LessWrong Political Prerequisites

Politics is the Mind-Killer

People go funny in the head when talking about politics. The evolutionary reasons for this are so obvious as to be worth belaboring: In the ancestral environment, politics was a matter of life and death. And sex, and wealth, and allies, and reputation . . . When, today, you get into an argument about whether "we" ought to raise the minimum wage, you're executing adaptations for an ancestral environment where being on the wrong side of the argument could get you killed. Being on the *right* side of the argument could let *you* kill your hated rival!

If you want to make a point about science, or rationality, then my advice is to not choose a domain from *contemporary* politics if you can possibly avoid it. If your point is inherently about politics, then talk about Louis XVI during the French Revolution. Politics is an important domain to which we should individually apply our rationality—but it's a terrible domain in which to *learn* rationality, or discuss rationality, unless all the discussants are already rational.

Politics is an extension of war by other means. Arguments are soldiers. Once you know which side you're on, you must support all arguments of that side, and attack all arguments that appear to favor the enemy side; otherwise it's like stabbing your soldiers in the back—providing aid and comfort to the enemy. People who would be level-headed about evenhandedly weighing all sides of an issue in their professional life as scientists, can suddenly turn into slogan-chanting zombies when there's a Blue or Green position on an issue.

In artificial intelligence, and particularly in the domain of nonmonotonic reasoning, there's a standard problem: "All Quakers are pacifists. All Republicans are not pacifists. Nixon is a Quaker and a Republican. Is Nixon a pacifist?"

What on Earth was the point of choosing this as an example? To rouse the political emotions of the readers and distract them from the main question? To make Republicans feel unwelcome in courses on artificial intelligence and discourage them from entering the field?[1]

Why would anyone pick such a *distracting* example to illustrate nonmonotonic reasoning? Probably because the author just couldn't resist getting in a good, solid dig at those hated Greens. It feels so *good* to get in a hearty punch, y'know, it's like trying to resist a chocolate cookie.

As with chocolate cookies, not everything that feels pleasurable is good for you.

I'm not saying that I think we should be apolitical, or even that we should adopt Wikipedia's ideal of the Neutral Point of View. But try to resist getting in those good, solid digs if you can possibly avoid it. If your topic legitimately relates to attempts to ban evolution in school curricula, then go ahead and talk about it—but don't blame it explicitly on the whole Republican Party; some of your readers may be Republicans, and they may feel that the problem is a few rogues, not the entire party. As with Wikipedia's NPOV, it doesn't matter whether (you think) the Republican Party really *is* at fault. It's just better for the spiritual growth of the community to discuss the issue without invoking color politics.

---

[1] And no, I am not a Republican. Or a Democrat.

Policy Debates Should Not Appear One-Sided

Robin Hanson proposed stores where banned products could be sold.[1] There are a number of excellent arguments for such a policy—an inherent right of individual liberty, the career incentive of bureaucrats to prohibit *everything*, legislators being just as biased as individuals. But even so (I replied), *some* poor, honest, not overwhelmingly educated mother of five children is going to go into these stores and buy a “Dr. Snakeoil’s Sulfuric Acid Drink” for her arthritis and die, leaving her orphans to weep on national television.

I was just making a factual observation. Why did some people think it was an argument in favor of regulation?

On questions of simple fact (for example, whether Earthly life arose by natural selection) there’s a legitimate expectation that the argument should be a one-sided battle; the facts themselves are either one way or another, and the so-called “balance of evidence” should reflect this. Indeed, under the Bayesian definition of evidence, “strong evidence” is just that sort of evidence which we only expect to find on one side of an argument.

But there is no reason for complex actions with many consequences to exhibit this onesidedness property. Why do people seem to want their *policy* debates to be one-sided?

Politics is the mind-killer. Arguments are soldiers. Once you know which side you’re on, you must support all arguments of that side, and attack all arguments that appear to favor the enemy side; otherwise it’s like stabbing your soldiers in the back. If you abide within that pattern, policy debates will also appear one-sided to you—the costs and drawbacks of your favored policy are enemy soldiers, to be attacked by any means necessary.

One should also be aware of a related failure pattern: thinking that the course of Deep Wisdom is to compromise with perfect evenness between whichever two policy positions receive the most airtime. A policy may legitimately have *lopsided* costs or benefits. If policy questions were not tilted one way or the other, we would be unable to make decisions about them. But there is also a human tendency to deny all costs of a favored policy, or deny all benefits of a disfavored policy; and people will therefore tend to think policy tradeoffs are tilted much further than they actually are.

If you allow shops that sell otherwise banned products, some poor, honest, poorly educated mother of five kids is going to buy something that kills her. This is a prediction about a factual consequence, and as a factual question it appears rather straightforward—a sane person should readily confess this to be true regardless of which stance they take on the policy issue. You may *also* think that making things illegal just makes them more expensive, that regulators will abuse their power, or that her individual freedom trumps your desire to meddle with her life. But, as a matter of simple fact, she’s still going to die.

We live in an unfair universe. Like all primates, humans have strong negative reactions to perceived unfairness; thus we find this fact stressful. There are two popular methods of dealing with the resulting cognitive dissonance. First, one may change one's view of the facts—deny that the unfair events took place, or edit the history to make it appear fair.[2] Second, one may change one's morality—deny that the events are unfair.

Some libertarians might say that if you go into a "banned products shop," passing clear warning labels that say THINGS IN THIS STORE MAY KILL YOU, and buy something that kills you, then it's your own fault and you deserve it. If that were a moral truth, there would be *no downside* to having shops that sell banned products. It wouldn't just be a *net benefit*, it would be a *one-sided* tradeoff with no drawbacks.

Others argue that regulators can be trained to choose rationally and in harmony with consumer interests; if those were the facts of the matter then (in their moral view) there would be *no downside* to regulation.

Like it or not, there's a birth lottery for intelligence—though this is one of the cases where the universe's unfairness is so extreme that many people choose to deny the facts. The experimental evidence for a purely genetic component of 0.6–0.8 is overwhelming, but even if this were to be denied, you don't choose your parental upbringing or your early schools either.

I was raised to believe that denying reality is a *moral wrong.* If I were to engage in wishful optimism about how Sulfuric Acid Drink was likely to benefit me, I would be doing something that I was *warned* against and raised to regard as unacceptable. Some people are born into environments—we won't discuss their genes, because that part is too unfair—where the local witch doctor tells them that it is *right* to have faith and *wrong* to be skeptical. In all goodwill, they follow this advice and die. Unlike you, they weren't raised to believe that people are responsible for their individual choices to follow society's lead. Do you really think you're so smart that you would have been a proper scientific skeptic even if you'd been born in 500 CE? Yes, there is a birth lottery, no matter what you believe about genes.

Saying "People who buy dangerous products deserve to get hurt!" is not tough-minded. It is a way of refusing to live in an unfair universe. Real tough-mindedness is saying, "Yes, sulfuric acid is a horrible painful death, and no, that mother of five children didn't deserve it, but we're going to keep the shops open anyway because we did this cost-benefit calculation." Can you imagine a politician saying that? Neither can I. But insofar as economists have the power to influence policy, it might help if they could think it privately—maybe even say it in journal articles, suitably dressed up in polysyllabismic obfuscationalization so the media can't quote it.

I don't think that when someone makes a stupid choice and dies, this is a cause for celebration. I count it as a tragedy. It is not always helping people, to save them from the consequences of their own actions; but I draw a moral line at capital punishment. If you're dead, you can't learn from your mistakes.

Unfortunately the universe doesn't agree with me. We'll see which one of us is still standing when this is over.

---

[1] Robin Hanson et al., "The Hanson-Hughes Debate on 'The Crack of a Future Dawn,'" 16, no. 1 (2007): 99–126, http://jetpress.org/v16/hanson.pdf.

[2] This is mediated by the affect heuristic and the just-world fallacy.

The Scales of Justice, the Notebook of Rationality

Lady Justice is widely depicted as carrying scales. A set of scales has the property that whatever pulls one side down pushes the other side up. This makes things very convenient and easy to track. It’s also usually a gross distortion.

In human discourse there is a natural tendency to treat discussion as a form of combat, an extension of war, a sport; and in sports you only need to keep track of how many points have been scored by each team. There are only two sides, and every point scored against one side is a point in favor of the other. Everyone in the audience keeps a mental running count of how many points each speaker scores against the other. At the end of the debate, the speaker who has scored more points is, obviously, the winner; so everything that speaker says must be true, and everything the loser says must be wrong.

“The Affect Heuristic in Judgments of Risks and Benefits” studied whether subjects mixed up their judgments of the possible benefits of a technology (e.g., nuclear power), and the possible risks of that technology, into a single overall good or bad feeling about the technology.[1] Suppose that I first tell you that a particular kind of nuclear reactor generates less nuclear waste than competing reactor designs. But then I tell you that the reactor is more unstable than competing designs, with a greater danger of melting down if a sufficient number of things go wrong simultaneously.

If the reactor is more likely to melt down, this seems like a “point against” the reactor, or a “point against” someone who argues for building the reactor. And if the reactor produces less waste, this is a “point for” the reactor, or a “point for” building it. So are these two facts opposed to each other? No. In the real world, no. These two facts may be cited by different sides of the same debate, but they are logically distinct; the facts don’t know whose side they’re on.

If it’s a physical fact about a reactor design that it’s passively safe (won’t go supercritical even if the surrounding coolant systems and so on break down), this doesn’t imply that the reactor will necessarily generate less waste, or produce electricity at a lower cost. All these things would be good, but they are not the same good thing. The amount of waste produced by the reactor arises from the properties of that reactor. Other physical properties of the reactor make the nuclear reaction more unstable. Even if some of the same design properties are involved, you have to separately consider the probability of meltdown, and the expected annual waste generated. These are two different physical questions with two different factual answers.

But studies such as the above show that people tend to judge technologies—and many other problems—by an overall good or bad feeling. If you tell people a reactor design produces less waste, they rate its probability of meltdown as lower. This means getting the *wrong answer* to physical questions with definite factual answers, because you have mixed up logically distinct questions—treated facts like human soldiers on different sides of a war, thinking that any soldier on one side can be used to fight any soldier on the other side.

A set of scales is not wholly inappropriate for Lady Justice if she is investigating a strictly factual question of guilt or innocence. Either John Smith killed John Doe, or not. We are taught (by E. T. Jaynes) that all Bayesian evidence consists of probability flows *between* hypotheses; there is no such thing as evidence that "supports" or "contradicts" a single hypothesis, except insofar as other hypotheses do worse or better. So long as Lady Justice is investigating a *single*, strictly *factual* question with a *binary* answer space, a set of scales would be an appropriate tool. If Justitia must consider any more complex issue, she should relinquish her scales or relinquish her sword.

Not all arguments reduce to mere up or down. Lady Rationality carries a notebook, wherein she writes down all the facts that aren't on anyone's side.

[1] Melissa L. Finucane et al., "The Affect Heuristic in Judgments of Risks and Benefits," *Journal of Behavioral Decision Making* 13, no. 1 (2000): 1–17.

Correspondence Bias

> The correspondence bias is the tendency to draw inferences about a person's unique and enduring dispositions from behaviors that can be entirely explained by the situations in which they occur.
>
> —Gilbert and Malone[1]

We tend to see far too direct a correspondence between others' actions and personalities. When we see someone else kick a vending machine for no visible reason, we assume they are "an angry person." But when you yourself kick the vending machine, it's because the bus was late, the train was early, your report is overdue, and now the damned vending machine has eaten your lunch money for the second day in a row. *Surely*, you think to yourself, *anyone would kick the vending machine, in that situation*.

We attribute our own actions to our *situations*, seeing our behaviors as perfectly normal responses to experience. But when someone else kicks a vending machine, we don't see their past history trailing behind them in the air. We just see the kick, for no reason *we* know about, and we think this must be a naturally angry person—since they lashed out without any provocation.

