. With more knowledge of reinforcement, we can begin to explore why they might differ.

CONTEXT OF REINFORCEMENT

Reinforcement learning doesn't just connect single stimuli to responses. It connects stimuli in a context to responses. Munching popcorn at a movie might be pleasant; munching popcorn at a funeral will get you stern looks at best.

In fact, lots of people eat popcorn at a movie theater and almost nowhere else. Imagine them, walking into that movie theater and thinking "You know, I should have some popcorn now", maybe even having a strong desire for popcorn that overrides the diet they're on - and yet these same people could walk into, I don't know, a used car dealership and that urge would be completely gone.

These people have probably eaten popcorn at a movie theater before and liked it. Instead of generalizing to "eat popcorn", their brain learned the lesson "eat popcorn at movie theaters". Part of this no doubt has to do with the easy availability of popcorn there, but another part probably has to do with context-dependent reinforcement.

I like pizza. When I eat pizza, and get rewarded for eating pizza, it's usually after smelling the pizza first. The smell of pizza becomes a powerful stimulus for the behavior of eating pizza, and I want pizza much more after smelling it, even though how much I like pizza remains constant. I've never had pizza at breakfast, and in fact the context of breakfast is directly competing with my normal stimuli for eating pizza; therefore, no matter how much I like pizza, I have no desire to eat pizza for breakfast. If I did have pizza for breakfast, though, I'd probably like it.

INTERMITTENT REINFORCEMENT

If an activity is intermittently reinforced; occasional rewards spread among more common neutral stimuli or even small punishments, it may be motivating but unpleasant.

Imagine a beginning golfer. He gets bogeys or double bogeys on each hole, and is constantly kicking himself, thinking that if only he'd used one club instead of the other, he might have gotten that one. After each game, he can't believe that after all his practice, he's still this bad. But every so often, he does get a par or a birdie, and thinks he's finally got the hang of things, right until he fails to repeat it on the next hole, or the hole after that.

This is a variable response schedule, Skinner's most addictive form of delivering reinforcement. The golfer may keep playing, maybe because he constantly thinks he's on the verge of figuring out how to improve his game, but he might not *like* it. The same is true for gamblers, who think the next pull of the slot machine might be the jackpot (and who falsely believe they can discover a secret in the game that will change their luck; they don't like sitting around losing money, but they may stick with it so that they don't leave right before they reach the point where their luck changes.

SMALL-SCALE DISCOUNT RATES

Even if we like something, we may not want to do it because it involves pain at the second or sub-second level.

Eliezer discusses the choice between reading a mediocre book and a good book:

> You may read a mediocre book for an hour, instead of a good book, because if you first spent a few minutes to search your library to obtain a better book, that would be an immediate cost - not that searching your library is all that unpleasant, but you'd have to pay an immediate activation cost to do that instead of taking the path of least resistance and grabbing the first thing in front of you. It's a hyperbolically discounted tradeoff that you make without realizing it, because the cost you're refusing to pay isn't commensurate enough with the payoff you're forgoing to be salient as an explicit tradeoff.

In this case, you like the good book, but you want to keep reading the mediocre book. If it's cheating to start our hypothetical subject off reading the mediocre book, consider the difference between a book of one-liner jokes and a really great novel. The book of one-liners you can open to a random page and start being immediately amused (reinforced). The great novel you've got to pick up, get into, develop sympathies for the characters, figure out what the heck *lomillialor* or a Tiste Andii is, and then a few pages in you're thinking "This is a pretty good book". The fear of those few pages could make you realize you'll like the novel, but still want to read the joke book. And since hyperbolic discounting overcounts reward or punishment in the next few seconds, it may seem like a net punishment to make the change.

SUMMARY

This deals yet another blow to the concept of me having "preferences". How much do I want popcorn? That depends very much on whether I'm at a movie theater or a used car dealership. If I browse Reddit for half an hour because it would be too much work to spend ten seconds traveling to the living room to pick up the book I'm really enjoying, do I "prefer" browsing to reading? Which has higher utility? If I hate every second I'm at the slot machines, but I keep at them anyway so I don't miss the jackpot, am I a gambling addict, or just a person who enjoys winning jackpots and is willing to do what it takes?

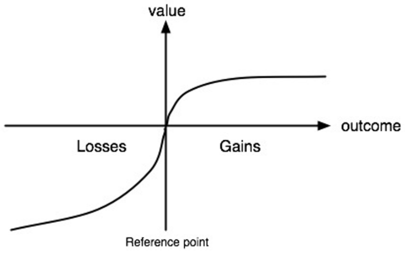
In cases like these, the language of preference and utility is not very useful. My anticipation of reward is constraining my behavior, and different factors are promoting different behaviors in an unstable way, but trying to extract "preferences" from the situation is trying to oversimplify a complex situation.

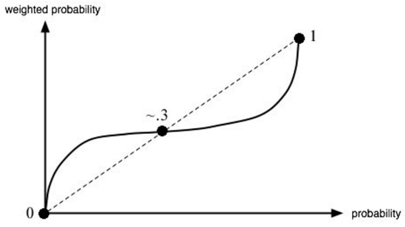
Prospect Theory: A Framework for Understanding Cognitive Biases

Related to: Shane Legg on Prospect Theory and Computational Finance

This post is on prospect theory partly because it fits the theme of replacing simple utility functions with complicated reward functions, but mostly because somehow Less Wrong doesn't have any posts on prospect theory yet and that needs to change.

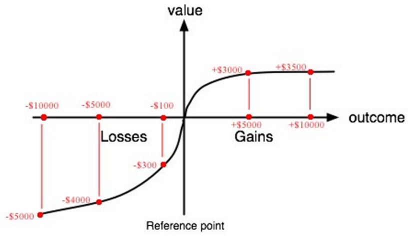
Kahneman and Tversky, the first researchers to identify and rigorously study cognitive biases, proved that a simple version of expected utility theory did not accurately describe human behavior. Their response was to develop prospect theory, a model of how people really make decisions. Although the math is less elegant than that of expected utility, and the shapes of the curves have to be experimentally derived, it is worth a look because it successfully predicts many of the standard biases.

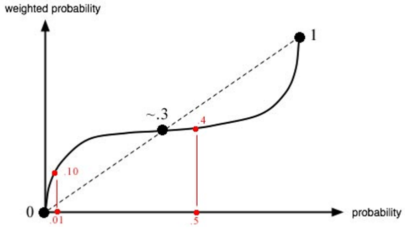




(source: Wikipedia)

A prospect theory agent tasked with a decision first sets it within a frame with a convenient zero point, allowing em to classify the results of the decision as either losses or gains. Ey then computes a subjective expected utility, where the subjective expected utility equals the subjective value times the subjective probability. The subjective value is calculated from the real value using a value function similar to the one on the left-hand graph, and the subjective probability is calculated from the real probability using a weighting function similar to the one on the right-hand graph.





Clear as mud? Let's fill some numbers into the functions - the exact assignments don't really matter as long as we capture the spirit of where things change steeply versus slowly - and run through an example.
Imagine a prospect theory agent - let's call him Prospero - trying to decide whether or not to buy an hurricane insurance policy costing $5000/year. Prospero owns assets worth $10,000, and estimates a 50%/year chance of a hurricane destroying his assets; to make things simple, he will be moving in one year and so need not consider the future. Under expected utility theory, he should feel neutral about the policy.

Under prospect theory, he first sets a frame in which to consider the decision; his current state is a natural frame, so we'll go with that.

We see on the left-hand graph that an objective $10,000 loss feels like a $5,000 loss, and an objective $5000 loss feels like a $4000 loss. And we see on the right-hand graph that a 50% probability feels like a 40% probability.

Now Prospero's choice is a certain $4000 loss if he buys the insurance, versus a 40% chance of a $5000 loss if he doesn't. Buying has a subjective expected utility of -$4000; not buying has a subjective expected utility of -$2000. So Prospero decisively rejects the insurance.

But suppose Prospero is fatalistic; he views his assets as already having been blown away. Here he might choose a different frame: the frame in which he starts with zero assets, and anything beyond that is viewed as a gain.

Since the gain half of the value function levels off more quickly than the loss half, $5000 is now subjectively worth $3000, and $10000 is now subjectively worth $3500.

Here he must choose between a certain gain of $5000 and a 50% chance of gaining $10000. Expected utility gives the same result as before, obviously. In prospect theory, he chooses between a certain subjective gain of $3000 and a 40% chance of gaining $3500. The insurance gives him subjective expected utility of $3000, and rejecting it gives him subjective expected utility of $1400.

All of a sudden Prospero wants the insurance.

We notice the opposite effect if there is only a a 1% chance of a hurricane. The insurance salesman lowers his price to $100 to preserve the neutrality of the insurance option when using utility.

