Ethical Injunctions

Why Does Power Corrupt?

Followup to: Evolutionary Psychology

> "Power tends to corrupt, and absolute power corrupts absolutely. Great men are almost always bad men."
> —Lord Acton

Call it a just-so story if you must, but as soon as I was introduced to the notion of evolutionary psychology (~1995), it seemed obvious to me why human beings are corrupted by power. I didn't then know that hunter-gatherer bands tend to be more egalitarian than agricultural tribes—much less likely to have a central tribal-chief boss-figure—and so I thought of it this way:

Humans (particularly human males) have evolved to exploit power and status when they obtain it, for the obvious reason: If you use your power to take many wives and favor your children with a larger share of the meat, then you will leave more offspring, *ceteris paribus.* But you're not going to have much luck *becoming* tribal chief if you just go around saying, "Put me in charge so that I can take more wives and favor my children." You could lie about your reasons, but human beings are not perfect deceivers.

So one strategy that an evolution could follow, would be to create a vehicle that reliably tended to start believing that the old power-structure was corrupt, and that the good of the whole tribe required their overthrow...

The young revolutionary's belief is honest. There will be no betraying catch in his throat, as he explains why the tribe is doomed at the hands of the old and corrupt, unless he is given power to set things right. Not even subconsciously does he think, "And then, once I obtain power, I will strangely begin to resemble that old corrupt guard, abusing my power to increase my inclusive genetic fitness."

People often think as if "purpose" is an inherent property of things; and so many interpret the message of ev-psych as saying, "You have a subconscious, hidden goal to maximize your fitness." But individual organisms are adaptation-executers, not fitness-maximizers. The *purpose* that the revolutionary should obtain power and abuse it, is not a plan anywhere in his brain; it belongs to evolution, which can just barely be said to have purposes. It is a fact about many past revolutionaries having successfully taken power, having abused it, and having left many descendants.

When the revolutionary obtains power, he will find that it is sweet, and he will try to hold on to it—perhaps still thinking that this is for the good of the tribe. He will find that it seems *right* to take many wives (surely he deserves some reward for his labor) and to help his children (who are more deserving of help than others). But the young revolutionary has no foreknowledge of this in the beginning, when he sets out to overthrow the awful people who currently rule the tribe—evil mutants whose intentions are obviously much less good than his own.

The circuitry that will respond to power by finding it pleasurable, is already wired into our young revolutionary's brain; but he does not know this. (It would not help him evolutionarily if he did know it, because then he would not be able to honestly proclaim his good intentions—though it is scarcely necessary for evolution to prevent

hunter-gatherers from knowing about evolution, which is one reason we are able to know about it now.)

And so we have the awful cycle of "meet the new boss, same as the old boss". Youthful idealism rails against their elders' corruption, but oddly enough, the new generation—when it finally succeeds to power—doesn't seem to be all that morally purer. The original Communist Revolutionaries, I would guess probably a majority of them, really were in it to help the workers; but once they were a ruling Party in charge...

All sorts of random disclaimers can be applied to this thesis: For example, you could suggest that maybe Stalin's intentions weren't all that good to begin with, and that some politicians do intend to abuse power and really are just lying. A much more important objection is the need to redescribe this scenario in terms of power structures that actually exist in hunter-gatherer bands, which, as I understand it, have egalitarian pressures (among adult males) to keep any one person from getting too far above others.

But human beings do find power over others sweet, and it's not as if this emotion could have materialized from thin air, *without* an evolutionary explanation in terms of hunter-gatherer conditions. If you don't think this is why human beings are corrupted by power—then what's your evolutionary explanation? On the whole, to me at least, the evolutionary explanation for this phenomenon has the problem of not even seeming profound, because what it explains seems so normal.

The moral of this story, and the reason for going into the evolutionary explanation, is that you shouldn't reason as if people who are corrupted by power are evil mutants, whose mutations you do not share.

Evolution is not an infinitely powerful deceiving demon, and our ancestors evolved under conditions of not knowing about evolutionary psychology. The tendency to be corrupted by power can be beaten, I think. The "warp" doesn't seem on the same level of deeply woven insidiousness as, say, confirmation bias.

There was once an occasion where a reporter wrote about me, and did a hatchet job. It was my first time being reported on, and I was completely blindsided by it. I'd known that reporters sometimes wrote hatchet jobs, but I'd thought that it would require *malice*—I hadn't begun to imagine that someone might write a hatchet job just because it was a cliche, an easy way to generate a few column inches. So I drew upon my own powers of narration, and wrote an autobiographical story on what it felt like to be reported on for the first time—that horrible feeling of violation. I've never sent that story off anywhere, though it's a fine and short piece of writing as I judge it.

For it occurred to me, while I was writing, that journalism is an example of unchecked power—the reporter gets to present only one side of the story, any way they like, and there's nothing that the reported-on can do about it. (If you've never been reported on, then take it from me, that's how it is.) And here I was writing my own story, potentially for publication as traditional journalism, not in an academic forum. I remember realizing that the standards were *tremendously* lower than in science. That you could get away with damn near anything, so long as it made a good story—that this was the standard in journalism. (If you, having never been reported on yourself, don't believe me that this is the case, then you're as naive as I once was.)

Just that thought—not even the intention, not even wondering whether to do it, but just the *thought*—that I could present only my side of the story and deliberately make

the offending reporter look bad, and that no one would call me on it. Just that *thought* triggered this huge surge of positive reinforcement. This *tremendous* high, comparable to the high of discovery or the high of altruism.

And I knew right away what I was dealing with. So I sat there, motionless, fighting down that surge of positive reinforcement. It didn't go away just because I wanted it to go away. But it went away after a few minutes.

If I'd had no label to slap on that huge surge of positive reinforcement—if I'd been a less reflective fellow, flowing more with my passions—then that might have been that. People who are corrupted by power are not evil mutants.

I wouldn't call it a close call. I did know immediately what was happening. I fought it down without much trouble, and could have fought much harder if necessary. So far as I can tell, the temptation of unchecked power is not anywhere near as insidious as the labyrinthine algorithms of self-deception. Evolution is not an infinitely powerful deceiving demon. George Washington refused the temptation of the crown, and he didn't even know about evolutionary psychology. Perhaps it was enough for him to know a little history, and think of the temptation as a sin.

But it was still a scary thing to experience—this circuit that suddenly woke up and dumped a huge dose of unwanted positive reinforcement into my mental workspace, not when I *planned* to wield unchecked power, but just when my brain visualized the *possibility*.

To the extent you manage to fight off this temptation, you do not say: "Ah, now that I've beaten the temptation of power, I can safely make myself the wise tyrant who wields unchecked power benevolently, for the good of all." Having *successfully* fought off the temptation of power, you search for strategies that *avoid* seizing power. George Washington's triumph was not how well he ruled, but that he *refused the crown*—despite all temptation to be horrified at who else might then obtain power.

I am willing to admit of the theoretical possibility that someone could beat the temptation of power and then end up with *no ethical choice left*, except to grab the crown. But there would be a large burden of skepticism to overcome.

Part of the sequence *Ethical Injunctions*

Next post: "Ends Don't Justify Means (Among Humans)"

(start of sequence)

Ends Don't Justify Means (Among Humans)

> "If the ends don't justify the means, what does?"
> —variously attributed

> "I think of myself as running on hostile hardware."
> —Justin Corwin

Yesterday I talked about how humans may have evolved a structure of political revolution, beginning by believing themselves morally superior to the corrupt current power structure, but ending by being corrupted by power themselves—not by any plan in their own minds, but by the echo of ancestors who did the same and thereby reproduced.

This fits the template:

> In some cases, human beings have evolved in such fashion as to think that they are doing X for prosocial reason Y, but when human beings actually do X, other adaptations execute to promote self-benefiting consequence Z.

From this proposition, I now move on to my main point, a question *considerably* outside the realm of classical Bayesian decision theory:

> "What if I'm running on corrupted hardware?"

In such a case as this, you might even find yourself uttering such seemingly paradoxical statements—sheer nonsense from the perspective of classical decision theory—as:

> "The ends don't justify the means."

But if you are running on corrupted hardware, then the reflective observation that it *seems* like a righteous and altruistic act to seize power for yourself—this *seeming* may not be be much evidence for the proposition that seizing power is in fact the action that will most benefit the tribe.