Yet consider the prior probabilities. There are more late buses in the world, than mutants born with unnaturally high anger levels that cause them to sometimes spontaneously kick vending machines. Now the average human is, in fact, a mutant. If I recall correctly, an average individual has two to ten somatically expressed mutations. But any *given* DNA location is very unlikely to be affected. Similarly, any given aspect of someone's disposition is probably not very far from average. To suggest otherwise is to shoulder a burden of improbability.

Even when people are informed explicitly of situational causes, they don't seem to properly discount the observed behavior. When subjects are told that a pro-abortion or anti-abortion speaker was *randomly assigned* to give a speech on that position, subjects still think the speakers harbor leanings in the direction randomly assigned.[2]

It seems quite intuitive to explain rain by water spirits; explain fire by a fire-stuff (phlogiston) escaping from burning matter; explain the soporific effect of a medication by saying that it contains a "dormitive potency." Reality usually involves more complicated mechanisms: an evaporation and condensation cycle underlying rain, oxidizing combustion underlying fire, chemical interactions with the nervous system for soporifics. But mechanisms sound more complicated than essences; they are harder to think of, less available. So when someone kicks a vending machine, we think they have an innate vending-machine-kicking-tendency.

Unless the "someone" who kicks the machine is us—in which case we're behaving perfectly normally, given our situations; surely anyone else would do the same. Indeed, we overestimate how likely others are to respond the same way we do—the "false consensus effect." Drinking students considerably overestimate the fraction of fellow students who drink, but nondrinkers considerably underestimate the fraction. The "fundamental attribution error" refers to our tendency to overattribute others' behaviors to their dispositions, while reversing this tendency for ourselves.

To understand why people act the way they do, we must first realize that everyone sees themselves as behaving normally. Don’t ask what strange, mutant disposition they were born with, which directly corresponds to their surface behavior. Rather, ask what situations people see themselves as being in. Yes, people do have dispositions—but there are not *enough* heritable quirks of disposition to directly account for all the surface behaviors you see.

Suppose I gave you a control with two buttons, a red button and a green button. The red button destroys the world, and the green button stops the red button from being pressed. Which button would you press? The green one. Anyone who gives a different answer is probably overcomplicating the question.[3]

And yet people sometimes ask me why I want to save the world.[4] Like I must have had a traumatic childhood or something. Really, it seems like a pretty obvious decision . . . if you see the situation in those terms.

I may have non-average views which call for explanation—why do I believe such things, when most people don’t?—but given those beliefs, my *reaction* doesn’t seem to call forth an exceptional explanation. Perhaps I am a victim of false consensus; perhaps I overestimate how many people would press the green button if they saw the situation in those terms. But y’know, I’d still bet there’d be at least a *substantial minority*.

Most people see themselves as perfectly normal, from the inside. Even people you hate, people who do terrible things, are not exceptional mutants. No mutations are required, alas. When you understand this, you are ready to stop being surprised by human events.

---

[1] Daniel T. Gilbert and Patrick S. Malone, “The Correspondence Bias,” *Psychological Bulletin* 117, no. 1 (1995): 21–38.

[2] Edward E. Jones and Victor A. Harris, “The Attribution of Attitudes,” *Journal of Experimental Social Psychology* 3 (1967): 1–24, http://www.radford.edu/~jaspelme/443/spring-2007/Articles/Jones_n_Harris_1967.pdf.

[3] Compare “Transhumanism as Simplified Humanism.” http://yudkowsky.net/singularity/simplified.

[4] See Eliezer Yudkowsky, “Artificial Intelligence as a Positive and Negative Factor in Global Risk,” in *Global Catastrophic Risks*, ed. Nick Bostrom and Milan M. Ćirković (New York: Oxford University Press, 2008), 308–345.

Are Your Enemies Innately Evil?

We see far too direct a correspondence between others' actions and their inherent dispositions. We see unusual dispositions that exactly match the unusual behavior, rather than asking after real situations or imagined situations that could explain the behavior. We hypothesize mutants.

When someone actually *offends* us—commits an action of which we (rightly or wrongly) disapprove—then, I observe, the correspondence bias redoubles. There seems to be a *very* strong tendency to blame evil deeds on the Enemy's mutant, evil disposition. Not as a moral point, but as a strict question of prior probability, we should ask what the Enemy might believe about their situation that would reduce the seeming bizarrity of their behavior. This would allow us to hypothesize a less exceptional disposition, and thereby shoulder a lesser burden of improbability.

On September 11th, 2001, nineteen Muslim males hijacked four jet airliners in a deliberately suicidal effort to hurt the United States of America. Now why do you suppose they might have done that? Because they saw the USA as a beacon of freedom to the world, but were born with a mutant disposition that made them hate freedom?

Realistically, most people don't construct their life stories with themselves as the villains. Everyone is the hero of their own story. The Enemy's story, as seen by the Enemy, *is not going to make the Enemy look bad.* If you try to construe motivations that *would* make the Enemy look bad, you'll end up flat wrong about what actually goes on in the Enemy's mind.

But politics is the mind-killer. Debate is war; arguments are soldiers. If the Enemy did have an evil disposition, that would be an argument in favor of your side. And *any* argument that favors your side must be supported, no matter how silly—otherwise you're letting up the pressure somewhere on the battlefront. Everyone strives to outshine their neighbor in patriotic denunciation, and no one dares to contradict. Soon the Enemy has horns, bat wings, flaming breath, and fangs that drip corrosive venom. If you deny any aspect of this on merely factual grounds, you are arguing the Enemy's side; you are a traitor. Very few people will understand that you aren't defending the Enemy, just defending the truth.

If it took a mutant to do monstrous things, the history of the human species would look very different. Mutants would be rare.

Or maybe the fear is that understanding will lead to forgiveness. It's easier to shoot down evil mutants. It is a more inspiring battle cry to scream, "Die, vicious scum!" instead of "Die, people who could have been just like me but grew up in a different environment!" You might feel guilty killing people who *weren't* pure darkness.

This looks to me like the deep-seated yearning for a one-sided policy debate in which the best policy has *no* drawbacks. If an army is crossing the border or a lunatic is coming at you with a knife, the policy alternatives are (a) defend yourself or (b) lie down and die. If you defend yourself, you may have to kill. If you kill someone who could, in another world, have been your friend, that is a tragedy. And it *is* a tragedy. The other option, lying down and dying, is also a tragedy. Why must there be a non-tragic option? Who says that the best policy available must have no downside? If someone has to die, it may as well be the initiator of force, to discourage future violence and thereby minimize the total sum of death.

If the Enemy has an average disposition, and is acting from beliefs about their situation that would make violence a typically human response, then that doesn't mean their beliefs are factually accurate. It doesn't mean they're justified. It means you'll have to shoot down someone who is the hero of their own story, and in their novel the protagonist will die on page 80. That is a tragedy, but it is better than the alternative tragedy. It is the choice that every police officer makes, every day, to keep our neat little worlds from dissolving into chaos.

When you accurately estimate the Enemy's psychology—when you know what is really in the Enemy's mind—that knowledge won't feel like landing a delicious punch on the opposing side. It won't give you a warm feeling of righteous indignation. It won't make you feel good about yourself. If your estimate makes you feel unbearably sad, you may be seeing the world as it really is. More rarely, an accurate estimate may send shivers of serious horror down your spine, as when dealing with true psychopaths, or neurologically intact people with beliefs that have utterly destroyed their sanity (Scientologists or Jesus Campers).

So let's come right out and say it—the 9/11 hijackers weren't evil mutants. They did not hate freedom. They, too, were the heroes of their own stories, and they died for what they believed was right—truth, justice, and the Islamic way. If the hijackers saw themselves that way, it doesn't mean their beliefs were true. If the hijackers saw themselves that way, it doesn't mean that we have to agree that what they did was justified. If the hijackers saw themselves that way, it doesn't mean that the passengers of United Flight 93 should have stood aside and let it happen. It does mean that in another world, if they had been raised in a different environment, those hijackers might have been police officers. And that is indeed a tragedy. Welcome to Earth.

Reversed Stupidity Is Not Intelligence

> ". . . then our people on that time-line went to work with corrective action. Here."
>
> He wiped the screen and then began punching combinations. Page after page appeared, bearing accounts of people who had claimed to have seen the mysterious disks, and each report was more fantastic than the last.
>
> "The standard smother-out technique," Verkan Vall grinned. "I only heard a little talk about the 'flying saucers,' and all of that was in joke. In that order of culture, you can always discredit one true story by setting up ten others, palpably false, parallel to it."
>
> —H. Beam Piper, *Police Operation*

Piper had a point. Pers'nally, I don't believe there are any poorly hidden aliens infesting these parts. But my disbelief has nothing to do with the awful embarrassing irrationality of flying saucer cults—at least, I hope not.

You and I believe that flying saucer cults arose in the total absence of any flying saucers. Cults can arise around almost any idea, thanks to human silliness. This silliness operates *orthogonally* to alien intervention: We would expect to see flying saucer cults whether or not there were flying saucers. Even if there were poorly hidden aliens, it would not be any *less* likely for flying saucer cults to arise. The conditional probability P(cults|aliens) isn't less than P(cults|¬aliens), unless you suppose that poorly hidden aliens would deliberately suppress flying saucer cults.[1] By the Bayesian definition of evidence, the observation "flying saucer cults exist" is not evidence *against* the existence of flying saucers. It's not much evidence one way or the other.

This is an application of the general principle that, as Robert Pirsig puts it, "The world's greatest fool may say the Sun is shining, but that doesn't make it dark out."[2]

If you knew someone who was wrong 99.99% of the time on yes-or-no questions, you could obtain 99.99% accuracy just by reversing their answers. They would need to do all the work of obtaining good evidence entangled with reality, and processing that evidence coherently, just to *anticorrelate* that reliably. They would have to be superintelligent to be that stupid.

A car with a broken engine cannot drive backward at 200 mph, even if the engine is *really really broken.*

If stupidity does not reliably anticorrelate with truth, how much less should human evil anticorrelate with truth? The converse of the halo effect is the horns effect: All perceived negative qualities correlate. If Stalin is evil, then everything he says should be false. You wouldn't want to agree with *Stalin*, would you?

Stalin also believed that 2 + 2 = 4. Yet if you defend any statement made by Stalin, even "2 + 2 = 4," people will see only that you are "agreeing with Stalin"; you must be on his side.

Corollaries of this principle:

- To argue against an idea honestly, you should argue against the best arguments of the strongest advocates. Arguing against weaker advocates proves *nothing*, because even the strongest idea will attract weak advocates. If you want to argue against transhumanism or the intelligence explosion, you have to directly challenge the arguments of Nick Bostrom or Eliezer Yudkowsky post-2003. The least convenient path is the only valid one.[3]
- Exhibiting sad, pathetic lunatics, driven to madness by their apprehension of an Idea, is no evidence against that Idea. Many New Agers have been made crazier by their personal apprehension of quantum mechanics.
- Someone once said, "Not all conservatives are stupid, but most stupid people are conservatives." If you cannot place yourself in a state of mind where this statement, true or false, seems *completely irrelevant* as a critique of conservatism, you are not ready to think rationally about politics.
- Ad hominem argument is not valid.
- You need to be able to argue against genocide without saying "Hitler wanted to exterminate the Jews." If Hitler *hadn't* advocated genocide, would it thereby become okay?
- In Hansonian terms: Your instinctive willingness to believe something will change along with your willingness to *affiliate* with people who are known for believing it—quite apart from whether the belief is actually *true.* Some people may be reluctant to believe that God does not exist, not because there is evidence that God *does* exist, but rather because they are reluctant to affiliate with Richard Dawkins or those darned "strident" atheists who go around publicly saying "God does not exist."
- If your current computer stops working, you can't conclude that everything about the current system is wrong and that you need a new system without an AMD processor, an ATI video card, a Maxtor hard drive, or case fans—even though your current system has all these things and it doesn't work. Maybe you just need a new power cord.
- If a hundred inventors fail to build flying machines using metal and wood and canvas, it doesn't imply that what you really need is a flying machine of bone and flesh. If a thousand projects fail to build Artificial Intelligence using electricity-based computing, this doesn't mean that electricity is the source of the problem. Until you understand the problem, hopeful reversals are exceedingly unlikely to hit the solution.[4]

[1] Read "P(cults|aliens)" as "the probability of UFO cults given that aliens have visited Earth," and read "P(cults|¬aliens)" as "the probability of UFO cults given that aliens have not visited Earth."