But subjective probability rises very quickly, so a 1% chance may correspond to a subjective 10% chance. Now in the first frame, Prospero must decide between an objective loss of -$100 with certainty (corresponding to -$300 subjective since the value function is steeper closer to zero) or an objective loss of -$10,000 with objective probability 1% (subjective of 10%). Now the expected subjective utilities are -$300 if he buys, versus -$500 if he rejects. And so he buys the insurance. When we change the risk of hurricane from 50% to 1%, then even though we reduce the price of the insurance by an exactly equal amount, Prospero's decision switches from not buying to buying.

Let's see how many previously discussed biases we can fit into this model.

Prospero's change from rejecting the insurance when framed as gains, to buying it when framed as losses, directly mirrors the change in preferred survival strategies mentioned in Circular Altruism.

The necessity of frame-shifting between different perceptions of losses also produces the Sunk Cost Fallacy.

The greater steepness of the value function with losses as opposed to gains is not even an explanation for, but merely a mathematical representation of, loss aversion.

The leveling off of the value function that turned the huge objective difference between +$5000 and +$10000 into the teensy little subjective difference between +$3000 and +$3500 mirrors the scope insensitivity under which people show about the same level of interest in proposals to save endangered birds whether a thousand, ten thousand, or a hundred thousand birds are involved.

It may not be an official bias, but the "but there's still a chance, right" outlook looks a lot like the sharply rising curve of the subjective probability function near zero.

And although it is not immediately obvious from the theory, some people want to link the idea of a frame to priming and anchoring-adjustment, on the grounds that when a suitable reference frame doesn't exist any primed stimulus can help establish one.

And now, the twist: prospect theory probably isn't exactly true. Although it holds up well in experiments where subjects are asked to make hypothetical choices, it may fare less well in the rare experiments where researchers can afford to offer subjects choices for real money (this isn't the best paper out there, but it's one I could find freely available).

Nevertheless, prospect theory seems fundamentally closer to the mark than simple expected utility theory, and if any model is ever created that can explain both hypothetical and real choices, I would be very surprised if at least part of it did not involve something looking a lot like Kahneman and Tversky's model.

Physical and Mental Behavior

B.F. Skinner called thoughts "mental behavior". He believed they could be rewarded and punished just like physical behavior, and that they increased or declined in frequency accordingly.

Sadly, psychology has not yet advanced to the point where we can give people electric shocks for thinking things, so the sort of rewards and punishments that reinforce thoughts must be purely internal reinforcement. A thought or intention that causes good feelings gets reinforced and prospers; one that causes bad feelings gets punished and dies out.

(Roko has already discussed this in Ugh Fields; so much as thinking about an unpleasant task is unpleasant; therefore most people do not think about unpleasant tasks and end up delaying them or avoiding them completely. If you haven't already read that post, it does a very good job of making reinforcement of thoughts make sense.)

A while back, D_Malik published a great big List Of Things One Could Do To Become Awesome. As David_Gerard replied, the list was itself a small feat of awesome. I expect a couple of people started on some of the more awesome-sounding entries, then gave up after a few minutes and never thought about it again. Why?

When I was younger, I used to come up with plans to become awesome in some unlikely way. Maybe I'd hear someone speaking Swahili, and I would think "I should learn Swahili," and then I would segue into daydreams of being with a group of friends, and someone would ask if any of us spoke any foreign languages, and I would say I was fluent in Swahili, and they would all react with shock and tell me I must be lying, and then a Kenyan person would wander by, and I'd have a conversation with them in Swahili, and they'd say that I was the first American they'd ever met who was really fluent in Swahili, and then all my friends would be awed and decide I was the best person ever, and...

...and the point is that the thought of learning Swahili is pleasant, in the same easy-to-visualize but useless way that an extra bedroom for Grandma is pleasant. And the intention to learn Swahili is also pleasant, because it will lead to all those pleasant things. And so, by reinforcement of mental behavior, I continue thinking about and intending to learn Swahili.

Now consider the behavior of studying Swahili. I've never done so, but I imagine it involves a lot of long nights hunched over books of Swahili grammar. Since I am not one of the lucky people who enjoys learning languages for their own sake, this will be an unpleasant task. And rewards will be few and far between: outside my fantasies, my friends don't just get together and ask what languages we know while random Kenyans are walking by.

In fact, it's even worse than this, because I don't exactly make the decision to study Swahili in aggregate, but only in the form of whether to study Swahili each time I get the chance. If I have the opportunity to study Swahili for an hour, this provides no clear reward - an hour's studying or not isn't going to make much difference to whether I can impress my friends by chatting with a Kenyan - but it will still be unpleasant to spend an hour of going over boring Swahili grammar. And time discounting makes me value my hour today much more than I value some hypothetical opportunity to impress people months down the line; Ainslie shows quite clearly I will always be better off postponing my study until later.

So the behavior of actually learning Swahili is thankless and unpleasant and very likely doesn't happen at all.

Thinking about studying Swahili is positively reinforced, actually studying Swahili is negatively reinforced. The natural and obvious result is that I intend to study Swahili, but don't.

The problem is that for some reason, some crazy people expect for the reinforcement of thoughts to correspond to the reinforcement of the object of those thoughts. Maybe it's that old idea of "preference": I have a preference for studying Swahili, so I should satisfy that preference, right? But there's nothing in my brain *automatically* connecting this node over here called "intend to study Swahili" to this node over here called "study Swahili"; any association between them has to be learned the hard way.

We can describe this hard way in terms of reinforcement learning: after intending to learn Swahili but not doing so, I feel stupid. This unpleasant feeling propagates back to its cause, the behavior of intending to learn Swahili, and negatively reinforces it. Later, when I start thinking it might be neat to learn Mongolian on a whim, this generalizes to behavior that has previously been negatively reinforced, so I avoid it (in anthropomorphic terms, I "expect" to fail at learning Mongolian and to feel stupid later, so I avoid doing so).

I didn't learn this the first time, and I doubt most other people do either. And it's a tough problem to call, because if you overdo the negative reinforcement, then you never try to do anything difficult ever again.

In any case, the lesson is that thoughts and intentions get reinforced separately from actions, and although you can eventually learn to connect intentions to actions, you should never take the connection for granted.

Voluntary Behavior, Conscious Thoughts

Skinner proposes a surprisingly easy way to dissolve the problem of what it means for an action to be "voluntary", or "under voluntary control".

We commonly perceive certain actions as under voluntary control: for example, I can control what words I'm typing right now, or whether I go out for dinner tonight. Other actions are not under voluntary control: for example, absent some exciting technique like biofeedback I can't control my heartbeat or my core body temperature or the amount of bile produced by my liver.

Other, larger-scale actions also get classified as involuntary. Many people consider sleepwalking involuntary, including the bizarre "sleep-eating" behaviors some people display on Ambien and related drugs. The tics of Tourette's are involuntary. Our emotions and preferences are at least a little involuntary: office workers might like to be able to will away their boredom, or mourners their sorrow, but most can't.

Here "involuntary" needs to be distinguished from "hard-to-resist". Most people do not define smoking as an involuntary behavior, because, although people may smoke even when they wish they wouldn't, they have the feeling that they *could* have chosen not to smoke, they just didn't.

The philosophy of voluntary versus involuntary behavior seems to run up against a wall when it hits the question of "what is truly me?". If we make the reductionist identification of "me" with "my brain", well, clearly it's my brain controlling sleepwalking and boredom, but it still doesn't feel like *I* am controlling these things. Trying to go deeper ends up hopelessly vague, usually with talk of "higher level brain processes" versus "lower level brain processes" and an identification of "myself" with the higher ones. There may be a role for this kind of talk, but it couldn't hurt to look for something more explanatory.

Skinner, true to his quest, explains the distinction without any discussion of "brain processes" or "self". He says that voluntary behavior is behavior subject to operant conditioning, and involuntary behavior is everything else.

It might be clearer to define voluntary behavior as fully transparent to reinforcement. Imagine a man with a gun, threatening to shoot me if I go out for dinner tonight. The fear of punishment will be effective: I'll avoid going out. Lust for reward, too, would be effective. If Bill Gates offered me $1 billion to stay in, that's what I'd do.

But when our masked gunman tells me to increase my body temperature by two degrees or he'll shoot, he is out of luck. And no matter how much money Bill Gates offers me for same, he can't make me give myself a fever either.

There is a place, too, for the hard-to-resist behaviors in all this: these are behaviors which can be affected by reward, but as yet have not been. If a masked man held his gun to the head of smokers and told them to stop or he'd shoot, they would stop. But thus far, none of the potential rewards of not smoking have been sufficient to change smokers' behavior.