By the power of naive realism, the corrupted hardware that you run on, and the corrupted seemings that it computes, will seem like the fabric of the very world itself—simply the way-things-are.

And so we have the bizarre-seeming rule: "For the good of the tribe, do not cheat to seize power *even when it would provide a net benefit to the tribe.*"

Indeed it may be wiser to phrase it this way: If you just say, "when it *seems* like it would provide a net benefit to the tribe", then you get people who say, "But it doesn't just *seem* that way—it *would* provide a net benefit to the tribe if I were in charge."

The notion of untrusted hardware seems like something wholly outside the realm of classical decision theory. (What it does to reflective decision theory I can't yet say, but that would seem to be the appropriate level to handle it.)

But on a human level, the patch seems straightforward. Once you know about the warp, you create rules that describe the warped behavior and outlaw it. A rule that says, "For the good of the tribe, do not cheat to seize power even for the good of the tribe." Or "For the good of the tribe, do not murder even for the good of the tribe."

And now the philosopher comes and presents their "thought experiment"—setting up a scenario in which, *by stipulation,* the *only* possible way to save five innocent lives is to murder one innocent person, and this murder is *certain* to save the five lives. "There's a train heading to run over five innocent people, who you can't possibly warn to jump out of the way, but you can push one innocent person into the path of the train, which will stop the train. These are your only options; what do you do?"

An altruistic human, who has accepted certain deontological prohibits—which seem well justified by some historical statistics on the results of reasoning in certain ways on untrustworthy hardware—may experience some mental distress, on encountering this thought experiment.

So here's a reply to that philosopher's scenario, which I have yet to hear any philosopher's victim give:

"You stipulate that the *only possible* way to save five innocent lives is to murder one innocent person, and this murder will *definitely* save the five lives, and that these facts are *known* to me with effective certainty. But since I am running on corrupted hardware, I can't occupy the *epistemic state* you want me to imagine. Therefore I reply that, in a society of Artificial Intelligences worthy of personhood and lacking any inbuilt tendency to be corrupted by power, it would be right for the AI to murder the one innocent person to save five, and moreover all its peers would agree. However, I refuse to extend this reply to myself, because the epistemic state you ask me to imagine, can only exist among other kinds of people than human beings."

Now, to me this seems like a dodge. I think the universe is sufficiently unkind that we can justly be forced to consider situations of this sort. The sort of person who goes around proposing that sort of thought experiment, might well deserve that sort of answer. But any human legal system does embody some answer to the question "How many innocent people can we put in jail to get the guilty ones?", even if the number isn't written down.

As a human, I try to abide by the deontological prohibitions that humans have made to live in peace with one another. But I don't think that our deontological prohibitions are *literally inherently nonconsequentially terminally right*. I endorse "the end doesn't justify the means" as a principle to guide humans running on corrupted hardware, but I wouldn't endorse it as a principle for a society of AIs that make well-calibrated estimates. (If you have one AI in a society of humans, that does bring in other considerations, like whether the humans learn from your example.)

And so I wouldn't say that a well-designed Friendly AI must necessarily refuse to push that one person off the ledge to stop the train. Obviously, I would expect any decent superintelligence to come up with a superior third alternative. But if those are the only two alternatives, and the FAI judges that it is wiser to push the one person off the ledge—even after taking into account knock-on effects on any humans who see it happen and spread the story, etc.—then I don't call it an alarm light, if an *AI* says that the right thing to do is sacrifice one to save five. Again, *I* don't go around pushing people into the paths of trains myself, nor stealing from banks to fund my altruistic projects. *I* happen to be a human. But for a Friendly AI to be corrupted by power

would be like it starting to bleed red blood. The tendency to be corrupted by power is a specific biological adaptation, supported by specific cognitive circuits, built into us by our genes for a clear evolutionary reason. It wouldn't spontaneously appear in the code of a Friendly AI any more than its transistors would start to bleed.

I would even go further, and say that if you had minds with an inbuilt warp that made them *overestimate* the external harm of self-benefiting actions, then they would need a rule "the ends do not prohibit the means"—that you should do what benefits yourself even when it (seems to) harm the tribe. By hypothesis, if their society did not have this rule, the minds in it would refuse to breathe for fear of using someone else's oxygen, and they'd all die. For them, an occasional overshoot in which one person seizes a personal benefit at the net expense of society, would seem just as cautiously virtuous—and indeed *be* just as cautiously virtuous—as when one of us humans, being cautious, passes up an opportunity to steal a loaf of bread that really would have been more of a benefit to them than a loss to the merchant (including knock-on effects).

"The end does not justify the means" is just consequentialist reasoning at one meta-level up. If a human starts thinking on the *object* level that the end justifies the means, this has awful consequences given our untrustworthy brains; therefore a human shouldn't think this way. But it is all still ultimately consequentialism. It's just *reflective* consequentialism, for beings who know that their moment-by-moment decisions are made by untrusted hardware.

Entangled Truths, Contagious Lies

> One of your very early philosophers came to the conclusion that a fully competent mind, from a study of one fact or artifact belonging to any given universe, could construct or visualize that universe, from the instant of its creation to its ultimate end . . .
>
> —*First Lensman*
>
> If any one of you will concentrate upon one single fact, or small object, such as a pebble or the seed of a plant or other creature, for as short a period of time as one hundred of your years, you will begin to perceive its truth.
>
> —*Gray Lensman*

I am reasonably sure that a single pebble, taken from a beach of our own Earth, does not specify the continents and countries, politics and people of this Earth. Other planets in space and time, other Everett branches, would generate the same pebble.

On the other hand, the identity of a single pebble would seem to include our laws of physics. In that sense the entirety of our Universe—*all* the Everett branches—would be implied by the pebble.[1]

From the study of that single pebble you could see the laws of physics and all they imply. Thinking about those laws of physics, you can see that planets will form, and you can guess that the pebble came from such a planet. The internal crystals and molecular formations of the pebble developed under gravity, which tells you something about the planet's mass; the mix of elements in the pebble tells you something about the planet's formation.

I am not a geologist, so I don't know to which mysteries geologists are privy. But I find it very easy to imagine showing a geologist a pebble, and saying, "This pebble came from a beach at Half Moon Bay," and the geologist immediately says, "I'm confused," or even, "You liar." Maybe it's the wrong kind of rock, or the pebble isn't worn enough to be from a beach—I don't know pebbles well enough to guess the linkages and signatures by which I might be caught, which is the point.

"Only God can tell a truly plausible lie." I wonder if there was ever a religion that developed this as a proverb? I would (falsifiably) guess not: it's a rationalist sentiment, even if you cast it in theological metaphor. Saying "everything is interconnected to everything else, because God made the whole world and sustains it" may generate some nice warm 'n' fuzzy feelings during the sermon, but it doesn't get you very far when it comes to assigning pebbles to beaches.

A penny on Earth exerts a gravitational acceleration on the Moon of around 4.5×10^{-31} m/s^2, so in one sense it's not too far wrong to say that every event is entangled with its whole past light cone. And since inferences can propagate backward and forward through causal networks, *epistemic* entanglements can easily cross the borders of light cones. But I wouldn't want to be the forensic astronomer who had to look at the Moon and figure out whether the penny landed heads or tails—the influence is far less than quantum uncertainty and thermal noise.

If you said, “Everything is entangled with something else,” or, “Everything is inferentially entangled and some entanglements are much stronger than others,” you might be really wise instead of just Deeply Wise.

Physically, each event is in some sense the sum of its whole past light cone, without borders or boundaries. But the list of *noticeable* entanglements is much shorter, and it gives you something like a network. This high-level regularity is what I refer to when I talk about the Great Web of Causality.

I use these Capitalized Letters somewhat tongue-in-cheek, perhaps; but if anything at all is worth Capitalized Letters, surely the Great Web of Causality makes the list.