[2] Robert M. Pirsig, *Zen and the Art of Motorcycle Maintenance: An Inquiry Into Values*, 1st ed. (New York: Morrow, 1974).

[3] See Scott Alexander, "The Least Convenient Possible World," *Less Wrong* (blog), December 2, 2018, http://lesswrong.com/lw/2k/the_least_convenient_possible_world/.

[4] See also "Selling Nonapples." http://lesswrong.com/lw/vs/selling_nonapples.

Argument Screens Off Authority

Scenario 1: Barry is a famous geologist. Charles is a fourteen-year-old juvenile delinquent with a long arrest record and occasional psychotic episodes. Barry flatly asserts to Arthur some counterintuitive statement about rocks, and Arthur judges it 90% probable. Then Charles makes an equally counterintuitive flat assertion about rocks, and Arthur judges it 10% probable. Clearly, Arthur is taking the speaker's *authority* into account in deciding whether to believe the speaker's assertions.

Scenario 2: David makes a counterintuitive statement about physics and gives Arthur a detailed explanation of the arguments, including references. Ernie makes an equally counterintuitive statement, but gives an unconvincing argument involving several leaps of faith. Both David and Ernie assert that this is the best explanation they can possibly give (to anyone, not just Arthur). Arthur assigns 90% probability to David's statement after hearing his explanation, but assigns a 10% probability to Ernie's statement.

It might seem like these two scenarios are roughly symmetrical: both involve taking into account useful evidence, whether strong versus weak authority, or strong versus weak argument.

But now suppose that Arthur asks Barry and Charles to make full technical cases, with references; and that Barry and Charles present equally good cases, and Arthur looks up the references and they check out. Then Arthur asks David and Ernie for their credentials, and it turns out that David and Ernie have roughly the same credentials—maybe they're both clowns, maybe they're both physicists.

Assuming that Arthur is knowledgeable enough to understand all the technical arguments—otherwise they're just impressive noises—it seems that Arthur should view David as having a great advantage in plausibility over Ernie, while Barry has at best a minor advantage over Charles.

Indeed, if the technical arguments are good enough, Barry's advantage over Charles may not be worth tracking. A good technical argument is one that *eliminates* reliance on the personal authority of the speaker.

Similarly, if we really believe Ernie that the argument he gave is the best argument he *could* give, which includes all of the inferential steps that Ernie executed, and all of the support that Ernie took into account—citing any authorities that Ernie may have listened to himself—then we can pretty much ignore any information about Ernie's credentials. Ernie can be a physicist or a clown, it shouldn't matter. (Again, this assumes we have enough technical ability to process the argument. Otherwise, Ernie is simply uttering mystical syllables, and whether we "believe" these syllables depends a great deal on his authority.)

So it seems there's an asymmetry between argument and authority. If we know authority we are still interested in hearing the arguments; but if we know the arguments fully, we have very little left to learn from authority.

Clearly (says the novice) authority and argument are fundamentally different kinds of evidence, a difference unaccountable in the boringly clean methods of Bayesian probability theory.[1] For while the strength of the evidences—90% versus 10%—is just the same in both cases, they do not behave similarly when combined. How will we account for this?

Here’s half a technical demonstration of how to represent this difference in probability theory. (The rest you can take on my personal authority, or look up in the references.)

If P(H|E1) = 90% and P(H|E2) = 9%, what is the probability P(H|E1,E2)? If learning E1 is true leads us to assign 90% probability to H, and learning E2 is true leads us to assign 9% probability to H, then what probability should we assign to H if we learn both E1 and E2? This is simply not something you can calculate in probability theory from the information given. No, the missing information is not the prior probability of H. The events E1 and E2 may not be independent of each other.

Suppose that H is “My sidewalk is slippery,” E1 is “My sprinkler is running,” and E2 is “It’s night.” The sidewalk is slippery starting from one minute after the sprinkler starts, until just after the sprinkler finishes, and the sprinkler runs for ten minutes. So if we know the sprinkler is on, the probability is 90% that the sidewalk is slippery. The sprinkler is on during 10% of the nighttime, so if we know that it’s night, the probability of the sidewalk being slippery is 9%. If we know that it’s night and the sprinkler is on—that is, if we know both facts—the probability of the sidewalk being slippery is 90%.

We can represent this in a graphical model as follows:

Whether or not it’s Night *causes* the Sprinkler to be on or off, and whether the Sprinkler is on *causes* the sidewalk to be Slippery or unSlippery.

The direction of the arrows is meaningful. Say we had:

This would mean that, if I *didn’t* know anything about the sprinkler, the probability of Nighttime and Slipperiness would be independent of each other. For example, suppose that I roll Die One and Die Two, and add up the showing numbers to get the Sum:

If you don’t tell me the sum of the two numbers, and you tell me the first die showed 6, this doesn’t tell me anything about the result of the second die, yet. But if you now also tell me the sum is 7, I know the second die showed 1.

Figuring out when various pieces of information are dependent or independent of each other, given various background knowledge, actually turns into a quite technical topic. The books to read are Judea Pearl’s *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference* and *Causality: Models, Reasoning, and Inference*. (If you only have time to read one book, read the first one.)

If you know how to read causal graphs, then you look at the dice-roll graph and immediately see:

$P(\mathit{Die\ 1},\mathit{Die\ 2}) = P(\mathit{Die\ 1}) \times P(\mathit{Die\ 2})$

$P(\mathit{Die\ 1},\mathit{Die\ 2}|\mathit{Sum}) \neq P(\mathit{Die\ 1})|\mathit{Sum}) \times P(\mathit{Die\ 2}|\mathit{Sum})$.

If you look at the correct sidewalk diagram, you see facts like:

$P(Slippery|Night) \neq P(Slippery)$

$P(Slippery|Sprinkler) \neq P(Slippery)$

$P(Slippery|Night,Sprinkler) = P(Slippery|Sprinkler)$.

That is, the probability of the sidewalk being Slippery, given knowledge about the Sprinkler and the Night, is the same probability we would assign if we knew only about the Sprinkler. Knowledge of the Sprinkler has made knowledge of the Night irrelevant to inferences about Slipperiness.

This is known as *screening off*, and the criterion that lets us read such conditional independences off causal graphs is known as *D-separation*.

For the case of argument and authority, the causal diagram looks like this:

If something is true, then it therefore tends to have arguments in favor of it, and the experts therefore observe these evidences and change their opinions. (In theory!)

If we see that an expert believes something, we infer back to the existence of evidence-in-the-abstract (even though we don’t know what that evidence is exactly), and from the existence of this abstract evidence, we infer back to the truth of the proposition.

But if we know the value of the Argument node, this D-separates the node “Truth” from the node “Expert Belief” by blocking all paths between them, according to certain technical criteria for “path blocking” that seem pretty obvious in this case. So even without checking the exact probability distribution, we can read off from the graph that:

$P(truth|argument,expert) = P(truth|argument)$.

This does not represent a contradiction of ordinary probability theory. It’s just a more compact way of expressing certain probabilistic facts. You could read the same equalities and inequalities off an unadorned probability distribution—but it would be harder to see it by eyeballing. Authority and argument don’t need two different kinds of probability, any more than sprinklers are made out of ontologically different stuff than sunlight.

In practice you can never *completely* eliminate reliance on authority. Good authorities are more likely to know about any counterevidence that exists and should be taken into account; a lesser authority is less likely to know this, which makes their arguments less reliable. This is not a factor you can eliminate merely by hearing the evidence they *did* take into account.

It’s also very hard to reduce arguments to *pure* math; and otherwise, judging the strength of an inferential step may rely on intuitions you can’t duplicate without the same thirty years of experience.

There is an ineradicable legitimacy to assigning *slightly* higher probability to what E. T. Jaynes tells you about Bayesian probability, than you assign to Eliezer Yudkowsky making the exact same statement. Fifty additional years of experience should not count for literally *zero* influence.

But this slight strength of authority is only *ceteris paribus*, and can easily be overwhelmed by stronger arguments. I have a minor erratum in one of Jaynes's books—because algebra trumps authority.

[1] See "What Is Evidence?" in *Map and Territory*.

Hug the Query

In the art of rationality there is a discipline of *closeness-to-the-issue*—trying to observe evidence that is as near to the original question as possible, so that it screens off as many other arguments as possible.

The Wright Brothers say, “My plane will fly.” If you look at their authority (bicycle mechanics who happen to be excellent amateur physicists) then you will compare their authority to, say, Lord Kelvin, and you will find that Lord Kelvin is the greater authority.

If you demand to see the Wright Brothers’ calculations, and you can follow them, and you demand to see Lord Kelvin’s calculations (he probably doesn’t have any apart from his own incredulity), then authority becomes much less relevant.

If you actually *watch the plane fly*, the calculations themselves become moot for many purposes, and Kelvin’s authority not even worth considering.

The more *directly* your arguments bear on a question, without intermediate inferences—the closer the observed nodes are to the queried node, in the Great Web of Causality—the more powerful the evidence. It’s a theorem of these causal graphs that you can never get *more* information from distant nodes, than from strictly closer nodes that screen off the distant ones.

Jerry Cleaver said: “What does you in is not failure to apply some high-level, intricate, complicated technique. It’s overlooking the basics. Not keeping your eye on the ball.”[1]

Just as it is superior to argue physics than credentials, it is also superior to argue physics than rationality. Who was more rational, the Wright Brothers or Lord Kelvin? If we can check their calculations, we don’t have to care! The virtue of a rationalist cannot *directly* cause a plane to fly.

If you forget this principle, learning about more biases will hurt you, because it will distract you from more direct arguments. It’s all too easy to argue that someone is exhibiting Bias #182 in your repertoire of fully generic accusations, but you can’t *settle* a factual issue without closer evidence. If there are biased reasons to say the Sun is shining, that doesn’t make it dark out.

Just as you can’t always experiment today, you can’t always check the calculations today.[2] Sometimes you don’t know enough background material, sometimes there’s private information, sometimes there just isn’t time. There’s a sadly large number of times when it’s worthwhile to judge the speaker’s rationality. You should always do it with a hollow feeling in your heart, though, a sense that something’s missing.

Whenever you can, dance as near to the original question as possible—press yourself up against it—get close enough to *hug the query!*

[1] Jerry Cleaver, *Immediate Fiction: A Complete Writing Course* (Macmillan, 2004).

[2] See also “Is Molecular Nanotechnology ’Scientific’?” http://lesswrong.com/lw/io/is_molecular_nanotechnology_scientific.

Rationality and the English Language

The other day, someone commented that my writing reminded them of George Orwell's "Politics and the English Language."[1] I was honored. Especially since I'd already thought of today's topic.

If you *really* want an artist's perspective on rationality, then read Orwell; he is mandatory reading for rationalists as well as authors. Orwell was not a scientist, but a writer; his tools were not numbers, but words; his adversary was not Nature, but human evil. If you wish to imprison people for years without trial, you must think of some other way to say it than "I'm going to imprison Mr. Jennings for years without trial." You must muddy the listener's thinking, prevent clear images from outraging conscience. You say, "Unreliable elements were subjected to an alternative justice process."

Orwell was the outraged opponent of totalitarianism and the muddy thinking in which evil cloaks itself—which is how Orwell's writings on language ended up as classic rationalist documents on a level with Feynman, Sagan, or Dawkins.