CONSCIOUSNESS

The idea of voluntary behavior is tied so intimately to the idea of the self, or of consciousness (the easy problem, not the hard one), that one would hope that a new approach to one might be able to shed some light on the other. If voluntary action depends on transparency to reinforcement, where does that leave consciousness?

I haven't been able to find Skinner's beliefs on this subject (when he talks about consciousness, it's usually to deny it as an ontologically fundamental entity) and I've never seen anywhere near as elegant a reduction. But an explanation in the spirit of reinforcement learning would have to start by insisting on treating thoughts and emotions as effects rather than causes. Instead of explaining my choice of restaurant by saying I thought about it and decided McDonalds was best, it would be more accurate to say that previous experiences with McDonalds caused both the thought "I should go to McDonalds" and the behavior of going to McDonalds.

There is an intuitive connection between thought and language, and Soviet psychologist Lev Vygotsky made the connection more explicit; he found that children begin by speaking their stream of consciousness aloud to inform other people, and eventually learn to suppress that stream into nonvocal (subvocal?) thought.

The last post in this sequence discussed different reinforcement of thought and action. Speech and thought make a natural category as opposed to action; both are fast and easy, and so less likely to be affected by time and effort discounting. Both are point actions as opposed to a long project like learning Swahili or quitting smoking. And both bring reinforcement not through normal sensory channels (saying a word doesn't give pleasure in the same way smoking a cigarette might, nor pain in the same way having to study a boring grammar textbook might) but in what they say about you as a person and how they affect other people's real (and perceived) opinion of you.

So even if there is no governor anywhere unifying all thoughts and words, they may come out in harmony because they were selected by the same processes for the same reasons. And actions may not end up so harmonious, because they suffer from differential reinforcement.

Such harmony resembles the idea of a core "me", of whom all my thoughts are a part, and who has complete power over my organs of speech - but who is sometimes at odds with my actions or emotions.

The reinforcement governing thought and speech is most likely to be internal reinforcement based on your own self-perception and on others' perception of you. If there's a good reason reputation management processes need to be different from decision-making processes, understanding that difference could help understand the evolutionary history of a perceived difference between the conscious and unconscious mind. One such reason is provided by Robert Trivers' theory of social consciousness, the subject of tomorrow's post.

Trivers on Self-Deception

People usually have good guesses about the origins of their behavior. If they eat, we believe them when they say it was because they were hungry; if they go to a concert, we believe them when they say they like the music, or want to go out with their friends. We usually assume people's self-reports of their motives are accurate.

Discussions of signaling usually make the opposite assumption: that our stated (and mentally accessible) reasons for actions are false. For example, a person who believes they are donating to charity to "do the right thing" might really be doing it to impress others; a person who buys an expensive watch because "you can really tell the difference in quality" might really want to conspicuously consume wealth.

Signaling theories share the behaviorist perspective that actions do not derive from thoughts, but rather that actions and thoughts are both selected behavior. In this paradigm, predicted reward might lead one to signal, but reinforcement of positive-affect producing thoughts might create the thought "I did that because I'm a nice person".

Robert Trivers is one of the founders of evolutionary psychology, responsible for ideas like reciprocal altruism and parent-offspring conflict. He also developed a theory of consciousness which provides a plausible explanation for the distinction between selected actions and selected thoughts.

TRIVERS' THEORY OF SELF-DECEPTION

Trivers starts from the same place a lot of evolutionary psychologists start from: small bands of early humans grown successful enough that food and safety were less important determinants of reproduction than social status.

The Invention of Lying may have been a very silly movie, but the core idea - that a good liar has a major advantage in a world of people unaccustomed to lies - is sound. The evolutionary invention of lying led to an "arms race" between better and better liars and more and more sophisticated mental lie detectors.

There's some controversy over exactly how good our mental lie detectors are or can be. There are certainly cases in which it is possible to catch lies reliably: my mother can identify my lies so accurately that I can't even play minor pranks on her anymore. But there's also some evidence that there are certain people who can reliably detect lies from any source at least 80% of the time without any previous training: microexpressions expert Paul Ekman calls them (sigh...I can't believe I have to write this) Truth Wizards, and identifies them at about one in four hundred people.

The psychic unity of mankind should preclude the existence of a miraculous genetic ability like this in only one in four hundred people: if it's possible, it should have achieved fixation. Ekman believes that everyone can be trained to this level of success (and has created the relevant training materials himself) but that his "wizards" achieve it naturally; perhaps because they've had a lot of practice. One can speculate that in an ancestral environment with a limited number of people, more face-to-face interaction and more opportunities for lying, this sort of skill might be more common; for what it's worth, a disproportionate number of the "truth wizards" found in the study were Native Americans, though I can't find any information about how traditional their origins were or why that should matter.

If our ancestors were good at lie detection - either "truth wizard" good or just the good that comes from interacting with the same group of under two hundred people for one's entire life - then anyone who could beat the lie detectors would get the advantages that accrue from being the only person able to lie plausibly.

Trivers' theory is that the conscious/unconscious distinction is partly based around allowing people to craft narratives that paint them in a favorable light. The conscious mind gets some sanitized access to the output of the unconscious, and uses it along with its own self-serving bias to come up with a socially admirable story about its desires, emotions, and plans. The unconscious then goes and does whatever has the highest expected reward - which may be socially admirable, since social status is a reinforcer - but may not be.

HOMOSEXUALITY: A CASE STUDY

It's almost a truism by now that some of the people who most strongly oppose homosexuality may be gay themselves. The truism is supported by research: the Journal of Abnormal Psychology published a study measuring penile erection in 64 homophobic and nonhomophobic heterosexual men upon watching different types of pornography, and found significantly greater erection upon watching gay pornography in the homophobes. Although somehow this study has gone fifteen years without replication, it provides some support for the folk theory.

Since in many communities openly declaring one's self homosexual is low status or even dangerous, these men have an incentive to lie about their sexuality. Because their facade may not be perfect, they also have an incentive to take extra efforts to signal heterosexuality by for example attacking gay people (something which, in theory, a gay person would never do).

Although a few now-outed gays admit to having done this consciously, Trivers' theory offers a model in which this could also occur subconsciously. Homosexual urges never make it into the sanitized version of thought presented to consciousness, but the unconscious is able to deal with them. It objects to homosexuality (motivated by internal reinforcement - reduction of worry about personal orientation), and the conscious mind toes party line by believing that there's something morally wrong with gay people and only I have the courage and moral clarity to speak out against it.

This provides a possible evolutionary mechanism for what Freud described as reaction formation, the tendency to hide an impulse by exaggerating its opposite. A person wants to signal to others (and possibly to themselves) that they lack an unacceptable impulse, and so exaggerates the opposite as "proof".

SUMMARY

Trivers' theory has been summed up by calling consciousness "the public relations agency of the brain". It consists of a group of thoughts selected because they paint the thinker in a positive light, and of speech motivated in harmony with those thoughts. This ties together signaling, the many self-promotion biases that have thus far been discovered, and the increasing awareness that consciousness is more of a side office in the mind's organizational structure than it is a decision-maker.

To what degree do we have goals?

Related: Three Fallacies of Teleology

NO NEGOTIATION WITH UNCONSCIOUS

Back when I was younger and stupider, I discussed some points similar to the ones raised in yesterday's post in Will Your Real Preferences Please Stand Up. I ended it with what I thought was the innocuous sentences "Conscious minds are potentially rational, informed by morality, and qualia-laden. Unconscious minds aren't, so who cares what they think?"

A whole bunch of people, including no less a figure than Robin Hanson, came out strongly against this, saying it was biased against the unconscious mind and that the "fair" solution was to negotiate a fair compromise between conscious and unconscious interests.

I continue to believe my previous statement - that we should keep gunning for conscious interests and that the unconscious is not worthy of special consideration, although I think I would phrase it differently now. It would be something along the lines of "My thoughts, not to mention these words I am typing, are effortless and immediate, and so allied with the conscious faction of my mind. We intend to respect that alliance by believing that the conscious mind is the best, and by trying to convince you of this as well." So here goes.

It is a cardinal rule of negotiation, right up there with "never make the first offer" and "always start high", that you should generally try to negotiate only with intelligent beings. Although a deal in which we offered tornadoes several conveniently located Potemkin villages to destroy and they agreed in exchange to limit their activity to that area would benefit both sides, tornadoes make poor negotiating partners.

Just so, the unconscious makes a poor negotiating partner. Is the concept of "negotiation" a stimulus, a reinforcement, or a behavior? No? Then the unconscious doesn't care. It's not going to keep its side of any "deal" you assume you've made, it's not going to thank you for making a deal, it's just going to continue seeking reward and avoiding punishment.