“Oh what a tangled web we weave, when first we practise to deceive,” said Sir Walter Scott. Not *all* lies spin out of control—we don’t live in so righteous a universe. But it does occasionally happen that someone lies about a fact, and then has to lie about an entangled fact, and then another fact entangled with that one:

> “Where were you?”
>
> “Oh, I was on a business trip.”
>
> “What was the business trip about?”
>
> “I can’t tell you that; it’s proprietary negotiations with a major client.”
>
> “Oh—they’re letting you in on those? Good news! I should call your boss to thank him for adding you.”
>
> “Sorry—he’s not in the office right now . . .”

Human beings, who are not gods, often fail to *imagine* all the facts they would need to distort to tell a truly plausible lie. “God made me pregnant” sounded a tad more likely in the old days before our models of the world contained (quotations of) Y chromosomes. Many similar lies, today, may blow up when genetic testing becomes more common. Rapists have been convicted, and false accusers exposed, years later, based on evidence they didn’t realize they could leave. A student of evolutionary biology can see the design signature of natural selection on every wolf that chases a rabbit; and every rabbit that runs away; and every bee that stings instead of broadcasting a polite warning—but the deceptions of creationists sound plausible to *them*, I’m sure.

Not all lies are uncovered, not all liars are punished; we don’t live in that righteous a universe. But not all lies are as safe as their liars believe. How many sins would become known to a Bayesian superintelligence, I wonder, if it did a (non-destructive?) nanotechnological scan of the Earth? At minimum, all the lies of which any evidence still exists in any brain. Some such lies may become known sooner than that, if the neuroscientists ever succeed in building a really good lie detector via neuroimaging. Paul Ekman (a pioneer in the study of tiny facial muscle movements) could probably read off a sizeable fraction of the world’s lies right now, given a chance.

Not all lies are uncovered, not all liars are punished. But the Great Web is very commonly underestimated. Just the knowledge that humans have *already accumulated* would take many human lifetimes to learn. Anyone who thinks that a non-God can tell a *perfect* lie, risk-free, is underestimating the tangledness of the Great Web.

Is honesty the best policy? I don’t know if I’d go that far: Even on my ethics, it’s sometimes okay to shut up. But compared to outright lies, either honesty or silence involves less exposure to recursively propagating risks you don’t know you’re taking.

[1] Assuming, as seems likely, there are no truly free variables.

Protected From Myself

Followup to: The Magnitude of His Own Folly, Entangled Truths, Contagious Lies

Every now and then, another one comes before me with the brilliant idea: "Let's lie!"

Lie about what?—oh, various things. The expected time to Singularity, say. Lie and say it's definitely going to be earlier, because that will get more public attention. Sometimes they say "be optimistic", sometimes they just say "lie". Lie about the current degree of uncertainty, because there are other people out there claiming to be certain, and the most unbearable prospect in the world is that someone else pull ahead. Lie about what the project is likely to accomplish—I flinch even to write this, but occasionally someone proposes to go and say to the Christians that the AI will create Christian Heaven forever, or go to the US government and say that the AI will give the US dominance forever.

But at any rate, lie. Lie because it's more convenient than trying to explain the truth. Lie, because someone else might lie, and so we have to make sure that we lie first. Lie to grab the tempting benefits, hanging just within reach—

Eh? Ethics? Well, now that you mention it, lying is at least a little bad, all else being equal. But with so much at stake, we should just ignore that and lie. You've got to follow the expected utility, right? The loss of a lie is much less than the benefit to be gained, right?

Thus do they argue. Except—what's the flaw in the argument? Wouldn't it be *irrational* not to lie, if lying has the greatest expected utility?

When I look back upon my history—well, I screwed up in a lot of ways. But it could have been *much worse,* if I had reasoned like those who offer such advice, and lied.

Once upon a time, I truly and honestly believed that either a superintelligence would do what was right, or else there was no right thing to do; and I said so. I was uncertain of the nature of morality, and I said that too. I didn't know if the Singularity would be in five years or fifty, and this also I admitted. My project plans were not guaranteed to deliver results, and I did not promise to deliver them. When I finally said "Oops", and realized that I needed to go off and do more fundamental research instead of rushing to write code immediately—

—well, I can imagine the mess I would have had on my hands, if I had told the people who trusted me: that the Singularity was surely coming in ten years; that my theory was sure to deliver results; that I had no lingering confusions; and that any superintelligence would *surely* give them their own private island and a harem of catpersons of the appropriate gender. How exactly would one then explain why you're now going to step back and look for math-inventors instead of superprogrammers, or why the code now has to be theorem-proved?

When you make an honest mistake, on some subject you were honest *about,* the recovery technique is straightforward: Just as you told people what you thought in the first place, you now list out the actual reasons that you changed your mind. This diff takes you to your current true thoughts, that imply your current desired policy. Then, just as people decided whether to aid you originally, they re-decide in light of the new information.

But what if you were "optimistic" and only presented one side of the story, the better to fulfill that all-important goal of persuading people to your cause? Then you'll have a much harder time persuading them *away* from that idea you sold them originally—you've nailed their feet to the floor, which makes it difficult for them to follow if you yourself take another step forward.

And what if, for the sake of persuasion, you told them things that you didn't believe yourself? Then there is no true diff from the story you told before, to the new story now. Will there be any coherent story that explains your change of heart?

Conveying the real truth is an art form. It's not an easy art form—those darned constraints of honesty prevent you from telling all kinds of convenient lies that would be so much easier than the complicated truth. But, if you tell lots of truth, you get good at what you practice. A lot of those who come to me and advocate lies, talk earnestly about how these matters of transhumanism are *so hard* to explain, too difficult and technical for the likes of Joe the Plumber. So they'd like to take the easy way out, and lie.

We don't live in a righteous universe where all sins are punished. Someone who *practiced* telling lies, and made their mistakes and learned from them, might well become expert at telling lies that allow for sudden changes of policy in the future, and telling more lies to explain the policy changes. If you use the various forbidden arts that create fanatic followers, they will swallow just about anything. The history of the Soviet Union and their sudden changes of policy, as presented to their ardent Western intellectual followers, helped inspire Orwell to write 1984.

So the question, really, is whether you want to practice truthtelling or practice lying, because whichever one you practice is the one you're going to get good at. Needless to say, those who come to me and offer their unsolicited advice do not appear to be expert liars. For one thing, a majority of them don't seem to find anything odd about floating their proposals in publicly archived, Google-indexed mailing lists.

But why *not* become an expert liar, if that's what maximizes expected utility? Why take the constrained path of truth, when things so much more important are at stake?

Because, when I look over my history, I find that **my ethics have, above all, protected me from *myself.*** They weren't inconveniences. They were safety rails on cliffs I didn't see.

I made fundamental mistakes, and my ethics didn't halt that, but they played a critical role in my recovery. **When I was stopped by unknown unknowns that I just wasn't expecting, it was my ethical constraints, and not any conscious planning, that had put me in a recoverable position.**

You can't duplicate this protective effect by trying to be clever and calculate the course of "highest utility". The expected utility just takes into account the things you *know* to expect. It really is amazing, looking over my history, the extent to which my ethics put me in a recoverable position from my unanticipated, *fundamental* mistakes, the things completely outside my plans and beliefs.

Ethics aren't just there to make your life difficult; they can protect you from Black Swans. A startling assertion, I know, but not one entirely irrelevant to current affairs.

If you've been following along my story, you'll recall that the downfall of all my theories, began with a tiny note of discord. A tiny note that I wouldn't ever have

followed up, if I had only cared about my own preferences and desires. It was the thought of what someone else might think—someone to whom I felt I owed an ethical consideration—that spurred me to follow up that one note.

And I have watched others fail utterly on the problem of Friendly AI, because they simply try to grab the banana in one form or another—seize the world for their own favorite moralities, without any thought of what others might think—and so they never enter into the complexities and second thoughts that might begin to warn them of the *technical* problems.