"Writers are told to avoid usage of the passive voice." A rationalist whose background comes *exclusively* from science may fail to see the flaw in the previous sentence; but anyone who's done a little writing should see it right away. I wrote the sentence in the passive voice, without telling you *who* tells authors to avoid passive voice. Passive voice removes the actor, leaving only the acted-upon. "Unreliable elements were subjected to an alternative justice process"—subjected by *whom*? What does an "alternative justice process" *do*? With enough static noun phrases, you can keep anything unpleasant from actually *happening*.

Journal articles are often written in passive voice. (Pardon me, *some scientists* write their journal articles in passive voice. It's not as if the articles are being written by no one, with no one to blame.) It sounds more authoritative to say "The subjects were administered Progenitorivox" than "I gave each college student a bottle of 20 Progenitorivox, and told them to take one every night until they were gone." If you remove the scientist from the description, that leaves only the all-important data. But in reality the scientist *is* there, and the subjects *are* college students, and the Progenitorivox wasn't "administered" but handed over with instructions. Passive voice obscures reality.

Judging from the comments I get, someone will protest that using the passive voice in a journal article is hardly a sin—after all, if you *think* about it, you can realize the scientist is there. It doesn't seem like a logical flaw. And this is why rationalists need to read Orwell, not just Feynman or even Jaynes.

Nonfiction conveys *knowledge*, fiction conveys *experience.* Medical science can extrapolate what would happen to a human unprotected in a vacuum. Fiction can make you live through it.

Some rationalists will try to analyze a misleading phrase, try to see if there *might possibly* be anything meaningful to it, try to *construct* a logical interpretation. They will be charitable, give the author the benefit of the doubt. Authors, on the other hand, are trained *not* to give themselves the benefit of the doubt. Whatever the audience *thinks* you said *is* what you said, whether you meant to say it or not; you can't argue with the audience no matter how clever your justifications.

A writer knows that readers will *not* stop for a minute to think. A fictional experience is a continuous stream of first impressions. A writer-rationalist pays attention to the *experience* words create. If you are evaluating the public rationality of a statement, and you analyze the words deliberatively, rephrasing propositions, trying out different meanings, searching for nuggets of truthiness, then you're losing track of the first impression—what the audience *sees*, or rather *feels.*

A novelist would notice the screaming wrongness of "The subjects were administered Progenitorivox." What life is here for a reader to live? This sentence creates a distant feeling of authoritativeness, and that's *all*—the *only* experience is the feeling of being told something reliable. A novelist would see nouns too abstract to show what actually happened—the postdoc with the bottle in their hand, trying to look stern; the student listening with a nervous grin.

My point is not to say that journal articles should be written like novels, but that a rationalist should become consciously aware of the *experiences* which words create. A rationalist must understand the mind and how to operate it. That includes the stream of consciousness, the part of yourself that unfolds in language. A rationalist must become consciously aware of the actual, experiential impact of phrases, beyond their mere propositional semantics.[2]

Or to say it more bluntly: *Meaning does not excuse impact!*

I don't care what rational interpretation you can *construct* for an applause light like "AI should be developed through democratic processes." That cannot excuse its irrational impact of signaling the audience to applaud, not to mention its cloudy question-begging vagueness.

Here is Orwell, railing against the *impact* of cliches, their effect on the experience of thinking:

> When one watches some tired hack on the platform mechanically repeating the familiar phrases—BESTIAL, ATROCITIES, IRON HEEL, BLOODSTAINED TYRANNY, FREE PEOPLES OF THE WORLD, STAND SHOULDER TO SHOULDER—one often has a curious feeling that one is not watching a live human being but some kind of dummy . . . A speaker who uses that kind of phraseology has gone some distance toward turning himself into a machine. The appropriate noises are coming out of his larynx, but his brain is not involved, as it would be if he were choosing his words for himself . . .
>
> What is above all needed is to let the meaning choose the word, and not the other way around. In prose, the worst thing one can do with words is surrender to them. When you think of a concrete object, you think wordlessly, and then, if you want to describe the thing you have been visualising you probably hunt about until you find the exact words that seem to fit it. When you think of something abstract you are more inclined to use words from the start, and unless you make a conscious effort to prevent it, the existing dialect will come rushing in and do the job for you, at the expense of blurring or even changing your meaning. Probably it is better to put off using words as long as possible and get one's meaning as clear as one can through pictures and sensations.

Charles Sanders Peirce might have written that last paragraph. More than one path can lead to the Way.

[1] Comment at http://lesswrong.com/lw/jb/applause_lights/f1t.

[2] Compare “Semantic Stopsigns” and “Applause Lights” in *Map and Territory*.

Human Evil and Muddled Thinking

George Orwell saw the descent of the civilized world into totalitarianism, the conversion or corruption of one country after another; the boot stamping on a human face, forever, and remember that it is forever. You were born too late to remember a time when the rise of totalitarianism seemed unstoppable, when one country after another fell to secret police and the thunderous knock at midnight, while the professors of free universities hailed the Soviet Union's purges as progress. It feels as alien to you as fiction; it is hard for you to take seriously. Because, in your branch of time, the Berlin Wall fell. And if Orwell's name is not carved into one of those stones, it should be.

Orwell saw the destiny of the human species, and he put forth a convulsive effort to wrench it off its path. Orwell's weapon was clear writing. Orwell knew that muddled language is muddled thinking; he knew that human evil and muddled thinking intertwine like conjugate strands of DNA:[1]

> In our time, political speech and writing are largely the defence of the indefensible. Things like the continuance of British rule in India, the Russian purges and deportations, the dropping of the atom bombs on Japan, can indeed be defended, but only by arguments which are too brutal for most people to face, and which do not square with the professed aims of the political parties. Thus political language has to consist largely of euphemism, question-begging and sheer cloudy vagueness. Defenceless villages are bombarded from the air, the inhabitants driven out into the countryside, the cattle machine-gunned, the huts set on fire with incendiary bullets: this is called PACIFICATION . . .

Orwell was clear on the goal of his clarity:

> If you simplify your English, you are freed from the worst follies of orthodoxy. You cannot speak any of the necessary dialects, and when you make a stupid remark its stupidity will be obvious, even to yourself.

To make our stupidity obvious, even to ourselves—this is the heart of *Overcoming Bias*.

Evil sneaks, hidden, through the unlit shadows of the mind. We look back with the clarity of history, and weep to remember the planned famines of Stalin and Mao, which killed tens of millions. We call this evil, because it was done by deliberate human intent to inflict pain and death upon innocent human beings. We call this evil, because of the revulsion that we feel against it, looking back with the clarity of history. For perpetrators of evil to avoid its natural opposition, the revulsion must remain latent. Clarity must be avoided at any cost. Even as humans of clear sight tend to oppose the evil that they see; so too does human evil, wherever it exists, set out to muddle thinking.

1984 sets this forth starkly: Orwell's ultimate villains are cutters and airbrushers of photographs (based on historical cutting and airbrushing in the Soviet Union). At the peak of all darkness in the Ministry of Love, O'Brien tortures Winston to admit that two plus two equals five:[2]

> “Do you remember,” he went on, “writing in your diary, ‘Freedom is the freedom to say that two plus two make four’?”
>
> “Yes,” said Winston.
>
> O’Brien held up his left hand, its back towards Winston, with the thumb hidden and the four fingers extended.
>
> “How many fingers am I holding up, Winston?”
>
> “Four.”
>
> “And if the party says that it is not four but five—then how many?”
>
> “Four.”
>
> The word ended in a gasp of pain. The needle of the dial had shot up to fifty-five. The sweat had sprung out all over Winston’s body. The air tore into his lungs and issued again in deep groans which even by clenching his teeth he could not stop. O’Brien watched him, the four fingers still extended. He drew back the lever. This time the pain was only slightly eased.

I am continually aghast at apparently intelligent folks—such as Robin Hanson’s colleague Tyler Cowen—who don’t think that overcoming bias is important.[3] This is your *mind* we’re talking about. Your human intelligence. It separates you from an orangutan. It built this world. You don’t think how the mind works is important? You don’t think the mind’s systematic malfunctions are important? Do you think the Inquisition would have tortured witches, if all were ideal Bayesians?

Tyler Cowen apparently feels that overcoming bias is just as biased as bias: “I view Robin’s blog as exemplifying bias, and indeed showing that bias can be very useful.” I *hope* this is only the result of thinking too abstractly while trying to sound clever. Does Tyler seriously think that scope insensitivity to the value of human life is on the same level with trying to create plans that will *really* save as many lives as possible?

Orwell was forced to fight a similar attitude—that to admit to any distinction is youthful naiveté:

> Stuart Chase and others have come near to claiming that all abstract words are meaningless, and have used this as a pretext for advocating a kind of political quietism. Since you don’t know what Fascism is, how can you struggle against Fascism?

Maybe overcoming bias doesn’t look quite exciting enough, if it’s framed as a struggle against mere accidental mistakes. Maybe it’s harder to get excited if there isn’t some clear evil to oppose. So let us be absolutely clear that where there is human evil in the world, where there is cruelty and torture and deliberate murder, there are biases enshrouding it. Where people of clear sight oppose these biases, the concealed evil fights back. The truth *does* have enemies. If *Overcoming Bias* were a newsletter in the old Soviet Union, every poster and commenter of *Overcoming Bias* would have been shipped off to labor camps.

In all human history, every great leap forward has been driven by a new clarity of thought. Except for a few natural catastrophes, every great woe has been driven by a stupidity. Our last enemy is ourselves; and this is a war, and we are soldiers.

[1] George Orwell, “Politics and the English Language,” *Horizon*, 1946.

[2] George Orwell, *1984* (Signet Classic, 1950).

[3] See Tyler Cowen, “How Important is Overcoming Bias?,” *Marginal Revolution* (blog), 2007, http://marginalrevolution.com/marginalrevolution/2007/08/how-important-i.html.

Politics is hard mode

Summary: I don't think 'politics is the mind-killer' works well rhetorically. I suggest 'politics is hard mode' instead.

Some people in and catawampus to the LessWrong community have objected to "*politics is the mind-killer*" as a framing (/ slogan / taunt). Miri Mogilevsky explained on Facebook:

> My usual first objection is that it seems odd to single politics out as a “mind-killer” when there’s plenty of evidence that tribalism happens *everywhere*. Recently, there has been a whole kerfuffle within the field of psychology about replication of studies. Of course, some key studies have failed to replicate, leading to accusations of “bullying” and “witch-hunts” and what have you. Some of the people involved have since walked their language back, but it was still a rather concerning demonstration of mind-killing in action. People took “sides,” people became upset at people based on their “sides” rather than their actual opinions or behavior, and so on.
>
> Unless this article refers specifically to electoral politics and Democrats and Republicans and things (not clear from the wording), “politics” is such a frightfully broad category of human experience that writing it off entirely as a mind-killer that cannot be discussed or else all rationality flies out the window effectively prohibits a large number of important issues from being discussed, by the very people who can, in theory, be counted upon to discuss them better than most. Is it “politics” for me to talk about my experience as a woman in gatherings that are predominantly composed of men? Many would say it is. But I’m sure that these groups of men stand to gain from hearing about my experiences, since some of them are concerned that so few women attend their events.
>
> In this article, Eliezer notes, “Politics is an important domain to which we should individually apply our rationality — but it’s a terrible domain in which to learn rationality, or discuss rationality, unless all the discussants are already rational.” But that means that we all have to individually, privately apply rationality to politics without consulting anyone who can help us do this well. After all, there is no such thing as a discussant who is “rational”; there is a reason the website is called “Less Wrong” rather than “Not At All Wrong” or “Always 100% Right.” Assuming that we are all trying to be more rational, there is nobody better to discuss politics with than each other.
>
> The rest of my objection to this meme has little to do with this article, which I think raises lots of great points, and more to do with the response that I’ve seen to it — an eye-rolling, condescending dismissal of politics itself and of anyone who cares about it. Of course, I’m totally fine if a given person isn’t interested in politics and doesn’t want to discuss it, but then they should say, “I’m not interested in this and would rather not discuss it,” or “I don’t think I can be rational in this discussion so I’d rather avoid it,” rather than sneeringly reminding me “You know, politics is the mind-killer,” as though I am an errant child. I’m well-aware of the dangers of politics to good thinking. I am also aware of the benefits

> of good thinking to politics. So I’ve decided to accept the risk and to try to apply good thinking there. [...]
>
> I’m sure there are also people who disagree with the article itself, but I don’t think I know those people personally. And to add a political dimension (heh), it’s relevant that most non-LW people (like me) initially encounter “politics is the mind-killer” being thrown out in comment threads, not through reading the original article. My opinion of the concept improved a lot once I read the article.

