
Supposed to be HPMOR! Squirrel standing in for Alexander the Great
Mangy mutts
Diogenes hangs out with
Diogenes as a badger

Hufflepuff Cynicism

Hufflepuff Cynicism

Summary: In response to catching a glimpse of the dark world, and especially of the extent of human hypocrisy with respect to the dark world, one might take a dim view of one's fellow humans. I describe an alternative, *Hufflepuff cynicism,* in which you lower your concept of what the standards were all along. I give arguments for and against this perspective.

This came out of conversations with Steve Rayhawk, Harmanas Chopra, Ziz, and Kurt Brown.

In See the Dark World, Nate describes a difficulty which people have in facing the worst of a situation. In one of his examples, in which Alice and Bob grapple with the idea that the market price of saving a life is $3000, Bob's response is to deny what this means about people -- to avoid concluding that humans aren't generally willing to pay more than 3k to save a life, Bob decides that money is not a relevant measure of how much people would sacrifice to save a life. This is a non-sequiter, but allows Bob to sleep at night.

Alice, on the other hand, decides that a life really is only worth about 3k. This, to me, is something like Hansonian cynicism -- when revealed preferences differ from stated preferences, assume that people are lying about what they value.

Nate calls both Alice and Bob's response "tolerification", and recommends undoing tolerification by entertaining the what-if question describing the dark world (following the leave-a-line-of-retreat method).

Ziz gives a somewhat similar analysis in social reality: when you recognize the distance between the social reality most people are willing to endorse and the real reality, you have three options:

> Make a buckets error where your map of reality overwrites your map of social reality, and you have the “infuriating perspective”, typified by less-cunning activists and people new to their forbidden truths. “No, it is not ‘a personal choice’, which means people can’t hide from the truth. I can call people out and win arguments”.
>
> Make a buckets error where your map of social reality overwrites your map of reality, and you have the “dehumanizing perspective” of someone who is a vegetarian for ethical reasons but believes truly feels it when they say “it’s a personal choice”, the atheist who respects religion-the-proposition, to some extent the trans person who feels the gender presentation they want would be truly out of line…
>
> *But it was all right, everything was all right, the struggle was finished. He had won the victory over himself. He loved Big Brother.*
>
> Learn to deeply track the two as separate, and you have the “isolating perspective”. It is isolating to let it entirely into your soul, the knowledge that “people are good and rational” is pretense.

This is importantly different from Nate's analysis in several ways, but I'll leave that as an exercise for the reader.

What's interesting to me is that in certain situations like this, I seem to take what Ziz calls the "isolating perspective" in a way which is not isolating at all -- I can cleanly seperate between the social reality and what's real, but it doesn't feel especially isolating. Humans are like cats. Sure, sometimes they claw up the upholstry, but it's no use being upset -- you have the choice of declawing them, or trying to train them by spraying them with water. Humans are not automatically strategic. Work with what's there.

I call this Hufflepuff cynicism because I think most people assume a cynic is like a slytherin (or, in Ziz's case, a sith) -- but hufflepuffs, above all, have to deal with people as they are. It seems to me that Hufflepuff cynicism is where you end up if you start out as a starry-eyed young Hufflepuff who just wants to do good, and then you become a teacher or a doctor and have to help with a whole bunch of people who are more ignorant on a subject than you. Helping people doesn't mean correcting every error. It means nudging people in the right direction as best you can. This perspective allows you to avoid creating Hufflepuff Traps.

ETA, more precise definition of Hufflepuff cynicism:

- Track social reality and real reality seperately, but don't get upset when this allows you to recognize ways that people don't live up to standards which they would claim to endorse.
- Adapt your sense of the standards to roughly conserve the amount of "violation of basic standards" you see, rather than letting it increase as you see better.
- Speak to others in the way you expect them to understand: work around truths which you suspect they're unable or unwilling to accept. Feel no negative judgement of them in doing so.
- Have a strong enough assumption of good faith around other people's beliefs and actions that you want to chesterton's-fence them before correcting them.

This certainly isn't a perfect response. One problem is that sometimes voicing my real thoughts lead to horrible awkwardness or social punishment. Tracking social reality and real reality seperately isn't very helpful in circumstances where I really want to be honest for one reason or another. However, I recently talked with some people about my hufflepuff cynicism, and they had some arguments against it which surprised me and which I found fairly compelling.

You Can't Choose between Lying to Others and Lying to Yourself

In Honesty: Beyond Internal Truth, Eliezer discusses two paths to rationality: scrupulous truth-telling, vs allowing yourself to lie to others so that you can keep your internal beliefs uncontaminated by social incentives. Eliezer preferred the first route, but granted the possibility of the second route. However, it seems like what really happens when you lie to others is that you start to believe the lies. Paul Christiano discusses how humans are bad at lying in If We Can't Lie to Others, We Will Lie to Ourselves. Scott Alexander recently discussed how, in hindsight, many of his attempts to seperately track what's true and what he can say have failed:

> Sometimes I can almost feel this happening. First I believe something is true, and say so. Then I realize it’s considered low-status and cringeworthy. Then I make a principled decision to avoid saying it – or say it only in a very careful way – in order to protect my reputation and ability to participate in society. Then when other people say it, I start looking down on them for being bad at public relations. Then I start looking down on them just for being low-status or cringeworthy. Finally the idea of “low-status” and “bad and wrong” have merged so fully in my mind that the idea seems terrible and ridiculous to me, and I only remember it’s true if I force myself to explicitly consider the question. And even then, it’s in a condescending way, where I feel like the people who say it’s true deserve low status for not being smart enough to remember not to say it. This is endemic, and I try to quash it when I notice it, but I don’t know how many times it’s slipped my notice all the way to the point where I can no longer remember the truth of the original statement.