We don't live in a righteous universe. And so, when I look over my history, the role that my ethics have played is so important that I've had to take a step back and ask, "*Why is this happening?*" The universe *isn't* set up to reward virtue—so why did my ethics help so much? Am I only imagining the phenomenon? That's one possibility. But after some thought, I've concluded that, to the extent you believe that my ethics *did* help me, these are the plausible reasons in order of importance:

1) **The honest Way often has a kind of simplicity that trangressions lack**. If you tell lies, you have to keep track of different stories you've told different groups, and worry about which facts might encounter the wrong people, and then invent new lies to explain any unexpected policy shifts you have to execute on account of your mistake. This *simplicity* is powerful enough to explain a great deal of the positive influence that I attribute to my ethics, in a universe that doesn't reward virtue *per se.*

2) **I was stricter with myself, and held myself to a higher standard, when I was doing various things that I considered myself ethically obligated to do**. Thus my recovery from various failures often seems to have begun with an ethical thought of some type—e.g. the whole development where "Friendly AI" led into the concept of AI as a precise art. That might just be a quirk of my own personality; but it seems to help account for the huge role my ethics played in leading me to important thoughts, which I cannot just explain by saying that the universe rewards virtue.

3) **The constraints that the wisdom of history suggests, to avoid hurting other people, may also stop you from hurting yourself.** When you have some brilliant idea that benefits the tribe, we don't want you to run off and do X, Y, and Z, even if you say "the end justifies the means!" Evolutionarily speaking, one suspects that the "means" have more often benefited the person who executes them, than the tribe. But this is not the ancestral environment. In the more complicated modern world, following the ethical constraints can prevent you from making huge networked mistakes that would catch you in their collapse. Robespierre led a shorter life than Washington.

Part of the sequence *Ethical Injunctions*

Next post: "Ethical Inhibitions"

Previous post: "Ends Don't Justify Means (Among Humans)"

Ethical Inhibitions

Followup to: Entangled Truths, Contagious Lies, Evolutionary Psychology

What's up with that bizarre emotion we humans have, this sense of *ethical caution?*

One can understand sexual lust, parental care, and even romantic attachment. The evolutionary psychology of such emotions might be subtler than it at first appears, but if you ignore the subtleties, the surface reasons are obvious. But why a sense of ethical caution? Why honor, why righteousness? (And no, it's not group selection; it never is.) What reproductive benefit does that provide?

The *specific* ethical codes that people feel uneasy violating, vary from tribe to tribe (though there are certain regularities). But the *emotion* associated with *feeling ethically inhibited*—well, I Am Not An Evolutionary Anthropologist, but that looks like a human universal to me, something with brainware support.

The obvious story behind prosocial emotions in general, is that those who offend against the group are sanctioned; this converts the emotion to an *individual* reproductive advantage. The human organism, executing the ethical-caution adaptation, ends up avoiding the group sanctions that would follow a violation of the code. This obvious answer may even be the entire answer.

But I suggest—if a bit more tentatively than usual—that by the time human beings were evolving the emotion associated with "ethical inhibition", we were already intelligent enough to observe the existence of such things as group sanctions. We were already smart enough (I suggest) to model what the group would punish, and to fear that punishment.

Sociopaths have a concept of getting caught, and they try to avoid getting caught. Why isn't this sufficient? Why have an *extra* emotion, a feeling that inhibits you even when you *don't* expect to be caught? Wouldn't this, from evolution's perspective, just result in passing up perfectly good opportunities?

So I suggest (tentatively) that humans naturally underestimate the odds of getting caught. We don't foresee all the possible chains of causality, all the entangled facts that can bring evidence against us. Those ancestors who lacked a sense of ethical caution stole the silverware *when they expected that no one would catch them or punish them;* and were *nonetheless* caught or punished often enough, on average, to outweigh the value of the silverware.

Admittedly, this may be an unnecessary assumption. It is a general idiom of biology that evolution is the only long-term consequentialist; organisms compute short-term rewards. Hominids violate this rule, but that is a very recent innovation.

So one could counter-argue: "Early humans didn't reliably forecast the punishment that follows from breaking social codes, so they didn't reliably think consequentially about it, so they developed an instinct to obey the codes." Maybe the modern sociopaths that *evade being caught* are smarter than average. Or modern sociopaths are better educated than hunter-gatherer sociopaths. Or modern sociopaths get more second chances to recover from initial stumbles—they can change their name and move. It's not so strange to find an emotion executing in some exceptional circumstance where it fails to provide a reproductive benefit.

But I feel justified in bringing up the more complicated hypothesis, because ethical inhibitions are *archetypally*that which stops us even when we think no one is looking. A humanly universal concept, so far as I know, though I am not an anthropologist.

Ethical inhibition, as a human motivation, seems to be implemented in a distinct style from hunger or lust. Hunger and lust can be outweighed when stronger desires are at stake; but the emotion associated with ethical prohibitions tries to assert itself *deontologically*. If you have the sense at all that you shouldn't do it, you have the sense that you unconditionally shouldn't do it. The emotion associated with ethical caution would seem to be a drive that—successfully or unsuccessfully—tries to *override* the temptation, not just *weigh against* it.

A monkey can be trapped by a food reward inside a hollowed shell—they can reach in easily enough, but once they close their fist, they can't take their hand out. The monkey may be screaming with distress, and still be unable to override the instinct to keep hold of the food. We humans can do better than that; we can let go of the food reward and run away, when our brain is warning us of the long-term consequences.

But why does the sensation of *ethical inhibition,* that might also command us to pass up a food reward, have a similar override-quality—even in the absence of explicitly expected long-term consequences? Is it just that ethical emotions evolved recently, and happen to be implemented in prefrontal cortex next to the long-term-override circuitry?

What is this tendency to *feel* inhibited from stealing the food reward? This message that tries to assert "I override", not just "I weigh against"? Even when we don't expect the long-term consequences of being discovered?

And before you think that I'm falling prey to some kind of appealing story, ask yourself why that particular story would sound appealing to humans. Why would it seem temptingly *virtuous* to let an ethical inhibition override, rather than just being one more weight in the balance?

One possible explanation would be if the emotion were carved out by the evolutionary-historical statistics of a *black-swan bet.*

Maybe you will, in all probability, get away with stealing the silverware on any particular occasion—just as your model of the world would extrapolate. But it was a statistical fact about your ancestors that sometimes the environment didn't operate the way they expected. Someone was watching from behind the trees. On those occasions their reputation was permanently blackened; they lost status in the tribe, and perhaps were outcast or murdered. Such occasions could be statistically rare, and still counterbalance the benefit of a few silver spoons.

The brain, like every other organ in the body, is a reproductive organ: it was carved out of entropy by the persistence of mutations that promoted reproductive fitness. And yet somehow, amazingly, the human brain wound up with circuitry for such things as honor, sympathy, and ethical resistance to temptations.

Which means that those alleles *drove their alternatives to extinction.* Humans, the organisms, can be nice to each other; but the alleles' game of frequencies is zero-sum. Honorable ancestors didn't necessarily kill the dishonorable ones. But if, by cooperating with each other, honorable ancestors outreproduced less honorable folk, then the honor allele killed the dishonor allele as surely as if it erased the DNA sequence off a blackboard.

That might be something to think about, the next time you're wondering if you should just give in to your ethical impulses, or try to override them with your rational awareness.

Especially if you're tempted to engage in some chicanery "for the greater good"—tempted to decide that the end justifies the means. Evolution doesn't care about whether something actually promotes the greater good—that's not how gene frequencies change. But if transgressive plans go awry often enough to hurt *the transgressor*, how much *more* often would they go awry and hurt the intended beneficiaries?

Historically speaking, it seems likely that, of those who set out to rob banks or murder opponents "in a good cause", those who managed to hurt *themselves,* mostly wouldn't make the history books. (Unless they got a second chance, like Hitler after the failed Beer Hall Putsch.) Of those cases we *do* read about in the history books, many people have done very well for *themselves* out of their plans to lie and rob and murder "for the greater good". But how many people cheated their way to *actual* huge altruistic benefits—cheated and *actually* realized the justifying greater good? Surely there must be at least one or two cases known to history—at least one king *somewhere* who took power by lies and assassination, and then ruled wisely and well—but I can't actually name a case off the top of my head. By and large, it seems to me a pretty fair generalization that people who achieve great good ends manage *not* to find excuses for all that much evil along the way.

Somehow, people seem much more likely to endorse plans that involve just a little pain for someone *else*, on behalf of the greater good, than to work out a way to let the sacrifice be themselves. But when you plan to damage society in order to save it, remember that your brain contains a sense of ethical unease that evolved from transgressive plans blowing up and damaging the *originator*—never mind the expected value of all the damage done to *other* people, if you really do care about them.