In the same thread, Andrew Mahone added, “Using it in that sneering way, Miri, seems just like a faux-rationalist version of ‘Oh, I don’t bother with *politics*.’ It’s just another way of looking down on any concerns larger than oneself as somehow dirty, only now, you know, *rationalist* dirty.” To which Miri replied: “Yeah, and what’s weird is that that *really* doesn’t seem to be Eliezer’s intent, judging by the eponymous article.”

Eliezer replied briefly, to clarify that he wasn't generally thinking of problems that can be directly addressed in local groups (but happen to be politically charged) as "politics":

> Hanson’s “Tug the Rope Sideways” principle, combined with the fact that large communities are hard to personally influence, explains a lot in practice about what I find suspicious about someone who claims that conventional national politics are the top priority to discuss. Obviously local community matters are exempt from that critique! I think if I’d substituted ‘national politics as seen on TV’ in a lot of the cases where I said ‘politics’ it would have more precisely conveyed what I was trying to say.

But that doesn't resolve the issue. Even if local politics is more instrumentally tractable, the worry about polarization and factionalization can still apply, and may still make it a poor epistemic training ground.

A subtler problem with banning “political” discussions on a blog or at a meet-up is that it’s hard to do fairly, because our snap judgments about what counts as “political” may themselves be affected by partisan divides. In many cases the status quo is thought of as apolitical, even though objections to the status quo are ‘political.’ (Shades of Pretending to be Wise.)

Because politics gets *personal* fast, it’s hard to talk about it successfully. But if you’re trying to build a community, build friendships, or build a movement, you can’t outlaw everything ‘personal.’

And *selectively* outlawing personal stuff gets even messier. Last year, daenerys shared anonymized stories from women, including several that discussed past experiences where the writer had been attacked or made to feel unsafe. If those discussions are made off-limits because they relate to gender and are therefore ‘political,’ some folks may take away the message that they aren’t allowed to talk about, e.g., some harmful or alienating norm they see at meet-ups. I haven’t seen enough discussions of this failure mode to feel super confident people know how to avoid it.

Since this is one of the LessWrong memes that’s most likely to pop up in cross-subcultural dialogues (along with the even more ripe-for-misinterpretation “policy debates should not appear one-sided“…), as a first (very small) step, my action proposal is to obsolete the ‘mind-killer’ framing. A better phrase for getting the same work done would be ‘**politics is hard mode**’:

1. ‘Politics is hard mode’ emphasizes that ‘mind-killing’ (= epistemic difficulty) is quantitative, not qualitative. Some things might instead fall under Middlingly Hard Mode, or under Nightmare Mode…

2. ‘Hard’ invites the question ‘hard for whom?’, more so than ‘mind-killer’ does. We’re used to the fact that some people and some contexts change what’s ‘hard’, so it’s a little less likely we’ll universally generalize.

3. ‘Mindkill’ connotes contamination, sickness, failure, weakness. In contrast, ‘Hard Mode’ doesn’t imply that a thing is low-status or unworthy. As a result, it’s less likely to create the impression (or reality) that LessWrongers or Effective Altruists dismiss out-of-hand the idea of hypothetical-political-intervention-that-isn’t-a-terrible-idea. Maybe some people do want to argue for the thesis that politics is always useless or icky, but if so it should be done in those terms, explicitly — not snuck in as a connotation.

4. ‘Hard Mode’ can’t readily be perceived as a personal attack. If you accuse someone of being ‘mindkilled’, with no context provided, that smacks of insult — you appear to be calling them stupid, irrational, deluded, or the like. If you tell someone they’re playing on ‘Hard Mode,’ that’s very nearly a compliment, which makes your advice that they change behaviors a lot likelier to go over well.

5. ‘Hard Mode’ doesn’t risk bringing to mind (e.g., gendered) stereotypes about communities of political activists being dumb, irrational, or overemotional.

6. ‘Hard Mode’ encourages a growth mindset. Maybe some topics are too hard to ever be discussed. Even so, ranking topics by difficulty encourages an approach where you try to do better, rather than *merely* withdrawing. It may be wise to eschew politics, but we should not *fear* it. (Fear is the mind-killer.)

7. **Edit:** One of the larger engines of conflict is that people are so much worse at noticing their own faults and biases than noticing others'. People will be relatively quick to dismiss others as 'mindkilled,' while frequently flinching away from or just-not-thinking 'maybe I'm a bit mindkilled about this.' Framing the problem as a challenge rather than as a failing might make it easier to be reflective and even-handed.

This is not an attempt to get more people to talk about politics. I think this is a better framing *whether or not* you trust others (or yourself) to have productive political conversations.

When I playtested this post, Ciphergoth raised the worry that 'hard mode' isn't scary-sounding enough. As dire warnings go, it's light-hearted—exciting, even. To which I say: *good*. Counter-intuitive fears should usually be *argued into* people (e.g., via Eliezer's politics sequence), not connotation-ninja'd or chanted at them. The cognitive content is more clearly conveyed by 'hard mode,' and if some group (people who love politics) stands to gain the most from internalizing this message, the message shouldn't cast that very group (people who love politics) in an obviously unflattering light. LW seems fairly memetically stable, so the main issue is what would make this meme infect friends and acquaintances who haven't read the sequences. (Or *Dune*.)

If you just want a scary personal mantra to remind yourself of the risks, I propose 'politics is SPIDERS'. Though 'politics is the mind-killer' is fine there too.

If you and your co-conversationalists haven’t yet built up a lot of trust and rapport, or if tempers are already flaring, conveying the message ‘I’m too rational to discuss politics’ or ‘You’re too irrational to discuss politics’ can make things worse. In that context, ‘politics is the mind-killer’ is the mind-killer. At least, it’s a needlessly mind-killing way of warning people about epistemic hazards.

‘Hard Mode’ lets you speak as the Humble Aspirant rather than the Aloof Superior. Strive to convey: ‘I’m worried I’m too low-level to participate in this discussion; could you have it somewhere else?’ Or: ‘Could we talk about something closer to Easy Mode, so we can level up together?’ More generally: If you’re worried that what you talk *about* will impact group epistemology, you should be even more worried about *how you talk about it*.

I Can Tolerate Anything Except The Outgroup

[Content warning: Politics, religion, social justice, spoilers for "The Secret of Father Brown". This isn't especially original to me and I don't claim anything more than to be explaining and rewording things I have heard from a bunch of other people. Unapologetically America-centric because I'm not informed enough to make it otherwise. Try to keep this off Reddit and other similar sorts of things.]

I.

In Chesterton's *The Secret of Father Brown*

, a beloved nobleman who murdered his good-for-nothing brother in a duel thirty years ago returns to his hometown wracked by guilt. All the townspeople want to forgive him immediately, and they mock the titular priest for only being willing to give a measured forgiveness conditional on penance and self-reflection. They lecture the priest on the virtues of charity and compassion.

Later, it comes out that the beloved nobleman did *not* in fact kill his good-for-nothing brother. The good-for-nothing brother killed the beloved nobleman (and stole his identity). *Now* the townspeople want to see him lynched or burned alive, and it is only the priest who – consistently – offers a measured forgiveness conditional on penance and self-reflection.

The priest tells them:

> It seems to me that you only pardon the sins that you don't really think sinful. You only forgive criminals when they commit what you don't regard as crimes, but rather as conventions. You forgive a conventional duel just as you forgive a conventional divorce. You forgive because there isn't anything to be forgiven.

He further notes that this is why the townspeople can self-righteously consider themselves more compassionate and forgiving than he is. Actual forgiveness, the kind the priest needs to cultivate to forgive evildoers, is really really hard. The fake forgiveness the townspeople use to forgive the people they like is really easy, so they get to boast not only of their forgiving nature, but of how much nicer they are than those mean old priests who find forgiveness difficult and want penance along with it.

After some thought I agree with Chesterton's point. There are a lot of people who say "I forgive you" when they mean "No harm done", and a lot of people who say "That was unforgiveable" when they mean "That was genuinely really bad". Whether or not forgiveness is *right* is a complicated topic I do not want to get in here. But since forgiveness is generally considered a virtue, and one that many want credit for having, I think it's fair to say you only earn the right to call yourself 'forgiving' if you forgive things that genuinely hurt you.

To borrow Chesterton's example, if you think divorce is a-ok, then you don't get to "forgive" people their divorces, you merely ignore them. Someone who thinks divorce is abhorrent can “forgive” divorce. *You* can forgive theft, or murder, or tax evasion, or something *you* find abhorrent.

I mean, from a utilitarian point of view, you are still doing the correct action of not giving people grief because they’re a divorcee. You can have all the Utility Points you want. All I’m saying is that if you “forgive” something you don’t care about, you don’t earn any Virtue Points.

(by way of illustration: a billionaire who gives $100 to charity gets as many Utility Points as an impoverished pensioner who donates the same amount, but the latter gets a lot more Virtue Points)

Tolerance is also considered a virtue, but it suffers the same sort of dimished expectations forgiveness does.

The Emperor summons before him Bodhidharma and asks: “Master, I have been tolerant of innumerable gays, lesbians, bisexuals, asexuals, blacks, Hispanics, Asians, transgender people, and Jews. How many Virtue Points have I earned for my meritorious deeds?”

Bodhidharma answers: “None at all”.

The Emperor, somewhat put out, demands to know why.

Bodhidharma asks: “Well, what do you think of gay people?”

The Emperor answers: “What do you think I am, some kind of homophobic bigot? Of course I have nothing against gay people!”

And Bodhidharma answers: “Thus do you gain no merit by tolerating them!”

II.

If I had to define “tolerance” it would be something like “respect and kindness toward members of an outgroup”.

And today we have an almost unprecedented situation.

We have a lot of people – like the Emperor – boasting of being able to tolerate everyone from every outgroup they can imagine, loving the outgroup, writing long paeans to how great the outgroup is, staying up at night fretting that somebody else might not like the outgroup enough.

This is really surprising. It’s a total reversal of everything we know about human psychology up to this point. No one did any genetic engineering. No one passed out weird glowing pills in the public schools. And yet suddenly we get an entire group of people who conspicuously promote and defend their outgroups, the outer the better.

What is going on here?

Let’s start by asking what exactly an outgroup is.

There’s a very boring sense in which, assuming the Emperor’s straight, gays are part of his “outgroup” ie a group that he is not a member of. But if the Emperor has curly hair, are straight-haired people part of his outgroup? If the Emperor’s name starts with the letter ‘A’, are people whose names start with the letter ‘B’ part of his outgroup?

Nah. I would differentiate between multiple different meanings of outgroup, where one is “a group you are not a part of” and the other is…something stronger.

I want to avoid a very easy trap, which is saying that outgroups are about how different you are, or how hostile you are. I don’t think that’s quite right.

Compare the Nazis to the German Jews and to the Japanese. The Nazis were very similar to the German Jews: they looked the same, spoke the same language, came from a similar culture. The Nazis were totally different from the Japanese: different race, different language, vast cultural gap. But the Nazis and Japanese mostly got along pretty well. Heck, the Nazis were actually moderately positively disposed to the *Chinese*, even when they were technically at war. Meanwhile, the conflict between the Nazis and the German Jews – some of whom didn’t even realize they were anything other than German until they checked their grandparents’ birth certificate – is the stuff of history and nightmares. Any theory of outgroupishness that naively assumes the Nazis’ natural outgroup is Japanese or Chinese people will be totally inadequate.