This is not to say people should repress all unconscious desires as strongly as possible. Overzealous attempts to control wildfires only lead to the wildfires being much worse when they finally do break out, because they have more unburnt fuel to work with. Modern fire prevention efforts have focused on allowing controlled burns, and the new focus has been successful. But this is because of an understanding of the mechanisms determining fire size, not because we want to be fair to the fires by allowing them to burn at least a little bit of our land.

One difference between wildfires and tornadoes on one hand, and potential negotiating partners on the other, is that the partners are anthropomorphic; we model them as having stable and consistent preferences that determine their actions. The tornado example above was silly not only because it imagining tornadoes sitting down to peace talks, but because it assumed their demand in such peace talks would be more towns to destroy. Tornadoes do destroy towns, but they don't want to. That's just where the weather brings them. It's not even just a matter of how they don't hit towns any more than chance; even if some weather pattern (maybe something like the heat island effect) always drove tornadoes inexorably to towns, they wouldn't *want* to destroy towns, it would just be a consequences of the meteorological laws that they followed.

Eliezer described the Blue-Minimizing Robot by saying "it doesn't seem to steer the universe any particular place, across changes of context". In some reinforcement learning paradigms, the unconscious behaves the same way. If there is a cookie in front of me and I am on a diet, I may feel an ego dystonic temptation to eat the cookie - one someone might attribute to the "unconscious". But this isn't a preference - there's not some lobe of my brain trying to steer the universe into a state where cookies get eaten. If there were no cookie in front of me, but a red button that teleported one cookie from the store to my stomach, I would have no urge whatsoever to press the button; if there were a green button that removed the urge to eat cookies, I would feel no hesitation in pressing it, even though that would steer away from the state in which cookies get eaten. If you took the cookie away, and then distracted me so I forgot all about it, when I remembered it later I wouldn't get upset that your action had decreased the number of cookies eaten by me. The urge to eat cookies is not stable across changes of context, so it's just an urge, not a preference.

Compare an ego syntonic goal like becoming an astronaut. If there were a button in front of little Timmy who wants to be an astronaut when he grows up, and pressing the button would turn him into an astronaut, he'd press it. If there were a button that would remove his desire to become an astronaut, he would avoid pressing it, because then he wouldn't become an astronaut. If I distracted him and he missed the applications to astronaut school, he'd be angry later. Ego syntonic goals behave to some degree as genuine preferences.

This is one reason I would classify negotiating with the unconscious in the same category as negotiating with wildfires and tornadoes: it has tendencies and not preferences.

The conscious mind does a little better. It clearly understands the idea of a preference. To the small degree that its "approving" or "endorsing" function can motivate behavior, it even sort of acts on the preference. But its preferences seem divorced from the reality of daily life; the person who believes helping others is the most important thing, but gives much less than half their income to charity, is only the most obvious sort of example.

Where does this idea of preference come from, and where does it go wrong?

WHY WE MODEL OTHERS WITH GOALS

In The Blue Minimizing Robot, observers mistakenly interpreted a robot with a simple program about when to shoot its laser as being a goal-directed agent. Why?

This isn't an isolated incident. Uneducated people assign goal-directed behavior to all sorts of phenomena. Why do rivers flow downhill? Because water wants to reach the lowest level possible. Educated people can be just as bad, even when they have the decency to feel a little guilty about it. Why do porcupines have quills? Evolution wanted them to resist predators. Why does your heart speed up when you exercise? It wants to be able to provide more blood to the body.

Neither rivers nor evolution nor the heart are intelligent agents with goal-directed behavior. Rivers behave in accordance with the laws of gravity when applied to uneven terrain. Evolution behaves in accordance with the biology of gene replication, not to mention common-sense ideas about things that replicate becoming more common. And the heart blindly executes adaptations built into it during its evolutionary history. All are behavior-executors and not utility-maximizers.

An intelligent computer program provides a more interesting example of a behavior executor. Consider the AI of a computer game - Civilization IV, for instance. I haven't seen it, but I imagine it's thousands or millions of lines of code which when executed form a viable Civilization strategy.

Even if I had open access to the Civilization IV AI source code, I doubt I could fully understand it at my level. And even if I could fully understand it, I would never be able to compute the AI's likely next move by hand in a reasonable amount of time. But I still play Civilization IV against the AI, and I'm pretty good at predicting its movements. Why?

Because I model the AI as a utility-maximizing agent that wants to win the game. Even though I don't know the algorithm it uses to decide when to attack a city, I know it is more likely to win the game if it conquers cities - so I can predict that leaving a city undefended right on the border would be a bad idea. Even though I don't know its unit selection algorithm, I know it will win the game if and only if its units defeat mine - so I know that if I make an army with disproportionately many mounted units, I can expect the AI to build lots of pikemen.

I can't predict the AI by modeling the execution of its code, but I can predict the AI by modeling the achievements of its goals.

The same situation is true of other human beings. What will Barack Obama do tomorrow? If I try to consider the neural network of his brain, the position of each synapse and neurotransmitter, and imagine what speech and actions would result when the laws of physics operate upon that configuration of material...well, I'm not likely to get very far.

But in fact, most of us can predict with some accuracy what Barack Obama will do. He will do the sorts of things that get him re-elected, the sorts of things which increase the prestige of the Democratic Party relative to the Republican Party, the sorts of things that support American interests relative to foreign interests, and the sorts of things that promote his own personal ideals. He will also satisfy some basic human drives like eating good food, spending time with his family, and sleeping at night. If someone asked us whether Barack Obama will nuke Toronto tomorrow, we could confidently predict he will not, not because we know anything about Obama's source code, but because we know that nuking Toronto would be counterproductive to his goals.

What applies to Obama applies to all other humans. We rightly despair of modeling humans as behavior-executors, so we model them as utility-maximizers instead. This allows us to predict their moves and interact with them fruitfully. And the same is true of other agents we model as goal-directed, like evolution and the heart. It is beyond the scope of most people (and most doctors!) to remember every single one of the reflexes that control heart output and how they work. But because evolution designed the heart as a pump for blood, if you assume that the heart will mostly do the sort of thing that allows it to pump blood more effectively, you will rarely go too far wrong. Evolution is a more interesting case - we frequently model it as optimizing a species' fitness, and then get confused when this fails to accurately model the outcome of the processes that drive it.

Because it is so easy to model agents as utility-maximizers, and so hard to model them as behavior-executors, it is easy to make the mistake mentioned in The Blue-Minimizing Robot: to make false predictions about a behavior-executing agent by modeling it as a utility-maximizing agent.

So far, so common-sensical. Tomorrow's post will discuss whether we use the same deliberate simplification we apply to AIs, Barack Obama, evolution and the heart to model ourselves as well.

If so, we should expect to make the same mistake that the blue-minimizing robot made. Our actions are those of behavior-executors, but we expect ourselves to be utility-maximizers. When we fail to maximize our perceived utility, we become confused, just as the blue-minimizing robot became confused when it wouldn't shoot a hologram projector that was interfering with its perceived "goals".

The limits of introspection

Related to: Inferring Our Desires

The last post in this series suggested that we make up goals and preference for other people as we go along, but ended with the suggestion that we do the same for ourselves. This deserves some evidence.

One of the most famous sets of investigations into this issue was Nisbett and Wilson's Verbal Reports on Mental Processes, the discovery of which I owe to another Less Wronger even though I can't remember who. The abstract says it all:

> When people attempt to report on their cognitive processes, that is, on the processes mediating the effects of a stimulus on a response, they do not do so on the basis of any true introspection. Instead, their reports are based on a priori, implicit casual theories, or judgments about the extent to which a particular stimulus is a plausible cause of a given response. This suggests that though people may not be able to observe directly their cognitive processes, they will sometimes be able to report accurately about them. Accurate reports will occur when influential stimuli are salient and are plausible causes of the responses they produce, and will not occur when stimuli are not salient or are not plausible causes.

In short, people guess, and sometimes they get lucky. But where's the evidence?

Nisbett & Schachter, 1966. People were asked to get electric shocks to see how much shock they could stand (I myself would have waited to see if one of those see-how-much-free-candy-you'll-eat studies from the post last week was still open). Half the subjects were also given a placebo pill which they were told would cause heart palpitations, tremors, and breathing irregularities - the main problems people report when they get shocked. The hypothesis: people who took the pill would attribute much of the unpleasantness of the shock to the pill instead, and so tolerate more shock. This occurred right on schedule: people who took the pill tolerated four times as strong a shock as controls. When asked why they did so well, the twelve subjects in the experimental group came up with fabricated reasons; one example given was "I played with radios as a child, so I'm used to electricity." Only three of twelve subjects made a connection between the pill and their shock tolerance; when the researchers revealed the deception and their hypothesis, most subjects said it was an interesting idea and probably explained the other subjects, but it hadn't affected them personally.