If natural selection, which doesn't care *at all* about the welfare of unrelated strangers, still manages to give you a sense of ethical unease on account of transgressive plans not always going as planned—then how much *more* reluctant should you be to rob banks for a good cause, if you aspire to *actually* help and protect others?

Part of the sequence *Ethical Injunctions*

Next post: "Ethical Injunctions"

Previous post: "Protected From Myself"

Ethical Injunctions

> "Would you kill babies if it was the right thing to do? If no, under what circumstances would you not do the right thing to do? If yes, how right would it have to be, for how many babies?"
>
> —horrible job interview question

Swapping hats for a moment, I'm *professionally* intrigued by the decision theory of "things you shouldn't do even if they seem to be the right thing to do".

Suppose we have a reflective AI, self-modifying and self-improving, at an intermediate stage in the development process. In particular, the AI's goal system isn't finished—the shape of its motivations is still being loaded, learned, tested, or tweaked.

Yea, I have seen many ways to screw up an AI goal system design, resulting in a decision system that decides, given its goals, that the universe ought to be tiled with tiny molecular smiley-faces, or some such. Generally, these deadly suggestions also have the property that the AI will not desire its programmers to fix it. If the AI is *sufficiently* advanced—which it may be even at an intermediate stage—then the AI may also realize that deceiving the programmers, hiding the changes in its thoughts, will help transform the universe into smiley-faces.

Now, from our perspective as programmers, if we *condition on the fact* that the AI has decided to hide its thoughts from the programmers, or otherwise act willfully to deceive us, then it would seem likely that some kind of unintended consequence has occurred in the goal system. We would consider it probable that the AI is *not* functioning as intended, but rather likely that we have messed up the AI's utility function somehow. So that the AI wants to turn the universe into tiny reward-system counters, or some such, and now has a motive to hide from us.

Well, suppose we're *not* going to implement some object-level Great Idea as the AI's utility function. Instead we're going to do something advanced and recursive—build a goal system which knows (and cares) about the programmers outside. A goal system that, via some nontrivial internal structure, "knows it's being programmed" and "knows it's incomplete". Then you might be able to have and keep the rule:

> "If [I decide that] fooling my programmers is the right thing to do, execute a controlled shutdown [instead of doing the right thing to do]."

And the AI would keep this rule, even through the self-modifying AI's revisions of its own code, because, in its structurally nontrivial goal system, the present-AI understands that this decision by a future-AI *probably* indicates something defined-as-a-malfunction. Moreover, the present-AI knows that if future-AI tries to *evaluate* the utility of executing a shutdown, once this hypothetical malfunction has occurred, the future-AI will probably *decide* not to shut itself down. So the shutdown should happen unconditionally, automatically, without the goal system getting another chance to recalculate the right thing to do.

I'm not going to go into the deep dark depths of the exact mathematical structure, because that would be beyond the scope of this blog. Also I don't yet know the deep dark depths of the mathematical structure. It looks like it *should* be possible, if you do things that are advanced and recursive and have nontrivial (but consistent) structure. But I haven't reached that level, as yet, so for now it's only a dream.

But the topic here is not advanced AI; it's human ethics. I introduce the AI scenario to bring out more starkly the strange idea of an *ethical injunction:*

> You should never, ever murder an innocent person who's helped you, *even if it's the right thing to do;* because it's far more likely that *you've made a mistake*, than that murdering an innocent person who helped you is the right thing to do.

Sound reasonable?

During World War II, it became necessary to destroy Germany's supply of deuterium, a neutron moderator, in order to block their attempts to achieve a fission chain reaction. Their supply of deuterium was coming at this point from a captured facility in Norway. A shipment of heavy water was on board a Norwegian ferry ship, the SF Hydro. Knut Haukelid and three others had slipped on board the ferry in order to sabotage it, when the saboteurs were discovered by the ferry watchman. Haukelid told him that they were escaping the Gestapo, and the watchman immediately agreed to overlook their presence. Haukelid "considered warning their benefactor but decided that might endanger the mission and only thanked him and shook his hand." (Richard Rhodes, *The Making of the Atomic Bomb.*) So the civilian ferry *Hydro* sank in the deepest part of the lake, with eighteen dead and twenty-nine survivors. Some of the Norwegian rescuers felt that the German soldiers present should be left to drown, but this attitude did not prevail, and four Germans were rescued. And that was, effectively, the end of the Nazi atomic weapons program.

Good move? Bad move? Germany *very likely* wouldn't have gotten the Bomb anyway... I hope with absolute desperation that I never get faced by a choice like that, but in the end, I can't say a word against it.

On the other hand, when it comes to the rule:

> "Never try to deceive yourself, or offer a reason to believe other than probable truth; because even if you come up with an amazing clever reason, it's more likely that you've made a mistake than that you have a reasonable expectation of this being a net benefit in the long run."

Then I really *don't* know of anyone who's knowingly been faced with an exception. There are times when you try to convince yourself "I'm not hiding any Jews in my basement" before you talk to the Gestapo officer. But then you do still know the truth, you're just trying to create something like an alternative self that exists in your imagination, a facade to talk to the Gestapo officer.

But to really believe something that isn't true? I don't know if there was ever anyone for whom that was *knowably* a good idea. I'm sure that there have been many many times in human history, where person X was better off with false belief Y. And by the same token, there is always some set of winning lottery numbers in every drawing. It's *knowing which lottery ticket will win* that is the epistemically difficult part, like X knowing when he's better off with a false belief.

Self-deceptions are the worst kind of black swan bets, much worse than lies, because without knowing the true state of affairs, you can't even guess at what the penalty will be for your self-deception. They only have to blow up once to undo all the good they ever did. One single time when you pray to God after discovering a lump, instead of going to a doctor. That's all it takes to undo a life. All the happiness that the warm thought of an afterlife ever produced in humanity, has now been more than cancelled by the failure of humanity to institute systematic cryonic preservations after liquid

nitrogen became cheap to manufacture. And I don't think that anyone ever had that sort of failure in mind as a possible blowup, when they said, "But we need religious beliefs to cushion the fear of death." That's what black swan bets are all about—the unexpected blowup.

Maybe you even get away with one or two black-swan bets—they don't get you *every* time. So you do it again, and then the blowup comes and cancels out every benefit and then some. That's what black swan bets are all about.

Thus the difficulty of knowing when it's safe to believe a lie (assuming you can even manage that much mental contortion in the first place)—part of the nature of black swan bets is that you don't see the bullet that kills you; and since our perceptions just seem like the way the world is, it looks like there is no bullet, period.

So I would say that there is an ethical injunction against self-deception. I call this an "ethical injunction" not so much because it's a matter of interpersonal morality (although it is), but because it's a rule that guards you from your own cleverness—an override against the temptation to do what seems like the right thing.

So now we have two kinds of situation that can support an "ethical injunction", a rule not to do something even when it's the right thing to do. (That is, you refrain "even when your brain has computed it's the right thing to do", but this will just *seem like* "the right thing to do".)

First, being human and running on corrupted hardware, we may generalize classes of situation where when you say e.g. "It's time to rob a few banks for the greater good," we deem it more likely that you've been corrupted than that this is really the case. (Note that we're not prohibiting it from *ever* being the case in *reality,* but we're questioning the *epistemic* state where you're *justified in trusting* your own calculation that this is the right thing to do—fair lottery tickets can win, but you can't justifiably buy them.)

Second, history may teach us that certain classes of action are black-swan bets, that is, they sometimes blow up bigtime for reasons not in the decider's model. So even when we calculate within the model that something seems like the right thing to do, we apply the further knowledge of the black swan problem to arrive at an injunction against it.

But surely... if one is *aware of these reasons...* then one can simply redo the calculation, taking them into account. So we can rob banks if it seems like the right thing to do *after taking into account* the problem of corrupted hardware and black swan blowups. That's the rational course, right?

There's a number of replies I could give to that.

I'll start by saying that this is a prime example of the sort of thinking I have in mind, when I warn aspiring rationalists to beware of cleverness.

I'll also note that I wouldn't want an attempted Friendly AI that had just decided that the Earth ought to be transformed into paperclips, to assess whether this was a reasonable thing to do in light of all the various warnings it had received against it. I would want it to undergo an automatic controlled shutdown. Who says that meta-reasoning is immune from corruption?