It is not at all clear that you can really make a decision to seperately track what you say to others and what is true. Ziz's "isolating perspective" and my "hufflepuff cynicism" may *sound* like the obvious path out of the cath-22, but perhaps this is not as sustainable as it may seem. Eliezer's path is much closer to Ziz's "infuriating perspective".

Becoming Complicit

A second argument against Hufflepuff cynicism is that you make yourself complicit in a bad social equilibrium. At its root, Hufflepuff cynicism is the claim that it doesn't make sense to hold people up to standards any higher than they're already mostly meeting. From one perspective, this is just common sense -- what are standards for, if they're being broken constantly? From another perspective, however, this is just *refusing to help.* And, in cases of hypocrisy, it may mean refusing to help kick a system out of a broken equilibrium.

The part of me that runs on Hufflepuff cynicism tends to correct someone *exactly once*, and if they don't change, forever assume that they're not interested in or open to advice on the matter. I'm often reluctant to do even that, because of a deeply-seated assumption that people mostly would have corrected an error if they were interested/able.

Here's an anecdote from Kurt Brown:

> The other day, someone was estimating time for me, and for a moment, it seemed that he thought I thought that it was 3 (rather than 3:30, the true time). His claim that something started in 2 hours (at 5) appeared to me to be intentionally propagating the error that I thought he thought I had made, presumably because he didn’t feel the need to correct me. I was disturbed and asked something like “this math doesn’t work out, and it looks like you should know this--are you huffcyn-ing me, bro??”
>
> He was not, he had simply made the same error in a confusing way. I draw no conclusions, but it was an interesting experience!

Refraining from correcting someone about what time it is isn't *exactly* like something my huffcyn algorithm would do, because there's no plausible story in which the person

is just incapable or unwilling to correct their time estimate. However, it feels worryingly close to things I might do.

The worst examples of this, and the most likely to actually happen, involve living within a corrupt system. The Hufflepuffian cynic is likely to pragmatically refuse to speak out, or speak out exactly once and then remain forever silent. This means a system in which all cynics are Hufflepuff cynics is not as self-correcting as one in which people are more likely to take the "infuriating perspective".

Conclusion?

I don't know whether to recommend Hufflepuff cynisism or not. It may help you face the dark world, bringing your beliefs closer in line with the truth. On the other hand, it may make you complicit in the perpetuation of the dark world itself. The way I see it now, there's a catch-22 whose resolution is unclear.

If you try to cleanly distinguish between what you can say and what's true, the latter may lose relevance over time until it is forgotten. If you try to only speak the truth, you may be ostracized or give in to the social incentives by editing your beliefs to be acceptable.

Hopefully, it is *at least* helpful to have the term in your lexicon, to identify Hufflepuff cynicism when it occurs (be it good or ill).

Hufflepuff Cynicism on Crocker's Rule

Yesterday, I mainly talked about Hufflepuff Cynicism from the cynic's end. However, there's a lot to be said about the receiving end. Hufflepuff cynicism can come off as a very patronizing strategy. Is this a point against it?

In the original conversation where I came up with the idea of Hufflepuff cynicism, I was talking about norms for aspiring rationalists around trying to get other people to be more rational. Maybe we agree that double crux is a good conversation procedure, but should we *try to convince* someone of that? Should we try to get them to double-crux with us about it? Maybe we believe you should bet or update when a disagreement hasn't been resolved, but what should we do with a disagreement about the bet-or-update rule?

My argument from the Hufflepuff Cynicism side was in favor of chesterton-fencing such disagreements. Don't try to convince others about rationality norms; at least, stop after the first explanation falls on deaf ears. Instead, figure out why the person isn't already following the norm. It seems likely that there's some important reason; if you can figure it out, maybe you can come up with a better norm which would address the concern (in much the same way bet-or-update addresses objections to the simpler strategies "bet on disagreements" or "talk out disagreements until you converge").

To my surprise, not everyone wants to be treated so carefully. Some people find this attitude patronizing or overly cautious, and request that I *just tell them what they are doing wrong*, possibly telling them *more than just one time* if they don't get it the first time. This is, more or less, an invocation of Crocker's Rule.