And this isn’t a weird exception. Freud spoke of the narcissism of small differences, saying that “it is precisely communities with adjoining territories, and related to each other in other ways as well, who are engaged in constant feuds and ridiculing each other”. Nazis and German Jews. Northern Irish Protestants and Northern Irish Catholics. Hutus and Tutsis. South African whites and South African blacks. Israeli Jews and Israeli Arabs. Anyone in the former Yugoslavia and anyone else in the former Yugoslavia.

So what makes an outgroup? Proximity plus small differences. If you want to know who someone in former Yugoslavia hates, don’t look at the Indonesians or the Zulus or the Tibetans or anyone else distant and exotic. Find the Yugoslavian ethnicity that lives closely intermingled with them and is most conspicuously similar to them, and chances are you’ll find the one who they have eight hundred years of seething hatred toward.

What makes an unexpected in-group? The answer with Germans and Japanese is obvious – a strategic alliance. In fact, the World Wars forged a lot of unexpected temporary pseudo-friendships. A recent article from War Nerd points out that the British, after spending centuries subjugating and despising the Irish and Sikhs, suddenly needed Irish and Sikh soldiers for World Wars I and II respectively. “Crush them beneath our boots” quickly changed to fawning songs about how “there never was a coward where the shamrock grows” and endless paeans to Sikh military prowess.

Sure, scratch the paeans even a little bit and you find condescension as strong as ever. But eight hundred years of the British committing genocide against the Irish and considering them literally subhuman turned into smiles and songs about shamrocks once the Irish started looking like useful cannon fodder for a larger fight. And the Sikhs, dark-skinned people with turbans and beards who pretty much exemplify the European stereotype of “scary foreigner”, were lauded by everyone from the news media all the way up to Winston Churchill.

In other words, outgroups may be the people who look exactly like you, and scary foreigner types can become the in-group on a moment’s notice when it seems convenient.

III.

There are certain theories of dark matter where it barely interacts with the regular world *at all*, such that we could have a dark matter planet exactly co-incident with Earth and never know. Maybe dark matter people are walking all around us and through us, maybe my house is in the Times Square of a great dark matter city, maybe a few meters away from me a dark matter blogger is writing on his dark matter computer about how weird it would be if there was a light matter person he couldn't see right next to him.

This is sort of how I feel about conservatives.

I don't mean the sort of light-matter conservatives who go around complaining about Big Government and occasionally voting for Romney. I see those guys all the time. What I mean is – well, take creationists. According to Gallup polls, about 46% of Americans are creationists. Not just in the sense of believing God helped guide evolution. I mean they think evolution is a vile atheist lie and God created humans exactly as they exist right now. That's half the country.

And I don't have a *single one of those people* in my social circle. It's not because I'm deliberately avoiding them; I'm pretty live-and-let-live politically, I wouldn't ostracize someone just for some weird beliefs. And yet, even though I probably know about a hundred fifty people, I am pretty confident that not one of them is creationist. Odds of this happening by chance? 1/2^150 = 1/10^45 = approximately the chance of picking a particular atom if you are randomly selecting among all the atoms on Earth.

About forty percent of Americans want to ban gay marriage. I think if I *really* stretch it, maybe ten of my top hundred fifty friends might fall into this group. This is less astronomically unlikely; the odds are a mere one to one hundred quintillion against.

People like to talk about social bubbles, but that doesn't even begin to cover one hundred quintillion. The only metaphor that seems really appropriate is the bizarre dark matter world.

I live in a Republican congressional district in a state with a Republican governor. The conservatives are definitely out there. They drive on the same roads as I do, live in the same neighborhoods. But they might as well be made of dark matter. I never meet them.

To be fair, I spend a lot of my time inside on my computer. I'm browsing sites like Reddit.

Recently, there was a thread on Reddit asking – Redditors Against Gay Marriage, What Is Your Best Supporting Argument? A Reddit user who didn't understand how anybody could be against gay marriage honestly wanted to know how other people who *were* against it justified their position. He figured he might as well ask one of the largest sites on the Internet, with an estimated user base in the tens of millions.

It soon became clear that nobody there was actually against gay marriage.

There were a bunch of posts saying "I of course support gay marriage but here are some reasons some other people might be against it," a bunch of others saying "my argument against gay marriage is the government shouldn't be involved in the marriage business at all", and several more saying "why would you even ask this question, there's no possible good argument and you're wasting your time". About halfway through the thread someone started saying homosexuality was unnatural and I *thought* they were going to be the first one to actually answer the question, but at the end they added “But it’s not my place to decide what is or isn’t natural, I’m still pro-gay marriage.”

In a thread with 10,401 comments, a thread *specifically* asking for people against gay marriage, I was eventually able to find *two* people who came out and opposed it, way near the bottom. Their posts started with “I know I’m going to be downvoted to hell for this…”

But I’m not only on Reddit. I also hang out on LW.

On last year’s survey, I found that of American LWers who identify with one of the two major political parties, 80% are Democrat and 20% Republican, which actually sounds pretty balanced compared to some of these other examples.

But it doesn’t last. Pretty much all of those “Republicans” are libertarians who consider the GOP the lesser of two evils. When allowed to choose “libertarian” as an alternative, only 4% of visitors continued to identify as conservative. But that’s still… some. Right?

When I broke the numbers down further, 3 percentage points of those are neoreactionaries, a bizarre sect that wants to be ruled by a king. Only *one percent* of LWers were normal everyday God-‘n-guns-but-not-George-III conservatives of the type that seem to make up about half of the United States.

It gets worse. My formative years were spent at a university which, if it was similar to other elite universities, had a faculty and a student body that skewed about 90-10 liberal to conservative – and we can bet that, like LW, even those few token conservatives are Mitt Romney types rather than God-n’-guns types. I get my news from vox.com, an Official Liberal Approved Site. Even when I go out to eat, it turns out my favorite restaurant, California Pizza Kitchen, is the most liberal restaurant in the United States.

I inhabit the same geographical area as *scores and scores* of conservatives. But without meaning to, I have created an *outrageously* strong bubble, a 10^45 bubble. Conservatives are all around me, yet I am about as likely to have a serious encounter with one as I am a Tibetan lama.

(Less likely, actually. One time a Tibetan lama came to my college and gave a really nice presentation, but if a conservative tried that, people would protest and it would be canceled.)

IV.

One day I realized that entirely by accident I was fulfilling *all* the Jewish stereotypes.

I’m nerdy, over-educated, good with words, good with money, weird sense of humor, don’t get outside much, I like deli sandwiches. And I’m a psychiatrist, which is about the most stereotypically Jewish profession short of maybe stand-up comedian or rabbi.

I’m not very religious. And I don’t go to synagogue. But *that’s* stereotypically Jewish too!

I bring this up because it would be a mistake to think “Well, a Jewish person is by definition someone who is born of a Jewish mother. Or I guess it sort of also means someone who follows the Mosaic Law and goes to synagogue. But I don’t care about

Scott’s mother, and I know he doesn’t go to synagogue, so I can’t gain any useful information from knowing Scott is Jewish.”

The defining factors of Judaism – Torah-reading, synagogue-following, mother-having – are the tip of a giant iceberg. Jews sometimes identify as a “tribe”, and even if you don’t attend synagogue, you’re still a member of that tribe and people can still (in a statistical way) infer things about you by knowing your Jewish identity – like how likely they are to be psychiatrists.

The last section raised a question – if people rarely select their friends and associates and customers explicitly for politics, how do we end up with such intense political segregation?

Well, in the same way “going to synagogue” is merely the iceberg-tip of a Jewish tribe with many distinguishing characteristics, so “voting Republican” or “identifying as conservative” or “believing in creationism” is the iceberg-tip of a conservative tribe with many distinguishing characteristics.

A disproportionate number of my friends are Jewish, because I meet them at psychiatry conferences or something – we self-segregate not based on explicit religion but on implicit tribal characteristics. So in the same way, political tribes self-segregate to an impressive extent – a 1/10^45 extent, I will never tire of hammering in – based on their implicit tribal characteristics.

The people who are actually into this sort of thing sketch out a bunch of speculative tribes and subtribes, but to make it easier, let me stick with two and a half.

The Red Tribe is most classically typified by conservative political beliefs, strong evangelical religious beliefs, creationism, opposing gay marriage, owning guns, eating steak, drinking Coca-Cola, driving SUVs, watching lots of TV, enjoying American football, getting conspicuously upset about terrorists and commies, marrying early, divorcing early, shouting “USA IS NUMBER ONE!!!”, and listening to country music.

The Blue Tribe is most classically typified by liberal political beliefs, vague agnosticism, supporting gay rights, thinking guns are barbaric, eating arugula, drinking fancy bottled water, driving Priuses, reading lots of books, being highly educated, mocking American football, feeling vaguely like they should like soccer but never really being able to get into it, getting conspicuously upset about sexists and bigots, marrying later, constantly pointing out how much more civilized European countries are than America, and listening to “everything except country”.

(There is a partly-formed attempt to spin off a Grey Tribe typified by libertarian political beliefs, Dawkins-style atheism, vague annoyance that the question of gay rights even comes up, eating paleo, drinking Soylent, calling in rides on Uber, reading lots of blogs, calling American football “sportsball”, getting conspicuously upset about the War on Drugs and the NSA, and listening to filk – but for our current purposes this is a distraction and they can safely be considered part of the Blue Tribe most of the time)

I think these “tribes” will turn out to be even stronger categories than politics. Harvard might skew 80-20 in terms of Democrats vs. Republicans, 90-10 in terms of liberals vs. conservatives, but maybe 99-1 in terms of Blues vs. Reds.

It’s the many, many differences between these tribes that explain the strength of the filter bubble – which *have I mentioned* segregates people at a strength of 1/10^45?

Even in something as seemingly politically uncharged as going to California Pizza Kitchen or Sushi House for dinner, I'm restricting myself to the set of people who like cute artisanal pizzas or sophsticated foreign foods, which are classically Blue Tribe characteristics.

Are these tribes based on geography? Are they based on race, ethnic origin, religion, IQ, what TV channels you watched as a kid? I don't know.

Some of it is certainly genetic – estimates of the genetic contribution to political association range from 0.4 to 0.6. Heritability of one's attitudes toward gay rights range from 0.3 to 0.5, which hilariously is a little more heritable than homosexuality itself.

(for an interesting attempt to break these down into more rigorous concepts like "traditionalism", "authoritarianism", and "in-group favoritism" and find the genetic loading for each see here. For an attempt to trace the specific genes involved, which mostly turn out to be NMDA receptors, see here)

But I don't think it's just genetics. There's something else going on too. The word "class" seems like the closest analogue, but only if you use it in the sophisticated Paul Fussell *Guide Through the American Status System* way instead of the boring "another word for how much money you make" way.

For now we can just accept them as a brute fact – as multiple coexisting societies that might as well be made of dark matter for all of the interaction they have with one another – and move on.

V.

The worst reaction I've ever gotten to a blog post was when I wrote about the death of Osama bin Laden. I've written all sorts of stuff about race and gender and politics and whatever, but that was the worst.

I didn't come out and say I was happy he was dead. But some people interpreted it that way, and there followed a bunch of comments and emails and Facebook messages about how could I possibly be happy about the death of another human being, even if he was a bad person? Everyone, even Osama, is a human being, and we should never rejoice in the death of a fellow man. One commenter came out and said:

> I'm surprised at your reaction. As far as people I casually stalk on the internet (ie, LJ and Facebook), you are the first out of the "intelligent, reasoned and thoughtful" group to be uncomplicatedly happy about this development and not to be, say, disgusted at the reactions of the other 90% or so.

This commenter was right. Of the "intelligent, reasoned, and thoughtful" people I knew, the overwhelming emotion was conspicuous disgust that other people could be happy about his death. I hastily backtracked and said I wasn't happy per se, just surprised and relieved that all of this was finally behind us.