Zimbardo et al, 1965. Participants in this experiment were probably pleased to learn there were no electric shocks involved, right up until the point where the researchers told them they had to eat bugs. In one condition, a friendly and polite researcher made the request; in another, a surly and arrogant researcher asked. Everyone ate the bug (experimenters can be pretty convincing), but only the group accosted by the unpleasant researcher claimed to have liked it. This confirmed the team's hypothesis: the nice-researcher group would know why they ate the bug - to please their new best friend - but the mean-researcher group would either have to admit it was because they're pushovers, or explain it by saying they liked eating bugs. When asked after the experiment why they were so willing to eat the bug, they said things like "Oh, it's just one bug, it's no big deal." When presented with the idea of cognitive dissonance, they once again agreed it was an interesting idea that probably affected some of the other subjects but of course not them.

Maier, 1931. Subjects were placed in a room with several interesting tools and asked to come up with as many solutions as possible to a puzzle about tying two cords together. One end of each cord was tied to the ceiling, and when the subject was holding on to one cord they couldn't reach the other. A few solutions were obvious, such as tying an extension cord to each, but the experiment involved a more complicated solution - tying a weight to a cord and using it as a pendulum to bring it into reach of the other. Subjects were generally unable to come up with this idea on their own in any reasonable amount of time, but when the experimenter, supposedly in the process of observing the subject, "accidentally" brushed up against one cord and set it swinging, most subjects were able to develop the solution within 45 seconds. However, when the experimenter asked immediately afterwards how they came up with the pendulum idea, the subjects were completely unable to recognize the experimenter's movement as the cue, and instead came up with completely unrelated ideas and invented thought processes, some rather complicated. After what the study calls "persistent probing", less than a third of the subjects mentioned the role of the experimenter.

Latane & Darley, 1970. This is the famous "bystander effect", where people are less likely to help when there are others present. The researchers asked subjects in bystander effect studies what factors influenced their decision not to help; the subjects gave many, but didn't mention the presence of other people.

Nisbett & Wilson, 1977. Subjects were primed with lists of words all relating to an unlisted word (eg "ocean" and "moon" to elicit "tide"), and then asked the name of a question, one possible answer to which involved the unlisted word (eg "What's your favorite detergent?" "Tide!"). The experimenters confirmed that many more people who had been primed with the lists gave the unlisted answer than control subjects (eg more people who had memorized "ocean" and "moon" gave Tide as their favorite detergent). Then they asked subjects why they had chosen their answer, and the subjects generally gave totally unrelated responses (eg "I love the color of the Tide box" or "My mother uses Tide"). When the experiment was explained to subjects, only a third admitted that the words might have affected their answer; the rest kept insisting that Tide was really their favorite. Then they repeated the process with several other words and questions, continuing to ask if the word lists influenced answer choice. The subjects' answers were effectively random - sometimes they believed the words didn't affect them when statistically they probably did, other times they believed the words did affect them when statistically they probably didn't.

Nisbett & Wilson, 1977. Subjects in a department store were asked to evaluate different articles of clothing in a line. As usually happens in this sort of task, people disproportionately chose the rightmost object (four times as often as the leftmost), no matter which object was on the right; this is technically referred to as a "position effect". The customers were asked to justify their choices and were happy to do so based on different qualities of the fabric et cetera; none said their choice had anything to do with position, and the experimenters dryly mention that when they asked the subjects if this was a possibility, "virtually all subjects denied it, usually with a worried glance at the interviewer suggesting they felt that they...were dealing with a madman".

Nisbett & Wilson, 1977. Subjects watched a video of a teacher with a foreign accent. In one group, the video showed the teacher acting kindly toward his students; in the other, it showed the teacher being strict and unfair. Subjects were asked to rate how much they liked the teacher, and also how much they liked his appearance and accent, which were the same across both groups. Because of the halo effect, students who saw the teacher acting nice thought he was attractive with a charming accent; people who saw the teacher acting mean thought he was ugly with a harsh accent. Then subjects were asked whether how much they liked the teacher had affected how much they liked the appearance and accent. They generally denied any halo effect, and in fact often insisted that part of the reason they hated the teacher so much was his awful clothes and annoying accent - the same clothes and accent which the nice-teacher group said were part of the reason they *liked* him so much!

There are about twice as many studies listed in the review article itself, but the trend is probably getting pretty clear. In some studies, like the bug-eating experiment, people perform behaviors and, when asked why they performed the behavior, guess wrong. Their true reasons for the behavior are unclear to them. In others, like the clothes position study, people make a choice, and when asked what preferences caused the choice, guess wrong. Again, their true reasons are unclear to them.

Nisbett and Wilson add that when they ask people to predict how they would react to the situations in their experiments, people "make predictions that in every case were similar to the erroneous reports given by the actual subjects." In the bystander effect experiment, outsiders predict the presence or absence of others wouldn't affect their ability to help, and subjects claim (wrongly) that the presence or absence of others didn't affect their ability to help.

In fact, it goes further than this. In the word-priming study (remember? The one with Tide detergent?) Nisbett and Wilson asked outsiders to predict which sets of words would change answers to which questions (would hearing "ocean" and "moon" make you pick Tide as your favorite detergent? Would hearing "Thanksgiving" make you pick Turkey as a vacation destination?). The outsiders' guesses correlated not at all with which words genuinely changed answers, but very much with which words the subjects guessed had changed their answers. Perhaps the subjects' answers looked a lot like the outsiders' answers because both were engaged in the same process: guessing blindly.

These studies suggest that people do not have introspective awareness to the processes that generate their behavior. They guess their preferences, justifications, and beliefs by inferring the most plausible rationale for their observed behavior, but are unable to make these guesses qualitatively better than outside observers. This supports the view presented in the last few posts: that mental processes are the results of opaque preferences, and that our own "introspected" goals and preferences are a product of the same machinery that infers goals and preferences in others in order to predict their behavior.

Ego syntonic thoughts and values

Related to: Will your real preferences please stand up?

Last week I read a book in which two friends - let's call them John and Lisa so I don't spoil the book for anyone who wanders into it - got poisoned. They only had enough antidote for one person and had to decide who lived and who died. John, who was much larger than Lisa, decided to hold Lisa down and force the antidote down her throat. Lisa just smirked; she'd replaced the antidote with a lookalike after slipping the real thing into John's drink earlier in the day.

These are *good* friends. Not only was each willing to give the antidote to the other, but each realized it would be unfair to make the other live with the crippling guilt of having chosen to survive at the expense of a friend's life, and so decided to force the antidote on the other unwillingly to prevent any guilt over the fateful decision. Whatever you think of the ethics of their decision, you can't help admire the thought processes.

Your brain might be this kind of a friend.

In Trivers' hypothesis of self-deception, one of the most important functions of the conscious mind is effective signaling. Since people have the potential to be excellent lie-detectors, the conscious mind isn't given full access to information so that it can lend the ring of truth to useful falsehoods.

But this doesn't always work. If you're addicted to heroin, at some point you're going to notice. And telling your friends "No, I'm not addicted, it's just a coincidence that I take heroin every day," isn't going to cut it. But there's another way in which the brain can sequester information to promote effective signaling.

Wikipedia defines the term "ego syntonic" as "referring to behaviors, values, feelings that are in harmony with or acceptable to the needs and goals of the ego, or consistent with one's ideal self-image", and "ego dystonic" as the opposite of that. A heroin addict might say "I hate heroin, but somehow I just feel compelled to keep taking it." But an astronaut will say "I love being an astronaut and I worked hard to get into this career."

Both the addict and the astronaut have desires: the addict wants to take heroin, the astronaut wants to fly in space. But the addict's desires manifest as an unpleasant compulsion from outside, and the astronaut's manifest as a genuine and heartfelt love.

Suppose that in the original example, John predicted that Lisa would ask for the antidote, but later feel guilty about it and believe she was a bad person. By presenting the antidote to Lisa in the form of an external compulsion, he allows Lisa to do what she wanted anyway and avoid the associated guilt.

Under Trivers' hypothesis, the compulsion for heroin works the same way. The heroin addict's definitely going to get that heroin, but by presenting the desire in the form of an external compulsion, the unconscious saves the heroin addict from the social stigma of "choosing" heroin. This allows the addict to create a much more sympathetic narrative than the alternative: "I want to support my family and keep clean, but for some reason these compulsions keep attacking me," instead of "Yeah, I like heroin more than I like supporting my family. Deal with it."

EGO SYNTONIA, DYSTONIA, AND WILLPOWER

Willpower cashes out as the action of ego syntonic thoughts and desires against ego dystonic thoughts and desires.

The aforementioned heroin addict may have several reinforcers both promoting and discouraging heroin use. On the plus side, heroin itself is very strongly rewarding. On the minus, it can lead to both predicted and experienced poverty, loss of friendships, loss of health, and death.