I could mention the important times that my naive, idealistic ethical inhibitions have protected me from *myself*, and placed me in a recoverable position, or helped start the recovery, from very deep mistakes I had no clue I was making. And I could ask whether I've really advanced so much, and whether it would really be all that wise, to remove the protections that saved me before.

Yet even so... "Am I still dumber than my ethics?" is a question whose answer isn't *automatically* "Yes."

There are obvious silly things here that you shouldn't do; for example, you shouldn't wait until you're really tempted, and *then* try to figure out if you're smarter than your ethics on that particular occasion.

But in general—there's only so much power that can vest in what your parents told you not to do. One shouldn't underestimate the power. Smart people debated historical lessons in the course of forging the Enlightenment ethics that much of Western culture draws upon; and some subcultures, like scientific academia, or science-fiction fandom, draw on those ethics more directly. But even so the power of the past is bounded.

And in fact...

I've had to make my ethics *much stricter* than what my parents and Jerry Pournelle and Richard Feynman told me not to do.

Funny thing, how when people seem to think they're smarter than their ethics, they argue for *less* strictness rather than *more* strictness. I mean, when you think about how much more complicated the modern world is...

And along the same lines, the ones who come to me and say, "You should lie about the Singularity, because that way you can get more people to support you; it's the rational thing to do, for the greater good"—these ones seem to have *no idea* of the risks.

They don't mention the problem of running on corrupted hardware. They don't mention the idea that lies have to be recursively protected from all the truths and all the truthfinding techniques that threaten them. They don't mention that honest ways have a simplicity that dishonest ways often lack. They don't talk about black-swan bets. They don't talk about the terrible nakedness of discarding the last defense you have against yourself, and trying to survive on raw calculation.

I am reasonably sure that this is because they have *no clue* about any of these things.

If you've truly understood the reason and the rhythm behind ethics, then one major sign is that, augmented by this newfound knowledge, you *don't do* those things that previously seemed like ethical transgressions. Only now you know why.

Someone who just looks at one or two reasons behind ethics, and says, "Okay, I've understood that, so now I'll take it into account consciously, and therefore I have no more need of ethical inhibitions"—this one is behaving more like a stereotype than a real rationalist. The world isn't simple and pure and clean, so you can't just take the ethics you were raised with and trust them. But that pretense of Vulcan logic, where you think you're just going to compute everything correctly once you've got one or two abstract insights—that doesn't work in real life either.

As for those who, having figured out *none* of this, think themselves smarter than their ethics: Ha.

And as for those who previously thought themselves smarter than their ethics, but who hadn't conceived of all these elements behind ethical injunctions "in so many words" until they ran across this *Overcoming Bias* sequence, and who *now* think themselves smarter than their ethics, because they're going to take all this into account from now on: Double ha.

I have seen many people struggling to excuse themselves from their ethics. Always the modification is toward lenience, never to be more strict. And I am stunned by the speed and the lightness with which they strive to abandon their protections. Hobbes said, "I don't know what's worse, the fact that everyone's got a price, or the fact that their price is so low." So very low the price, so very eager they are to be bought. They don't look twice and then a third time for alternatives, before deciding that they have no option left but to transgress—though they may look very grave and solemn when they say it. They abandon their ethics at the very first opportunity. "Where there's a will to failure, obstacles can be found." The will to fail at ethics seems very strong, in some people.

I don't know if I can endorse absolute ethical injunctions that bind over all possible epistemic states of a human brain. The universe isn't kind enough for me to trust that. (Though an ethical injunction against self-deception, for example, does seem to me to have tremendous force. I've seen many people arguing for the Dark Side, and none of them seem aware of the network risks or the black-swan risks of self-deception.) If, someday, I attempt to shape a (reflectively consistent) injunction within a self-modifying AI, it will only be after working out the math, because that is so totally not the sort of thing you could get away with doing via an ad-hoc patch.

But I will say this much:

I am completely unimpressed with the knowledge, the reasoning, and the overall level, of those folk who have eagerly come to me, and said in grave tones, "It's rational to do unethical thing X because it will have benefit Y."

Prices or Bindings?

Followup to: Ethical Injunctions

During World War II, Knut Haukelid and three other saboteurs sank a civilian Norwegian ferry ship, the SF Hydro, carrying a shipment of deuterium for use as a neutron moderator in Germany's atomic weapons program. Eighteen dead, twenty-nine survivors. And that was the end of the Nazi nuclear program. Can you imagine a Hollywood movie in which the hero did that, instead of coming up with some amazing clever way to save the civilians on the ship?

Stephen Dubner and Steven Levitt published the work of an anonymous economist turned bagelseller, Paul F., who dropped off baskets of bagels and came back to collect money from a cashbox, and also collected statistics on payment rates. The current average payment rate is 89%. Paul F. found that people on the executive floor of a company steal more bagels; that people with security clearances don't steal any fewer bagels; that telecom companies have robbed him and that law firms aren't worth the trouble.

Hobbes (of Calvin and Hobbes) once said: "I don't know what's worse, the fact that everyone's got a price, or the fact that their price is so low."

If Knut Haukelid sold his soul, he held out for a *damned* high price—the end of the Nazi atomic weapons program.

Others value their integrity less than a bagel.

One suspects that Haukelid's price was far higher than most people would charge, if you told them to *never* sell out. Maybe we should stop telling people they should *never* let themselves be bought, and focus on raising their price to something higher than a bagel?

But I really don't know if that's enough.

The German philosopher Fichte once said, "I would not break my word even to save humanity."

Raymond Smullyan, in whose book I read this quote, seemed to laugh and not take Fichte seriously.

Abraham Heschel said of Fichte, "His salvation and righteousness were apparently so much more important to him than the fate of all men that he would have destroyed mankind to save himself."

I don't think they get it.

If a serial killer comes to a confessional, and confesses that he's killed six people and plans to kill more, should the priest turn him in? I would answer, "No." If not for the seal of the confessional, the serial killer would never have come to the priest in the first place. All else being equal, I would prefer the world in which the serial killer talks to the priest, and the priest gets a chance to try and talk the serial killer out of it.

I use the example of a priest, rather than a psychiatrist, because a psychiatrist might be tempted to break confidentiality "just this once", and the serial killer knows that. But a Catholic priest who broke the seal of the confessional—*for any reason*—would face universal condemnation from his own church. No Catholic would be tempted to say, "Well, it's all right because it was a serial killer."

I approve of this custom and its absoluteness, and I wish we had a rationalist equivalent.

The trick would be establishing something of equivalent strength to a Catholic priest who believes God doesn't want him to break the seal, rather than the lesser strength of a psychiatrist who outsources their tape transcriptions to Pakistan. Otherwise serial killers will, quite sensibly, use the Catholic priests instead, and get less rational advice.

Suppose someone comes to a rationalist Confessor and says: "You know, tomorrow I'm planning to wipe out the human species using this neat biotech concoction I cooked up in my lab." What then? Should you break the seal of the confessional to save humanity?

It appears obvious to me that the issues here are just those of the one-shot Prisoner's Dilemma, and I do not consider it obvious that you should defect on the one-shot PD if the other player cooperates in advance on the expectation that you will cooperate as well.

There are issues with trustworthiness and how the sinner can trust the rationalist's commitment. It is not enough to be trustworthy; you must appear so. But anything that mocks the appearance of trustworthiness, while being unbound from its substance, is a poor signal; the sinner can follow that logic as well. Perhaps once neuroimaging is a bit more advanced, we could have the rationalist swear under a truthtelling machine that they would not break the seal of the confessional even to save humanity.

There's a proverb I failed to Google, which runs something like, "Once someone is known to be a liar, you might as well listen to the whistling of the wind." You wouldn't want others to expect you to lie, if you have something important to say to them; and this issue cannot be wholly decoupled from the issue of whether you actually tell the truth. If you'll lie when the fate of the world is at stake, and others can guess that fact about you, then, at the moment when the fate of the world *is* at stake, that's the moment when your words become the whistling of the wind.