And I genuinely believed that day that I had found some unexpected good in people – that everyone I knew was so humane and compassionate that they were unable to rejoice even in the death of someone who hated them and everything they stood for.

Then a few years later, Margaret Thatcher died. And on my Facebook wall – made of these same "intelligent, reasoned, and thoughtful" people – the most common response was to quote some portion of the song “Ding Dong, The Witch Is Dead”. Another popular response was to link the videos of British people spontaneously throwing parties in the street, with comments like “I wish I was there so I could join in”. From this exact same group of people, not a single expression of disgust or a “c’mon, guys, we’re all human beings here.”

I gently pointed this out at the time, and mostly got a bunch of “yeah, so what?”, combined with links to an article claiming that “the demand for respectful silence in the wake of a public figure’s death is not just misguided but dangerous”.

And that was when something clicked for me.

You can talk all you want about Islamophobia, but my friend’s “intelligent, reasoned, and thoughtful people” – her name for the Blue Tribe – can’t get together enough energy to really hate Osama, let alone Muslims in general. We understand that what he did was bad, but it didn’t anger us personally. When he died, we were able to very rationally apply our better nature and our Far Mode beliefs about how it’s never right to be happy about anyone else’s death.

On the other hand, that same group absolutely *loathed* Thatcher. Most of us (though not all) can agree, if the question is posed explicitly, that Osama was a worse person than Thatcher. But in terms of actual gut feeling? Osama provokes a snap judgment of “flawed human being”, Thatcher a snap judgment of “scum”.

I started this essay by pointing out that, despite what geographical and cultural distance would suggest, the Nazis’ outgroup was not the vastly different Japanese, but the almost-identical German Jews.

And my hypothesis, stated plainly, is that if you’re part of the Blue Tribe, then your outgroup isn’t al-Qaeda, or Muslims, or blacks, or gays, or transpeople, or Jews, or atheists – it’s the Red Tribe.

VI.

“But racism and sexism and cissexism and anti-Semitism are these giant all-encompassing social factors that verge upon being human universals! Surely you’re not arguing that mere *political* differences could ever come close to them!”

One of the ways we *know* that racism is a giant all-encompassing social factor is the Implicit Association Test. Psychologists ask subjects to quickly identify whether words or photos are members of certain gerrymandered categories, like “either a white person’s face or a positive emotion” or “either a black person’s face and a negative emotion”. Then they compare to a different set of gerrymandered categories, like “either a black person’s face or a positive emotion” or “either a white person’s face or a negative emotion.” If subjects have more trouble (as measured in latency time) connecting white people to negative things than they do white people to positive things, then they probably have subconscious positive associations with white people. You can try it yourself here.

Of course, what the test famously found was that even white people who claimed to have no racist attitudes at all usually had positive associations with white people and negative associations with black people on the test. There are very many claims and counterclaims about the precise meaning of this, but it ended up being a big part of the evidence in favor of the current consensus that all white people are at least a little racist.

Anyway, three months ago, someone finally had the bright idea of doing an Implicit Association Test with political parties, and they found that people's unconscious partisan biases were *half again as strong* as their unconscious racial biases (h/t Bloomberg. For example, if you are a white Democrat, your unconscious bias against blacks (as measured by something called a d-score) is 0.16, but your unconscious bias against Republicans will be 0.23. The Cohen's *d* for racial bias was 0.61, by the book a "moderate" effect size; for party it was 0.95, a "large" effect size.

Okay, fine, but we know race has *real world* consequences. Like, there have been several studies where people sent out a bunch of identical resumes except sometimes with a black person's photo and other times with a white person's photo, and it was noticed that employers were much more likely to invite the fictional white candidates for interviews. So just some stupid Implicit Association Test results can't compare to that, right?

Iyengar and Westwood also decided to do the resume test for parties. They asked subjects to decide which of several candidates should get a scholarship (subjects were told this was a genuine decision for the university the researchers were affiliated with). Some resumes had photos of black people, others of white people. And some students listed their experience in Young Democrats of America, others in Young Republicans of America.

Once again, discrimination on the basis of party was much stronger than discrimination on the basis of race. The size of the race effect for white people was only 56-44 (and in the reverse of the expected direction); the size of the party effect was about 80-20 for Democrats and 69-31 for Republicans.

If you want to see their third experiment, which applied *yet another* classic methodology used to detect racism and *once again* found partyism to be much stronger, you can read the paper.

I & W did an unusually thorough job, but this sort of thing isn't new or groundbreaking. People have been studying "belief congruence theory" – the idea that differences in beliefs are more important than demographic factors in forming ingroups and outgroups – for decades. As early as 1967, Smith et al were doing surveys all over the country and finding that people were more likely to accept friendships across racial lines than across beliefs; in the forty years since then, the observation has been replicated scores of times. Insko, Moe, and Nacoste's 2006 review Belief Congruence And Racial Discrimination concludes that:

> . The literature was judged supportive of a weak version of belief congruence theory which states that in those contexts in which social pressure is nonexistent or ineffective, belief is more important than race as a determinant of racial or ethnic discrimination. Evidence for a strong version of belief congruence theory (which states that in those contexts in which social pressure is nonexistent, or ineffective, belief is the only determinant of racial or ethnic discrimination) and was judged much more problematic.

One of the best-known examples of racism is the "Guess Who's Coming To Dinner" scenario where parents are scandalized about their child marrying someone of a different race. Pew has done some good work on this and found that only 23% of conservatives and 1% (!) of liberals admit they would be upset in this situation. But Pew *also* asked how parents would feel about their child marrying someone of a different *political party*. Now 30% of conservatives and 23% of liberals would get upset. Average them out, and you go from 12% upsetness rate for race to 27% upsetness rate for party – more than double. Yeah, people do lie to pollsters, but a picture is starting to come together here.

(Harvard, by the way, is a tossup. There are more black students – 11.5% – than conservative students – 10% – but there are more conservative faculty than black faculty.)

Since people will delight in misinterpreting me here, let me overemphasize what I am *not* saying. I'm not saying people of either party have it "worse" than black people, or that partyism is more of a *problem* than racism, or any of a number of stupid things along those lines which I am sure I will nevertheless be accused of believing. Racism is worse than partyism because the two parties are at least kind of balanced in numbers and in resources, whereas the brunt of an entire country's racism falls on a few underprivileged people. I am saying that the *underlying attitudes that produce* partyism are stronger than the underlying attitudes that produce racism, with no necessary implications on their social effects.

But if we want to look at people's psychology and motivations, partyism and the particular variant of tribalism that it represents are going to be fertile ground.

VII.

Every election cycle like clockwork, conservatives accuse liberals of not being sufficiently pro-America. And every election cycle like clockwork, liberals give extremely unconvincing denials of this.

"It's not that we're, like, *against* America per se. It's just that…well, did you know Europe has much better health care than we do? And much lower crime rates? I mean, come on, how did they get so awesome? And we're just sitting here, can't even get the gay marriage thing sorted out, seriously, what's wrong with a country that can't… sorry, what were we talking about? Oh yeah, America. They're okay. Cesar Chavez was really neat. So were some other people outside the mainstream who became famous precisely by criticizing majority society. That's *sort of* like America being great, in that I think the parts of it that point out how bad the rest of it are often make excellent points. Vote for me!"

(sorry, I make fun of you because I love you)

There was a big brouhaha a couple of years ago when, as it first became apparent Obama had a good shot at the Presidency, Michelle Obama said that "for the first time in my adult life, I am proud of my country."

Republicans pounced on the comment, asking why she hadn't felt proud before, and she backtracked saying of course she was proud all the time and she loves America with the burning fury of a million suns and she was just saying that the Obama campaign was *particularly* inspiring.

As unconvincing denials go, this one was pretty far up there. But no one really held it against her. Probably most Obama voters felt vaguely the same way. *I* was an Obama voter, and I have proud memories of spending my Fourth of Julys as a kid debunking people's heartfelt emotions of patriotism. Aaron Sorkin:

> [What makes America the greatest country in the world?] It's not the greatest country in the world! We're seventh in literacy, 27th in math, 22nd in science,

> 49th in life expectancy, 178th in infant mortality, third in median household income, No. 4 in labor force, and No. 4 in exports. So when you ask what makes us the greatest country in the world, I don't know what the f*** you're talking about.

(Another good retort is "We're number one? Sure – number one in incarceration rates, drone strikes, and making new parents go back to work!")

All of this is true, of course. But it's weird that it's such a classic interest of members of the Blue Tribe, and members of the Red Tribe never seem to bring it up.

("We're number one? Sure – number one in levels of sexual degeneracy! Well, I guess probably number two, after the Netherlands, but they're really small and shouldn't count.")

My hunch – both the Red Tribe and the Blue Tribe, for whatever reason, identify "America" with the Red Tribe. Ask people for typically "American" things, and you end up with a very Red list of characteristics – guns, religion, barbecues, American football, NASCAR, cowboys, SUVs, unrestrained capitalism.

That means the Red Tribe feels intensely patriotic about "their" country, and the Blue Tribe feels like they're living in fortified enclaves deep in hostile territory.

Here is a popular piece published on a major media site called America: A Big, Fat, Stupid Nation. Another: America: A Bunch Of Spoiled, Whiny Brats. Americans are ignorant, scientifically illiterate religious fanatics whose "patriotism" is actually just narcissism. You Will Be Shocked At How Ignorant Americans Are, and we should Blame The Childish, Ignorant American People.

Needless to say, every single one of these articles was written by an American and read almost entirely by Americans. Those Americans very likely enjoyed the articles very much and did not feel the least bit insulted.

And look at the sources. HuffPo, Salon, Slate. Might those have anything in common?

On both sides, "American" can be either a normal demonym, or a code word for a member of the Red Tribe.

VIII.

The other day, I logged into OKCupid and found someone who looked cool. I was reading over her profile and found the following sentence:

> Don't message me if you're a sexist white guy

And my first thought was "Wait, so a sexist black person would be okay? Why?"

(The girl in question was white as snow)

Around the time the Ferguson riots were first starting, there were a host of articles with titles like Why White People Don't Seem To Understand Ferguson, Why It's So Hard For Whites To Understand Ferguson, and White Folks Listen Up And Let Me Tell You What Ferguson Is All About, this last of which says:

> Social media is full of people on both sides making presumptions, and believing what they want to believe. But it's the white folks that don't understand what this is all about. Let me put it as simply as I can for you […]

> No matter how wrong you think Trayvon Martin or Michael Brown were, I think we can all agree they didn't deserve to die over it. I want you white folks to understand that this is where the anger is coming from. You focused on the looting…."

And on a hunch I checked the author photos, and every single one of these articles was written by a white person.

White People Are Ruining America? White. White People Are Still A Disgrace? White. White Guys: We Suck And We're Sorry? White. Bye Bye, Whiny White Dudes? White. Dear Entitled Straight White Dudes, I'm Evicting You From My Life? White. White Dudes Need To Stop Whitesplaining? White. Reasons Why Americans Suck #1: White People? White.

We've all seen articles and comments and articles like this. Some unsavory people try to use them to prove that white people are the *real* victims or the media is biased against white people or something. Other people who are very nice and optimistic use them to show that some white people have developed some self-awareness and are willing to engage in self-criticism.

But I think the situation with "white" is much the same as the situation with "American" – it can either mean what it says, or be a code word for the Red Tribe.

(except on the blog Stuff White People Like, where it obviously serves as a code word for the *Blue* tribe. I don't know, guys. I didn't do it.)

I realize that's making a strong claim, but it would hardly be without precedent. When people say things like "gamers are misogynist", do they mean the 52% of gamers who are women? Do they mean every one of the 59% of Americans from every walk of life who are known to play video or computer games occasionally? No. "Gamer" is a coded reference to the Gray Tribe, the half-branched-off collection of libertarianish tech-savvy nerds, and everyone knows it. As well expect that when people talk about "fedoras", they mean Indiana Jones. Or when they talk about "urban youth", they mean freshmen at NYU. Everyone knows exactly who we mean when we say "urban youth", and them being young people who live in a city has only the most tenuous of relations to the actual concept.