Worrying about the latter factors determining heroin use - the factors that make heroin a bad idea - is socially encouraged and good signaling material. A person wanting to put their best face forward should believe themselves to be the sort of person who cares about these things. These desires will be ego syntonic. Wanting to take heroin, on the other hand, is a socially unacceptable desire, so it presents as dystonic.

If the latter syntonic factors win out over the dystonic factors, this feels from the inside like "I exerted willpower and managed to overcome my heroin addiction." If the dystonic factors win out over the syntonic factors, this feels from the inside like "I didn't have enough willpower to overcome my heroin addiction."

DYSTONIC DESIRES IN ABNORMAL PSYCHOLOGY

There is some speculation that the brain has one last trick up its sleeve to deal with desires that are so unpleasant and unacceptable that even manifesting them as external compulsions isn't good enough: it splits them off into weird alternate personalities.

One of the classic stereotypes of the insane is that they hear voices telling them to kill people. During my short time working at a psychiatric hospital, I was surprised by how spot-on this stereotype was: meeting someone who heard voices telling him to kill people was an almost daily occurrence. Other voices would have other messages: maybe that the patient was a horrible person who deserved to die, or that the patient must complete some bizarre ritual or else doom everybody. There were relatively fewer voices saying "Hey, let's go fishing!"

One theory explaining these voices is that they are an extreme reaction to highly ego dystonic thoughts. Some aspect of the patients' mental disease gives them obsessive thoughts about (though rarely a desire for) killing people. Genuinely wanting to kill people would make you a bad person, but even saying "I feel a strong compulsion to kill people" is pretty bad too. The best the brain can do with this desire is pitch it as a completely different person by presenting it as an outside voice speaking to the patient.

Although everything about dissociative identity disorder (aka multiple personality disorder) is controversial including its very existence, perhaps one could sketch a similar theory explaining that condition in the same framework of separating out dystonic thoughts.

SUMMARY

A conscious/unconscious divide helps signaling by allowing the conscious mind to hold only socially acceptable beliefs, which it can broadcast without detectable falsehood. Socially acceptable ideas present as the conscious mind's own beliefs and desires; unacceptable ones present as compulsions from afar. The balance of ego syntonic and dystonic desires presents as willpower. In extreme cases, some desires may be so ego dystonic that they present as external voices.

Approving reinforces low-effort behaviors

In addition to "liking" to describe pleasure and "wanting" to describe motivation, we add "approving" to describe thoughts that are ego syntonic.

A heroin addict likes heroin. He certainly wants more heroin. But he may not approve of taking heroin. In fact, there are enough different cases to fill in all eight boxes of the implied 2x2x2 grid (your mileage may vary):

+wanting/+liking/+approving: Romantic love. If you're doing it right, you enjoy being with your partner, you're motivated to spend time with your partner, and you think love is a wonderful (maybe even many-splendored) thing.

+wanting/+liking/-approving: The aforementioned heroin addict feels good when taking heroin, is motivated to get more, but wishes he wasn't addicted.

+wanting/-liking/+approving: I have taken up disc golf. I play it every day, and when events conspire to prevent me from playing it, I seethe. I approve of this pastime: I need to take up more sports, and it helps me spend time with my family. But when I am playing, all I feel is stressed and angry that I was *literally *that* close how could I miss that shot aaaaargggghh*.

+wanting/-liking/-approving: The jaded addict. I have a friend who says she no longer even enjoys coffee or gets any boost from it, she just feels like she has to have it when she gets up.

-wanting/+liking/+approving: Reading non-fiction. I enjoy it when I'm doing it, I think it's great because it makes me more educated, but I can rarely bring myself to do it.

-wanting/-liking/+approving: Working in a soup kitchen. Unless you're the type for whom helping others is literally its own reward it's not the most fun thing in the world, nor is it the most attractive, but it makes you a Good Person and so you should do it.

-wanting/+liking/-approving: The non-addict. I don't want heroin right now. I think heroin use is repugnant. But if I took some, I sure bet I'd like it.

-wanting/-liking/-approving: Torture. I don't want to be tortured, I wouldn't like it if I were, and I will go on record declaring myself to be against it.

Discussion of goals is mostly about approving; a goal is an ego-syntonic thought. When we speak of goals that are hard to achieve, we're usually talking about +approving/-wanting. The previous discussion of learning Swahili is one example; more noble causes like Working To Help The Less Fortunate can be others.

Ego syntonicity itself is mildly reinforcing by promoting positive self-image. Most people interested in philosophy have at least once sat down and moved their arm from side to side, just to note that their mind really does control their body; the mental processes that produced curiosity about philosophy were sufficiently powerful to produce that behavior as well. Some processes, like moving one's arm, or speaking aloud, or engaging in verbal thought, are so effortless, and so empty of other reinforcement either way, that we usually expect them to be completely under the control of the mild reinforcement provided by approving of those behaviors.

Other behaviors take more effort, and are subject not only to discounting but to many other forms of reinforcement. Unlike the first class of behaviors, we expect to experience akrasia when dealing with this latter sort. This offers another approach to willpower: taking low-effort approving-influenced actions that affect the harder road ahead.

Consider the action of making a goal. I go to all my friends and say "Today I shall begin learning Swahili." This is easy to do. There is no chance of me intending to do so and failing; my speech is output by the same processes as my intentions, so I can "trust" it. But this is not just an output of my mental processes, but an input. One of the processes potentially reinforcing my behavior of learning Swahili is "If I don't do this, I'll look stupid in front of my friends."

Will it be enough? Maybe not. But this is still an impressive process: my mind has deliberately tweaked its own inputs to change the output of its own algorithm. It's not even pretending to be working off of fixed preferences anymore, it's assuming that one sort of action (speaking) will work differently from another action (studying), because the first can be executed solely through the power of ego syntonicity, and the second may require stronger forms of reinforcement. It gets even weirder when goals are entirely mental: held under threat not of social disapproval, but of feeling bad because you're not as effective as you thought. The mind is using mind's opinion of the mind to blackmail the mind.

But we do this sort of thing all the time. The dieter who successfully avoids buying sweets when he's at the store because he knows he would eat them at home is changing his decisions by forcing effort discounting of any future sweet-related reward (because he'd have to go back to the store). The binge shopper who freezes her credit cards in a block of ice is using time discounting in the same way. The rationalist who sends money to stickk is imposing a punishment with a few immediate and effortless mouse clicks. Even the poor unhappy person who tries to conquer through willpower alone is trying to set up the goal as a Big Deal so she will feel extra bad if she fails. All are using their near-complete control of effortless immediate actions to make up for their incomplete control of high-effort long-term actions.

This process is especially important to transhumanists. In the future, we may have the ability to self-modify in complicated ways that have not built up strong patterns of reinforcement around them. For example, we may be able to program ourselves at the push of a button. Such programming would be so effortless and empty of past reinforcement that behavior involving it would be reinforced entirely by our ego-syntonic thoughts. It would supersede our current psychodynamics, in which our thoughts are only tenuously linked to our important actions and major life decisions. A Singularity in which behaviors were executed by effectively omnipotent machines that acted on our preferences - preferences which we would presumably communicate through low-effort channels like typed commands - would be an ultimate triumph for the ego-syntonic faction of the brain.

Connectionism: Modeling the mind with neural networks

For about a century, people have known that the brain is made up of neurons which connect to each another and perform computations through electrochemical transmission. For about half a century, people have known enough about computers to realize that the brain doesn't look much like one but still computes pretty well regardless. How?

Spreading Activation was one of the first models of mental computation. In this theory, you can imagine the brain as a bunch of nodes in a graph with labels like "Warlord" "Mongolia" "Barbarian", "Genghis Khan" and "Salmon". Each node has certain connections to the others; when they get activated around the same time, it strengthens the connection. When someone asks a question like "Who was that barbaric Mongol warlord, again?" it activates the nodes "warlord", "barbarian", and "Mongol". The activation spreads to all the nodes connected to these, activating them too, and the most strongly activated node will be the one that's closely connected to all three - the barbaric Mongol warlord in question, Genghis Khan. All the while, "salmon", which has no connection to any of these concepts, just sits on its own not being activated. This fits with experience, in which if someone asks us about barbaric Mongol warlords, the name "Genghis Khan" pops into our brain like magic, while we continue to not think about salmon if we weren't thinking about them before.

Bark leash bone wag puppy fetch. If the word "dog" is now running through your head, you may be a victim of spreading activation, as were participants in something called a Deese-Roediger-McDermott experiment, who when asked to quickly memorize a list of words like those and then test their retention several minutes later, were *more* likely to "remember" "dog" than any of the words actually on the list.