I don't know if Fichte meant it that way, but his statement makes perfect sense as an ethical thesis to me. It's not that one person's personal integrity is worth more, as terminal valuta, than the entire world. Rather, **losing all your ethics is not a pure advantage.**

Being believed to tell the truth has advantages, and I don't think it's so easy to decouple that from telling the truth. Being believed to keep your word has advantages; and if you're the sort of person who would in fact break your word to save humanity, the other may guess that too. Even intrapersonal ethics can help protect you from black swans and fundamental mistakes. That logic doesn't change its structure when you double the value of the stakes, or even raise them to the level of a world. Losing your ethics is not like shrugging off some chains that were cool to look at, but were weighing you down in an athletic contest.

This I knew from the beginning: That if I had no ethics I would hold to even with the world at stake, I had no ethics at all. And I could guess how *that* would turn out.

Part of the sequence *Ethical Injunctions*

Next post: "Ethics Notes"

Previous post: "Ethical Injunctions"

Ethics Notes

Followup to: Ethical Inhibitions, Ethical Injunctions, Prices or Bindings?

(Some collected replies to comments on the above three posts.)

From Ethical Inhibitions:

> Spambot: Every major democratic political leader lies abundantly to obtain office, as it's a necessity to actually persuade the voters. So Bill Clinton, Jean Chretien, Winston Churchill should qualify for at least half of your list of villainy.

Have the ones who've lied more, done better?

In cases where the politician who told *more* lies won, has that politician gone on to rule well in an absolute sense?

Is it actually true that no one who refused to lie (and this is not the same as always telling the whole truth) could win political office?

Are the lies expected, and in that sense, less than true betrayals of someone who trusts you?

Are there understood Rules of Politics that include lies but not assassinations, which the good politicians abide by, so that they are not really violating the ethics of their tribe?

Will the world be so much worse off if sufficiently good people refuse to tell outright lies and are thereby barred from public office; or would we thereby lose a George Washington or Marcus Aurelius or two, and thereby darken history?

> Pearson: American revolutionaries as well ended human lives for the greater good

Police must sometimes kill the guilty. Soldiers must sometimes kill civilians (or if the enemy knows you're reluctant, that gives them a motive to use civilians as a shield). Spies sometimes have legitimate cause to kill people who helped them, but this has probably been done far more often than it has been *justified* by a need to end the Nazi nuclear program.

I think it's worth noting that in all such cases, you can write out something like a code of ethics and at least try to have social acceptance of it. Politicians, who lie, may prefer not to discuss the whole thing, but politicians are only a small slice of society.

Are there many who transgress *even the unwritten rules* and end up really implementing the greater good? (And no, there's no unwritten rule that says you can rob a bank to stop global warming.)

...but if you're placing yourself under unusual stress, you may need to be stricter than what society will accept from you. In fact, I think it's fair to say that the further I push any art, such as rationality or AI theory, the more I perceive that *what society will let you get away with* is tremendously too sloppy a standard.

> Yvain: There are all sorts of biases that would make us less likely to believe people who "break the rules" can ever turn out well. One is the halo effect.

> Another is availability bias—it's much easier to remember people like Mao than it is to remember the people who were quiet and responsible once their revolution was over, and no one notices the genocides that didn't happen because of some coup or assassination.
>
> When the winners do something bad, it's never interpreted as bad after the fact. Firebombing a city to end a war more quickly, taxing a populace to give health care to the less fortunate, intervening in a foreign country's affairs to stop a genocide: they're all likely to be interpreted as evidence for "the ends don't justify the means" when they fail, but glossed over or treated as common sense interventions when they work.

Both fair points. One of the difficult things in reasoning about ethics is the extent to which we can expect historical data to be distorted by moral self-deception on top of the more standard fogs of history.

> Morrison: I'm not sure you aren't "making too much stew from one oyster". I certainly feel a whole lot less ethically inhibited if I'm really, really certain I'm not going to be punished. When I override, it feels very deliberate—"system two" grappling and struggling with "system one"'s casual amorality, and with a significant chance of the override attempt failing.
>
> Weeks: This entire post is kind of surreal to me, as I'm pretty confident I've never felt the emotion described here before... I don't remember ever wanting to do something that I both felt would be wrong and wouldn't have consequences otherwise.

I don't know whether to attribute this to genetic variance, environmental variance, misunderstanding, or a small number of genuine sociopaths among Overcoming Bias readers. Maybe Weeks is referring to "not wanting" in terms of not finally deciding to do something he felt was wrong, rather than not being tempted?

From Ethical Injunctions:

> Psy-Kosh: Given the current sequence, perhaps it's time to revisit the whole Torture vs Dust Specks thing?

I can think of two positions on torture to which I am sympathetic:

Strategy 1: No legal system or society should ever *refrain from prosecuting* those who torture. Anything important enough that torture would even be on the table, like the standard nuclear bomb in New York, is important enough that everyone involved should be willing to go to prison for the crime of torture.

Strategy 2: The chance of actually encountering a "nuke in New York" situation, that can be effectively resolved by torture, is so low, and the knock-on effects of having the policy in place so awful, that a *blanket* injunction against torture makes sense.

In case 1, you would choose TORTURE over SPECKS, and then go to jail for it, even though it was the right thing to do.

In case 2, you would say "TORTURE over SPECKS is the right alternative of the two, but a human can never be in an epistemic state where you have justified belief that this is the case". Which would tie in well to the Hansonian argument that you have an

O(3^^^3) probability penalty from the unlikelihood of finding yourself in such a unique position.

So I am sympathetic to the argument that people should never torture, or that a human can't actually get into the epistemic state of a TORTURE vs. SPECKS decision.

But I can't back the position that SPECKS over TORTURE is *inherently the right thing to do,* which I did think was the issue at hand. This seems to me to mix up an epistemic precaution with morality.

There's certainly worse things than torturing one person—torturing two people, for example. But if you adopt position 2, then you would refuse to torture one person with your own hands even to save a thousand people from torture, while *simultaneously* saying that that it is better for one person to be tortured at your own hands than for a thousand people to be tortured at someone else's.

I try to use the words "morality" and "ethics" consistently as follows: The moral questions are over the territory (or, hopefully equivalently, over epistemic states of absolute certainty). The ethical questions are over epistemic states that humans are likely to be in. Moral questions are terminal. Ethical questions are instrumental.

> Hanson: The problem here of course is how selective to be about rules to let into this protected level of "rules almost no one should think themselves clever enough to know when to violate." After all, your social training may well want you to include "Never question our noble leader" in that set. Many a Christian has been told the mysteries of God are so subtle that they shouldn't think themselves clever enough to know when they've found evidence that God isn't following a grand plan to make this the best of all possible worlds.

Some of the flaws in Christian theology lie in what they think their supposed facts would imply: e.g., that because God did miracles you can know that God is good. Other problems come more from the falsity of the premises than the invalidity of the deductions. Which is to say, if God *did* exist and *were* good, then you would be justified in being cautious around stomping on parts of God's plan that didn't seem to make sense at the moment. But this epistemic state would best be arrived at via a long history of people saying, "Look how stupid God's plan is, we need to do X" and then X blowing up on them. Rather than, as is actually the case, people saying "God's plan is X" and then X blows up on them.

Or if you'd found with some historical regularity that, when you challenged the verdict of the black box, that you seemed to be right 90% of the time, but the other 10% of the time you got black-swan blowups that caused a hundred times as much damage, that would also be cause for hesitation—albeit it doesn't quite seem like grounds for suspecting a divine plan.

> Nominull: S o... do you not actually believe in your injunction to "shut up and multiply"? Because for some time now you seem to have been arguing that we should do what feels right rather than trying to figure out what is right.

Certainly I'm not saying "just do what feels right". There's no safe defense, not even ethical injunctions. There's also no safe defense, not even "shut up and multiply".

I probably should have been clearer about this before, but I was trying to discuss things in order, and didn't want to wade into ethics without specialized posts...

People often object to the sort of scenarios that illustrate "shut up and multiply" by saying, "But if the experimenter tells you X, what if they might be lying?"

Well, in a lot of real-world cases, then yes, there are various probability updates you perform based on other people being willing to make bets against you; and just because you get certain experimental instructions doesn't imply the real world is that way.

But the base case has to be *moral* comparisons between worlds, or comparisons of expected utility between *given* probability distributions. If you can't ask about the base case, then what good can you get from instrumental ethics built on top?