And I'm saying words like "American" and "white" work the same way. Bill Clinton was the "first black President", but if Herman Cain had won in 2012 he'd have been the 43rd white president. And when an angry white person talks at great length about how much he hates "white dudes", *he is not being humble and self-critical*.

IX.

Imagine hearing that a liberal talk show host and comedian was so enraged by the actions of ISIS that he'd recorded and posted a video in which he shouts at them for ten minutes, cursing the "fanatical terrorists" and calling them "utter savages" with "savage values".

If *I* heard that, I'd be kind of surprised. It doesn't fit my model of what liberal talk show hosts do.

But the story I'm *actually* referring to is liberal talk show host / comedian Russell Brand making that same rant against Fox News for *supporting war against* the Islamic State, adding at the end that "Fox is worse than ISIS".

That fits my model perfectly. You wouldn’t celebrate Osama’s death, only Thatcher’s. And you wouldn’t call ISIS savages, only Fox News. Fox is the outgroup, ISIS is just some random people off in a desert. You hate the outgroup, you don’t hate random desert people.

I would go further. Not only does Brand not feel much like hating ISIS, he has a strong incentive not to. That incentive is: the Red Tribe is known to hate ISIS loudly and conspicuously. Hating ISIS would signal Red Tribe membership, would be the equivalent of going into Crips territory with a big Bloods gang sign tattooed on your shoulder.

But this might be unfair. What would Russell Brand answer, if we asked him to justify his decision to be much angrier at Fox than ISIS?

He might say something like “Obviously Fox News is not literally worse than ISIS. But here I am, talking to my audience, who are mostly white British people and Americans. These people already know that ISIS is bad; they don’t need to be told that any further. In fact, at this point being angry about how bad ISIS is, is less likely to genuinely change someone’s mind about ISIS, and more likely to promote Islamophobia. The sort of people in my audience are at zero risk of becoming ISIS supporters, but at a very real risk of Islamophobia. So ranting against ISIS would be counterproductive and dangerous.

On the other hand, my audience of white British people and Americans is very likely to contain many Fox News viewers and supporters. And Fox, while not quite as evil as ISIS, is still pretty bad. So here’s somewhere I have a genuine chance to reach people at risk and change minds. Therefore, I think my decision to rant against Fox News, and maybe hyperbolically say they were ‘worse than ISIS’ is justified under the circumstances.”

I have a lot of sympathy to hypothetical-Brand, especially to the part about Islamophobia. It *does* seem really possible to denounce ISIS’ atrocities to a population that already hates them in order to weak-man a couple of already-marginalized Muslims. We need to fight terrorism and atrocities – therefore it’s okay to shout at a poor girl ten thousand miles from home for wearing a headscarf in public. Christians are being executed for their faith in Sudan, therefore let’s picket the people trying to build a mosque next door.

But my sympathy with Brand ends when he acts like his audience is likely to be fans of Fox News.

In a world where a negligible number of Redditors oppose gay marriage and 1% of Less Wrongers identify conservative and I know 0/150 creationists, how many of the people who visit the YouTube channel of a well-known liberal activist with a Che-inspired banner, a channel whose episode names are things like “War: What Is It Good For?” and “Sarah Silverman Talks Feminism” – how many of them do you think are big Fox News fans?

In a way, Russell Brand would have been *braver* taking a stand against ISIS than against Fox. If he attacked ISIS, his viewers would just be a little confused and uncomfortable. Whereas every moment he’s attacking Fox his viewers are like “HA HA! YEAH! GET ‘EM! SHOW THOSE IGNORANT BIGOTS IN THE OUTGROUP WHO’S BOSS!”

Brand acts as if there are just these countries called “Britain” and “America” who are receiving his material. Wrong. There are two parallel universes, and he’s only broadcasting to one of them.

The result is exactly what we predicted would happen in the case of Islam. Bombard people with images of a far-off land they already hate and tell them to hate it more, and the result is ramping up the intolerance on the couple of dazed and marginalized representatives of that culture who have ended up stuck on your half of the divide. Sure enough, if industry or culture or community gets Blue enough, Red Tribe members start getting harassed, fired from their jobs (Brendan Eich being the obvious example) or otherwise shown the door.

Think of Brendan Eich as a member of a tiny religious minority surrounded by people who hate that minority. Suddenly firing him doesn’t seem very noble.

If you mix together Podunk, Texas and Mosul, Iraq, you can prove that Muslims are scary and very powerful people who are executing Christians all the time – and so we have a great excuse for kicking the one remaining Muslim family, random people who never hurt anyone, out of town.

And if you mix together the open-source tech industry and the parallel universe where you can’t wear a FreeBSD t-shirt without risking someone trying to exorcise you, you can prove that Christians are scary and very powerful people who are persecuting everyone else all the time, and you have a great excuse for kicking one of the few people willing to affiliate with the Red Tribe, a guy who never hurt anyone, out of town.

When a friend of mine heard Eich got fired, she didn’t see anything wrong with it. “I can tolerate anything except intolerance,” she said.

“Intolerance” is starting to look like another one of those words like “white” and “American”.

“I can tolerate anything except the outgroup.” Doesn’t sound quite so noble now, does it?

X.

We started by asking: millions of people are conspicuously praising every outgroup they can think of, while conspicuously condemning their own in-group. This seems contrary to what we know about social psychology. What’s up?

We noted that outgroups are rarely literally “the group most different from you”, and in fact far more likely to be groups very similar to you sharing *almost* all your characteristics and living in the same area.

We then noted that although liberals and conservatives live in the same area, they might as well be two totally different countries or universe as far as level of interaction were concerned.

Contra the usual idea of them being marked only by voting behavior, we described them as very different tribes with totally different cultures. You can speak of “American culture” only in the same way you can speak of “Asian culture” – that is, with a lot of interior boundaries being pushed under the rug.

The outgroup of the Red Tribe is occasionally blacks and gays and Muslims, more often the Blue Tribe.

The Blue Tribe has performed some kind of very impressive act of alchemy, and transmuted *all* of its outgroup hatred to the Red Tribe.

This is not surprising. Ethnic differences have proven quite tractable in the face of shared strategic aims. Even the Nazis, not known for their ethnic tolerance, were able to get all buddy-buddy with the Japanese when they had a common cause.

Research suggests Blue Tribe / Red Tribe prejudice to be much stronger than better-known types of prejudice like racism. Once the Blue Tribe was able to enlist the blacks and gays and Muslims in their ranks, they became allies of convenience who deserve to be rehabilitated with mildly condescending paeans to their virtue. “There never was a coward where the shamrock grows.”

Spending your entire life insulting the other tribe and talking about how terrible they are makes you look, well, tribalistic. It is definitely not high class. So when members of the Blue Tribe decide to dedicate their entire life to yelling about how terrible the Red Tribe is, they make sure that instead of saying “the Red Tribe”, they say “America”, or “white people”, or “straight white men”. That way it’s *humble self-criticism*. They are *so* interested in justice that they are willing to critique *their own beloved side*, much as it pains them to do so. We know they are not exaggerating, because one might exaggerate the flaws of an enemy, but that anyone would exaggerate their *own* flaws fails the criterion of embarrassment.

The Blue Tribe always has an excuse at hand to persecute and crush any Red Tribers unfortunate enough to fall into its light-matter-universe by defining them as all-powerful domineering oppressors. They appeal to the fact that this is definitely the way it works in the Red Tribe’s dark-matter-universe, and that’s in the same country so it has to be the same community for all intents and purposes. As a result, every Blue Tribe institution is permanently licensed to take whatever emergency measures are necessary against the Red Tribe, however disturbing they might otherwise seem.

And so how virtuous, how noble the Blue Tribe! Perfectly tolerant of all of the different groups that just so happen to be allied with them, never intolerant unless it happen to be against intolerance itself. Never stooping to engage in petty tribal conflict like that awful Red Tribe, but always nobly criticizing their own culture and striving to make it better!

Sorry. But I hope this is at least a *little* convincing. The weird dynamic of outgroup-philia and ingroup-phobia isn’t anything of the sort. It’s just good old-fashioned in-group-favoritism and outgroup bashing, a little more sophisticated and a little more sneaky.

XI.

This essay is bad and I should feel bad.

I should feel bad because I made *exactly* the mistake I am trying to warn everyone else about, and it wasn’t until I was almost done that I noticed.

How virtuous, how noble I must be! Never stooping to engage in petty tribal conflict like that silly Red Tribe, but always nobly criticizing my own tribe and striving to make it better.

Yeah. Once I’ve written a ten thousand word essay savagely attacking the Blue Tribe, either I’m a very special person or they’re my outgroup. And I’m not *that* special.

Just as you can pull a fast one and look humbly self-critical if you make your audience assume there’s just one American culture, so maybe you can trick people by assuming there’s only one Blue Tribe.

I’m pretty sure I’m not Red, but I did talk about the Grey Tribe above, and I show all the risk factors for being one of them. That means that, although my critique of the Blue Tribe may be right or wrong, in terms of *motivation* it comes from the same place as a Red Tribe member talking about how much they hate al-Qaeda or a Blue Tribe member talking about how much they hate ignorant bigots. And when I boast of being able to tolerate Christians and Southerners whom the Blue Tribe is mean to, I’m not being tolerant at all, just noticing people so far away from me they wouldn’t make a good outgroup anyway.

I had *fun* writing this article. People do not have fun writing articles savagely criticizing their in-group. People can criticize their in-group, it’s not *humanly impossible*, but it takes nerves of steel, it makes your blood boil, you should sweat blood. It shouldn’t be *fun*.

You can bet some white guy on Gawker who week after week churns out “Why White People Are So Terrible” and “Here’s What Dumb White People Don’t Understand” is having fun and not sweating any blood at all. He’s not criticizing his in-group, he’s never even *considered* criticizing his in-group. I can’t blame him. Criticizing the in-group is a really difficult project I’ve barely begun to build the mental skills necessary to even consider.

I can think of criticisms of my own tribe. Important criticisms, true ones. But the thought of writing them makes my blood boil.

I imagine might I feel like some liberal US Muslim leader, when he goes on the O’Reilly Show, and O’Reilly ambushes him and demands to know why he and other American Muslims haven’t condemned beheadings by ISIS more, demands that he criticize them right there on live TV. And you can see the wheels in the Muslim leader’s head turning, thinking something like “Okay, obviously beheadings are terrible and I hate them as much as anyone. But you don’t care even *the slightest bit* about the victims of beheadings. You’re just looking for a way to score points against me so you can embarass all Muslims. And I would rather personally behead every single person in the world than give a smug bigot like you a single microgram more stupid self-satisfaction than you’ve already got.”

That is how I feel when asked to criticize my own tribe, even for correct reasons. If you think you’re criticizing your own tribe, and your blood is not at that temperature, consider the possibility that you aren’t.

But if I want Self-Criticism Virtue Points, criticizing the Grey Tribe is the only honest way to get them. And if I want Tolerance Points, my own personal cross to bear right now is tolerating the Blue Tribe. I need to remind myself that when they are bad people, they are merely Osama-level bad people instead of Thatcher-level bad people. And when they are good people, they are powerful and necessary crusaders against the evils of the world.

The worst thing that could happen to this post is to have it be used as convenient feces to fling at the Blue Tribe whenever feces are necessary. Which, given what has happened to my last couple of posts along these lines and the obvious biases of my own subconscious, I already expect it will be.

But the best thing that could happen to this post is that it makes a lot of people, especially myself, figure out how to be more tolerant. Not in the “of course I’m tolerant, why shouldn’t I be?” sense of the Emperor in Part I. But in the sense of “being tolerant makes me see red, makes me sweat blood, but darn it *I am going to be tolerant anyway*.”