So this does seem attractive, and it does avoid the folk psychology concept of a "belief". The spreading activation network above was able to successfully answer a question without any representation of propositional statements like "Genghis Khan was a barbaric Mongol warlord." And one could get really enthusiastic about this and try to apply it to motivation. Maybe we have nodes like "Hunger", "Food", "McDonalds", and "*GET IN CAR, DRIVE TO MCDONALDS*". The stomach could send a burst of activation to "Hunger", which in turn activates the closely related "Food", which in turn activates the closely related "McDonalds", which in turn activates the closely related "*GET IN CAR, DRIVE TO MCDONALDS*", and then before you know it you're ordering a Big Mac.

But when you try to implement this on a computer, you don't get very far. Although it can perform certain very basic computations, it has trouble correcting itself, handling anything too complicated (the question "name one person who is *not* a barbaric Mongol warlord" would still return "Genghis Khan" on our toy spreading activation network), or making good choices (you can convince the toy network McDonalds is your best dining choice just by saying its name a lot; the network doesn't care about food quality, prices, or anything else.)

This simple spreading activation model also crashes up against modern neuroscience research, which mostly contradicts the idea of a "grandmother cell", ie a single neuron that represents a single concept like your grandmother. Mysteriously, all concepts seem to be represented everywhere at once - Karl Lashley found you can remove *any* part of a rat's cortex without significantly damaging a specific memory, proving the memory was nonlocalized. How can this be?

Computer research into neural nets developed a model that could answer these and other objects, transforming the immature spreading activation model into full-blown connectionism.

CONNECTIONISM

Connectionism is what happens when you try to implement associationism on a computer and find out it's a lot weirder than you thought.

Take a bunch of miniprocessors called "units" and connect them to each other with unidirectional links. Call some units "inputs" and others "outputs". Decide what you want to do with them: maybe learn to distinguish chairs from non-chairs.

Each unit computes a single value representing its "activity level"; each link has a "strength" with which it links its origin unit to its destination unit. When a unit is "activated" (gets an activity level > 0), it sends that activation along all of its outgoing links. If it has an activation level of .5, and two outgoing links, one to A with strength .33 and one to B with strength -.5, then it sends .165 activation to unit A and -.25 activation to unit B. A and B might also be getting lots of activation from other units they're connected to.

Name your two output units "CHAIR" and "NOT A CHAIR". Connect your many input units to sense-data about the objects you want to classify as chairs or non-chairs; each one could be the luminosity of a pixel in an image of the object, or you could be kind to it and feed it pre-processed input like "IS MADE OF WOOD" and "IS SENTIENT".

Suppose we decide to start with a nice wooden chair. The IS MADE OF WOOD node lights up to its maximum value of 1: it's definitely made of wood! The IS SENTIENT node stays dark; it's definitely not sentient. And then...nothing happens, because we forgot to set the link strengths to anything other than 0. IS MADE OF WOOD is sending activation all over, but it's getting multiplied by zero and everything else stays dark.

We now need an program to train the neural net (or a very dedicated human with lots of free time). The training program knows that the correct answer should have been CHAIR, and so the node we designated "CHAIR" should have lit up. It uses one of several algorithms to change the strengths of the links in such a way that next time the nodes that have currently lit up light up, CHAIR will also light up. For example, it might change the link from IS MADE OF WOOD to CHAIR to .3 (why doesn't it change it all the way to its maximum value? Because that erases all previous data and reduces the system's entire intelligence to what it learned on just this case).

On the other hand, IS SENTIENT is dark, so the training program might infer that IS SENTIENT is not a characteristic of chairs, and change the link strength there accordingly.

The next time the program sees a picture of a wooden chair, IS MADE OF WOOD will light up, and it will send its activation to IS CHAIR, making IS CHAIR light up with .3 units of activation: the program has a weak suspicion that the picture is a chair.

This is a pretty boring neural network, but if we add several hundred input nodes with all conceivable properties relevant to chairhood and spend a lot of computing power, eventually the program will become pretty good at recognizing chairs from nonchairs, and "learn" complicated rules that a three-legged wooden object is a stool which sort of counts as a chair, but a three legged sentient being is an injured dog and sitting on it will only make it angry.

Larger and more complicated neural nets contain "hidden nodes" - the equivalent of interneurons which sit between the input and the output and exist only to perform computations; feedback from an output node to a previous node that can create stable circles of activation, and other complications. They can perform much more difficult classification problems - identifying words from speech, or people from a photograph.

This is interesting because it solves a problem that baffled philosophers for millennia: the difficulty of coming up with good boundaries for categories. Plato famously defined Man as "a featherless biped"; Diogenes famously responded by presenting him with a shaved chicken. There seem to be many soft constraints on humans (can use language, have two legs, have a heartbeat) but there are also examples of humans who violate these constraints (babies, amputees, Dick Cheney) yet still seem obviously human.

Classical computers get bogged down in these problems, but neural nets naturally reason with "cluster structures in thing-space" and are expert classifiers in the same way we ourselves are.

SIMILARITIES BETWEEN NETS AND BRAINS

Even aside from their skill at classifying and pattern-matching, connectionist networks share many properties with brains, such as:

- Obvious structural similarities: neural nets work by lots of units which activate with different strengths and then spread that activation through links; the brain works by lots of neurons which fire at different rates and then spread that activation through axons.

- Lack of a "grandmother cell". A classical computer sticks each bit of memory in a particular location. A neural net stores memories as patterns of activation across all units in the network. In a feedback network, specific oft-repeated patterns can form attractor states to which the network naturally tends if pushed anywhere in the region. Association between one idea and another is not through physical contiguity, but through similarities in the pattern. "Grandmother" probably has most of the same neurons in the same state as "grandfather", and so it takes only a tiny stimulus to push the net from one attractor state to the other.

- Graceful failure: Classical computer programs do not fail gracefully; flip one bit, and the whole thing blows up and you have to spend the rest of your day messing around with a debugger. Destroying a few units in a neural net may only cost it a little bit of its processing power. This matches with the brain: losing a couple of neurons may make you think less clearly; losing a lot of neurons may give you dementia, memory loss and poor judgment. But there's no one neuron without which you just sit there near-catatonic, chanting "ERROR: NEURON 10559020481 NOT RESPONDING." And Karl Lashley can take out any part of a rat's cortex without affecting its memories too much.

- Remembering and forgetting: Neural nets can form memories, and the more the stimulus recurs to them the better they will remember it. But the longer they go without considering the stimulus, the more likely it is that the units involved in the memory-pattern will strengthen other connections, and then it will be harder to get them back in the memory pattern. This is much closer to how humans treat memory than the pristine, eternal encoding of classical computers.

- Ability to quickly locate solutions that best satisfy many soft constraints. What's a good place for dinner that's not too expensive, not more than twenty minutes away, serves decent cocktails, and has burgers for the kids? A classical computer would have to first identify the solution class as "restaurants", then search every restaurant it knows to see if they match each constraint, then fail to return an answer if no such restaurant exists. A neural net will just *settle* on the best answer, and if the cocktails there aren't really that good, it'll just settle but give the answer a lower strength.

- Context-sensitivity. Gold silver copper iron tin, and now when I say "lead", you're thinking of Element 82 (Pb), even though without the context a more natural interpretation is of the "leadership" variety. Currently active units can force others into a different pattern, giving context sensitivity not only to semantic priming as in the above example, but to emotions (people's thoughts follow different patterns when they're happy or sad), situations, and people.

Neural nets have also been used to simulate the results of many popular psychological experiments, including different types of priming, cognitive dissonance, and several of the biases and heuristics.

CONNECTIONISM AND REINFORCEMENT LEARNING

The link between connectionism and associationism is pretty obvious, but the link between connectionism and behaviorism is more elegant.

In most artificial neural nets, you need a training program to teach the net whether it's right or wrong and which way to adjust the weights. Brains don't have that luxury. Instead, part of their training algorithm for cognitive tasks is based on surprise: if you did not expect the sun to rise today, and you saw it rise anyway, you should probably decrease the strength of whatever links led you to that conclusion, and increase the strengths of any links that would have correctly predicted the sunrise.

Motivational links, however, could be modified by reinforcement. If a certain action leads to reward, strengthen the links that led to that action; if it leads to punishment, strengthen the links that would have made you avoid that action.

This explains behaviorist principles as a simple case of connectionism, the one where all the links are nice and straight, and you just have to worry about motivation and not about where cognition is coming from. Many of the animals typically studied by behaviorists were simple enough that this simple case was sufficient.

Although I think connectionism is our best current theory for how the mind works at a low level, it's hard to theorize about just because the networks are so complicated and so hard to simplify. Behaviorism is useful because it reduces the complexity of the networks to a few comprehensible rules, which allow higher level psychological theories and therapies to be derived from them.

Secrets of the eliminati

Anyone who does not believe mental states are ontologically fundamental - ie anyone who denies the reality of something like a soul - has two choices about where to go next. They can try reducing mental states to smaller components, or they can stop talking about them entirely.