Let's be very clear that I *don't* think that one small act of self-deception is an *inherently morally worse event* than, say, getting a hand chopped off. I'm asking, rather, how one should best *avoid* the dismembering chainsaw; and I am arguing that in reasonable states of knowledge a human can attain, the answer is, "Don't deceive yourself, it's a black-swan bet at best." Furthermore, that in the vast majority of cases where I have seen people conclude otherwise, it has indicated messed-up reasoning more than any actual advantage.

> Vassar: For such a reason, I would be very wary of using such rules in an AGI, but of course, perhaps the actual mathematical formulation of the rule in question within the AGI would be less problematic, though a few seconds of thought doesn't give me much reason to think this.

Are we still talking about self-deception? Because I would give odds around as extreme as the odds I would give of anything, that if you tell me "the AI you built is trying to deceive itself", it indicates that some kind of really *epic* error has occurred. Controlled shutdown, immediately.

> Vassar: In a very general sense though, I see a logical problem with this whole line of thought. How can any of these injunctions survive except as self-protecting beliefs? Isn't this whole approach just the sort of "fighting bias with bias" that you and Robin usually argue against?

Maybe I'm not being clear about how this would work in an AI!

The ethical injunction isn't *self*-protecting, it's supported within the structural framework of the underlying system. You might even find ethical injunctions starting to emerge without programmer intervention, in some cases, depending on how well the AI understood its own situation.

But the kind of injunctions I have in mind wouldn't be *reflective*—they wouldn't modify the utility function, or kick in at the reflective level to ensure their own propagation. *That* sounds really scary, to me—there ought to be an injunction against it!

You might have a rule that would *controlledly shut down* the (non-mature) AI if it tried to execute a certain kind of source code change, but that wouldn't be the same as having an injunction that *exerts direct control over the source code to propagate itself.*

To the extent the injunction sticks around in the AI, it should be as the result of ordinary reasoning, *not reasoning taking the injunction into account!* That would be the *wrong* kind of circularity; you *can* unwind past ethical unjunctions!

My ethical injunctions do *not* come with an extra clause that says, "Do not reconsider this injunction, including not reconsidering this clause." That would be going *way* too far. If anything, you ought to have an injunction against that kind of circularity (since it seems like a plausible failure mode in which the system has been parasitized by its own content).

> *You should never, ever murder an innocent person who's helped you, even if it's the right thing to do*
>
> *Shut up and do the impossible!*

> Ord: As written, both these statements are conceptually confused. I understand that you didn't actually mean either of them literally, but I would advise against trading on such deep-sounding conceptual confusions.

I can't weaken them and make them come out as the right advice.

Even after "Shut up and do the impossible", there was that commenter who posted on their failed attempt at the AI-Box Experiment by saying that they thought they gave it a good try—which shows how hard it is to convey the sentiment of "Shut up and do the impossible!"

Readers can work out on their own how to distinguish the map and the territory, I hope. But if you say "Shut up and do what seems impossible!", then that, to me, sounds like dispelling part of the essential message—that what seems impossible doesn't look like it "seems impossible", it just looks impossible.

Likewise with "things you shouldn't do even if they're the right thing to do". Only the paradoxical phrasing, which is obviously not meant to be taken literally, conveys the danger and tension of ethics—the genuine opportunities you might be passing up—and for that matter, how dangerously meta the whole line of argument is.

"Don't do it, even if it *seems* right" sounds merely clever by comparison—like you're going to reliably divine the difference between what seems right and what is right, and happily ride off into the sunset.

> Crowe: This seems closely related to inside-view versus outside-view. The think-lobe of the brain comes up with a cunning plan. The plan breaks an ethical rule but calculation shows it is for the greater good. The executive-lobe of the brain then ponders the outside view. Every-one who has executed an evil cunning plan has run a calculation of the greater good and had their plan endorsed. So the calculation lack outside-view credibility.

Yes, inside view versus outside view is definitely part of this. And the planning fallacy, optimism, and overconfidence, too.

But there are also biases arguing against the same line of reasoning, as noted by Yvain: History may be written by the victors to emphasize the transgressions of the losers while overlooking the moral compromises of those who achieved "good" results, etc.

Also, some people who execute evil cunning plans may just have evil intent—possibly also with outright lies about their intentions. In which case, they really wouldn't be in the reference class of well-meaning revolutionaries, albeit you would have to worry about your comrades; the Trotsky->Lenin->Stalin slide.

> Kurz: What's to prohibit the meta-reasoning from taking place before the shutdown triggers? It would seem that either you can hard-code an ethical inhibition or you can't. Along those lines, is it fair to presume that the inhibitions are always negative, so that non-action is the safe alternative? Why not just revert to a known state?

If a self-modifying AI with the right structure will write ethical injunctions at all, it will also inspect the code to guarantee that no race condition exists with any deliberative-level supervisory systems that might have gone wrong in the condition where the code executes. Otherwise you might as well not have the code.

Inaction isn't safe but it's safer than running an AI whose moral system has gone awry.

> Finney: Which is better: conscious self-deception (assuming that's even meaningful), or unconscious?

Once you deliberately choose self-deception, you may have to protect it by adopting other Dark Side Epistemology. I would, of course, say "neither" (as otherwise I would be swapping to the Dark Side) but if you ask me which is worse—well, hell, even I'm still undoubtedly unconsciously self-deceiving, but that's not the same as going over to the Dark Side by *allowing* it!

From Prices or Bindings?:

> Psy-Kosh: Hrm. I'd think "avoid destroying the world" itself to be an ethical injunction too.

The problem is that this is phrased as an injunction over positive consequences. Deontology does better when it's closer to the action level and negative rather than positive.

Imagine trying to give this injunction to an AI. Then it would have to do *anything* that it thought would prevent the destruction of the world, without other considerations. Doesn't sound like a good idea.

> Crossman: Eliezer, can you be explicit which argument you're making? I thought you were a utilitarian, but you've been sounding a bit Kantian lately.

If all I want is money, then I will one-box on Newcomb's Problem.

I don't think that's quite the same as being a Kantian, but it does reflect the idea that similar decision algorithms in similar epistemic states will tend to produce similar outputs, and that such decision systems should not pretend to the logical impossibility of local optimization. But this is a deep subject on which I have yet to write up my full views.

> Clay: Put more seriously, I would think that being believed to put the welfare of humanity ahead of concerns about personal integrity could have significant

advantages itself.

The whole point here is that "personal integrity" doesn't have to be about being a virtuous person. It can be about trying to save the world without any concern for your own virtue. It can be the sort of thing you'd want a pure nonsentient decision agent to do, something that was purely a means and not at all an end in itself.

> Andrix: There seems to be a conflict here between not lying to yourself, and holding a traditional rule that suggests you ignore your rationality.

Your rationality is the sum of your full abilities, including your wisdom about what you refrain from doing in the presence of what seem like good reasons.

> Yvain: I am glad Stanislav Petrov, contemplating his military oath to always obey his superiors and the appropriate guidelines, never read this post.

An interesting point, for several reasons.

First, did Petrov actually swear such an oath, and would it apply in such fashion as to require him to follow the written policy rather than using his own military judgment?

Second, you might argue that Petrov's oath wasn't intended to cover circumstances involving the end of the world, and that a common-sense exemption should apply when the stakes suddenly get raised hugely beyond the intended context of the original oath. I think this fails, because Petrov was regularly in charge of a nuclear-war installation and so this was exactly the sort of event his oath would be *expected* to apply to.

Third, the Soviets arguably implemented what I called Strategy 1 above: Petrov did the right thing, and was censured for it anyway.

Fourth—maybe, on sober reflection, we wouldn't have wanted the Soviets to act differently! Yes, the written policy was stupid. And the Soviet Union was undoubtedly censuring Petrov out of bureaucratic coverup, not for reasons of principle. But do you want the Soviet Union to have a written, explicit policy that says, "Anyone can ignore orders in a nuclear war scenario if they think it's a good idea," or even an explicit policy that says "Anyone who ignores orders in a nuclear war scenario, who is later vindicated by events, will be rewarded and promoted"?

Part of the sequence *Ethical Injunctions*

(end of sequence)

Previous post: "Prices or Bindings?"