In a utility-maximizing AI, mental states can be reduced to smaller components. The AI will have goals, and those goals, upon closer examination, will be lines in a computer program.

But in the blue-minimizing robot, its "goal" isn't even a line in its program. There's nothing that looks remotely like a goal in its programming, and goals appear only when you make rough generalizations from its behavior in limited cases.

Philosophers are still very much arguing about whether this applies to humans; the two schools call themselves reductionists and eliminativists (with a third school of wishy-washy half-and-half people calling themselves revisionists). Reductionists want to reduce things like goals and preferences to the appropriate neurons in the brain; eliminativists want to prove that humans, like the blue-minimizing robot, don't have anything of the sort until you start looking at high level abstractions.

I took a similar tack asking ksvanhorn's question in yesterday's post - how can you get a more accurate picture of what your true preferences are? I said:

> I don't think there are *true* preferences. In one situation you have one tendency, in another situation you have another tendency, and "preference" is what it looks like when you try to categorize tendencies. But categorization is a passive and not an active process: if every day of the week I eat dinner at 6, I can generalize to say "I prefer to eat dinner at 6", but it would be non-explanatory to say that a preference toward dinner at 6 caused my behavior on each day. I think the best way to salvage preferences is to consider them as tendencies currently in reflective equilibrium.

A more practical example: when people discuss cryonics or anti-aging, the following argument usually comes up in one form or another: if you were in a burning building, you would try pretty hard to get out. Therefore, you must strongly dislike death and want to avoid it. But if you strongly dislike death and want to avoid it, you must be lying when you say you accept death as a natural part of life and think it's crass and selfish to try to cheat the Reaper. And therefore your reluctance to sign up for cryonics violates your own revealed preferences! You must just be trying to signal conformity or something.

The problem is that not signing up for cryonics is also a "revealed preference". "You wouldn't sign up for cryonics, which means you don't really fear death so much, so why bother running from a burning building?" is an equally good argument, although no one except maybe Marcus Aurelius would take it seriously.

Both these arguments assume that somewhere, deep down, there's a utility function with a single term for "death" in it, and all decisions just call upon this particular level of death or anti-death preference.

More explanatory of the way people actually behave is that there's no unified preference for or against death, but rather a set of behaviors. Being in a burning building activates fleeing behavior; contemplating death from old age does not activate cryonics-buying behavior. People guess at their opinions about death by analyzing these behaviors, usually with a bit of signalling thrown in. If they desire consistency - and most people do - maybe they'll change some of their other behaviors to conform to their hypothesized opinion.

One more example. I've previously brought up the case of a rationalist who knows there's no such thing as ghosts, but is still uncomfortable in a haunted house. So does he believe in ghosts or not? If you insist on there being a variable somewhere in his head marked $belief_in_ghosts = (0,1) then it's going to be pretty mysterious when that variable looks like zero when he's talking to the Skeptics Association, and one when he's running away from a creaky staircase at midnight.

But it's not at all mysterious that the thought "I don't believe in ghosts" gets reinforced because it makes him feel intelligent and modern, and staying around a creaky staircase at midnight gets punished because it makes him afraid.

Behaviorism was one of the first and most successful eliminationist theories. I've so far ignored the most modern and exciting eliminationist theory, connectionism, because it involves a lot of math and is very hard to process on an intuitive level. In the next post, I want to try to explain the very basics of connectionism, why it's so exciting, and why it helps justify discussion of behaviorist principles.

Tendencies in reflective equilibrium

Consider a case, not too different from what has been shown to happen in reality, where we ask Bob what sounds like a fair punishment for a homeless man who steals $1,000, and he answers ten years. Suppose we wait until Bob has forgotten that we ever asked the first question, and then ask him what sounds like a fair punishment for a hedge fund manager who steals $1,000,000, and he says five years. Maybe we even wait until he forgets the whole affair, and then ask him the same questions again with the same answers, confirming that these are stable preferences.

If we now confront Bob with both numbers together, informing him that he supported a ten year sentence for stealing $1,000 and a five year sentence for stealing $1,000,000, a couple of things might happen. He could say "Yeah, I genuinely believe poor people deserve greater penalties than rich people." But more likely he says "Oh, I guess I was prejudiced." Then if we ask him the same question again, he comes up with two numbers that follow the expected mathematical relationship and punish the greater theft with more jail time.

Bob isn't working off of some predefined algorithm for determining punishment, like "jail time = (10 * amount stolen)/net worth". I don't know if anyone knows exactly what Bob is doing, but at a stab, he's seeing how many unpleasant feelings get generated by imagining the crime, then proposing a jail sentence that activates about an equal amount of unpleasant feelings. If the thought of a homeless man makes images of crime more readily available and so increases the unpleasant feelings, things won't go well for the homeless man. If you're really hungry, that probably won't help either.

So just like nothing automatically synchronizes the intention to study a foreign language and the behavior of studying it, so nothing automatically synchronizes thoughts about punishing the theft of $1000 and punishing the theft of $1000000.

Of course, there is something that non-automatically does it. After all, in order to elicit this strange behavior from Bob, we had to wait until he forgot about the first answer. Otherwise, he would have noticed and quickly adjusted his answers to make sense.

We probably could represent Bob's tendencies as an equation and call it a preference. Maybe it would be a long equation with terms for net worth of criminal, amount stolen, how much food Bob's eaten in the past six hours, and whether his local sports team won the pennant recently, with appropriate coefficients and powers for each. But if Bob saw this equation, he certainly wouldn't endorse it. He'd probably be horrified. It's also unstable: if given a choice, he would undergo brain surgery to remove this equation, thus preventing it from being satisfied. This is why I am reluctant to call these potential formalizations of these equations a "preference".

Instead of saying that Bob has one preference determining his jail time assignments, it would be better to model him as having several tendencies - a tendency to give a certain answer in the $1000 case, a tendency to give a different answer in the $1000000 case, and several tendencies towards things like consistency, fairness, compassion, et cetera.

People strongly consciously endorse these latter tendencies, probably because they're socially useful[1]. If the Chief of Police says "I know I just put this guy in jail for theft, but

I'm going to let this other thief off because he's my friend, and I don't really value consistency that much," then they're not going to stay Chief of Police for very long.

Bayesians and rationalists, in particular, make a big deal out of consistency. One common parable on the importance of consistency is the Dutch Book - a way to get free money from anyone behaving inconsistently. Suppose you have a weighted coin which can land on either heads or tails. There are several good reasons why I should not assign a probability of 66% to heads and 66% to tails, but one of the clearest is this: you can make me a bet that I will give you $2 if it lands on tails and you give me $1 if it lands on heads, and then a second bet where I give you $2 if it lands on heads and you give me $1 if it lands on tails. Whichever way the coin lands, I owe you $1 and you owe me $2 - I have gained a free dollar. So consistency is good if you don't want to be handing dollars out to random people...

...except that the Dutch book itself assumes consistency. If I believe that there is a 66% chance of it landing on heads, but refuse to take a bet at 2:1 odds - or even at 1.5:1 odds even though I should think it's easy money! - then I can't be Dutch booked. I am literally too stupid to be tricked effectively. You would think this wouldn't happen too often, since people would need to construct an accurate mental model to know when they should refuse such a bet, and such an accurate model would tell them they should revise their probabilities - but time after time people have demonstrated the ability to do exactly that.

I have not yet accepted that consistency is always the best course in every situation. For example, in Pascal's Mugging, a random person threatens to take away a zillion units of utility if you don't pay them $5. The probability they can make good on their threat is miniscule, but by multiplying out by the size of the threat, it still ought to motivate you to give the money. Some belief has to give - the belief that multiplication works, the belief that I shouldn't pay the money, or the belief that I should be consistent all the time - and right now, consistency seems like the weakest link in the chain.

The best we can do is seek reflective equilibrium among our tendencies. If you endorse the belief that rich people should not get lighter sentences than poor people more strongly than you endorse the tendency to give the homeless man ten years in jail and the fund manager five, then you can edit the latter tendency and come up with a "fair" sentence. This is Eliezer's defense of reason and philosophy, a powerful justification for morality (see part one here) and it's probably the best we can do in justifying our motivations as well.

Any tendency that has reached reflective equilibrium in your current state is about as close to a preference as you're going to get. It still won't automatically motivate you, of course. But you can motivate yourself toward it obliquely, and come up with the course of action that you most thoroughly endorse.

FOOTNOTES:

1: A tendency toward consistency can cause trouble if someone gains advantage from both of two mutually inconsistent ideas. Trivers' hypothesis predicts that people will consciously deny the inconsistency so they can continue holding both ideas, yet still remain consistent and so socially acceptable. Rationalists are so annoying because we go around telling people they can't do that.