To quote the sl4 wiki on Crocker's Rules:

> Declaring yourself to be operating by "Crocker's Rules" means that other people are allowed to optimize their messages for information, not for being nice to you. Crocker's Rules means that you have accepted full responsibility for the operation of your own mind - if you're offended, it's your fault. Anyone is allowed to call you a moron and claim to be doing you a favor. (Which, in point of fact, they would be. One of the big problems with this culture is that everyone's afraid to tell you you're wrong, or they think they have to dance around it.) Two people using Crocker's Rules should be able to communicate all relevant information in the minimum amount of time, without paraphrasing or social formatting. Obviously, don't declare yourself to be operating by Crocker's Rules unless you have that kind of mental discipline.
>
> **Note that Crocker's Rules does *not* mean you can insult people; it means that *other* people don't have to worry about whether *they* are insulting *you*.** Crocker's Rules are a discipline, not a privilege. Furthermore, taking advantage of Crocker's Rules does not imply reciprocity. How could it? Crocker's Rules are something you do for yourself, to maximize information received - *not* something you grit your teeth over and do as a favor.

(This seems like really just one rule to me, so I tend to call it Crocker's Rule.)

A problem I've encountered, which has reinforced my Hufflepuff Cynicism, is that people can invoke Crocker's Rule and then get upset about feedback I give anyway.

My advice is this: Crocker's Rule is a promise not to punish others for giving you negative feedback. If you don't want to be patronized by Hufflepuff Cynics like myself, don't make that promise unless you're sure you can keep it. Instead, *show others through your words and actions over an extended period of time* that you are both able and happy to accept negative feedback. Don't make a standing request for negative feedback. Ask for it explicitly again and again, and thank others for giving it to you. If you can't thank them genuinely, you've learned something about yourself and can adapt accordingly.

But, maybe I'm too much of a Hufflepuff Cynic. I don't know. Maybe I should... hold people to their own standards, sometimes...?

Hufflepuff Cynicism on Hypocrisy

Epistemic Status: I've been thinking about this for a number of years, looking for steelmen of the position against hypocrisy. I haven't found anything satisfying yet, but maybe you can tell me why hypocrisy is actually bad? Barring that, I'm rather confident in the view expressed here.

"Hypocrisy is bad" is a deeply-rooted assumption in our culture. If I'm trying to give someone advice, I'll flinch away from saying things I don't do myself. For example, I don't have a driver's license, so I would hesitate before suggesting that someone else get one. If I *do* suggest it, I'll get kind of apologetic, and hedge my statement with "but I haven't gotten one yet". This, despite the fact that whether I've gotten one is almost irrelevant to whether they should get one. It seems to me that this habit is universal in American culture, and I'd be surprised (and intrigued!) to hear about any culture where it isn't.

I think the anti-hypocrisy norm is likely based on a blame/praise model of advice. If advice is always norm-enforcing criticism, and is tied to how many "social points" you score, then the anti-hypocrisy norm prevents a particular type of social exploit. A hypocrite can enforce, and claim to believe in, norms which they themselves break. They score social points, gaining status and power, based on an unfair self-favoring application of rules. Rulers who are above the laws they themselves enforce violate our fairness norms; they are the very image of corruption.

However, this problem can be addressed in a more specific way by calling out unfairness, building power structures with transparency and accountability in mind, and deserving trust as a community. The cost of the anti-hypocrisy norm is too high; it throws out too much useful advice, constraining the directions in which we think when we're trying to help one another.

Rob Bensinger already wrote a whole post on why hypocrisy is a bad concept, which I also endorse. Ironically, though, he has a different concept of hypocrisy than me, so my argument against the concept is somewhat different. He treats "hypocrisy" as "inconsistency", and points out that winning an argument by pointing out inconsistency is not very informative: showing that someone holds two incompatible views doesn't tell you which of the two is right.

I think of "hypocrisy" as specifically referring to an inconsistency between words and actions.

Put simply: inconsistency between words and actions is no big deal. Why should your best estimate about good strategies be anchored to what you're already doing? The anti-hypocrisy norm seems to implicitly assume we're already perfect; it leaves no room for people who are in the process of trying to improve. We *know* people have akrasia. Also, akrasia isn't necessarily the only reason why actions may differ from words. It's important to be able to think and talk about better ways of doing things without necessarily changing courses at the drop of a hat. This is especially true in group coordination situations. Scott Alexander argues that anti-hypocrisy norms for journalists prevent them from suggesting improvements to society.

Flinching away from giving advice you don't yourself follow is accompanied by a knee-jerk reaction which discounts advice from others if we realize that the advice was hypocritical. Even though I've been trying to root out this mental habit for a long time,

I still catch myself updating away from advice when I realize that the person who gave it doesn't follow it. It *feels* relevant; decisive, even. Upon examination, though, it isn't.

One counter-argument is: if the hypocrite lacks experience with what they preach, they likely don't know what they're talking about. There may be obstacles to following their advice which they simply haven't experienced.

This may be relevant, but if so, I emphasize that it should be considered *in and of itself,* not as part of an anti-hypocrisy flinch. Why? Because *personal experience is not the only way to gain knowledge.* The anti-hypocrisy norm is founded in part on a non-Bayesian model of knowledge which emphasises personal experience and vivid stories from close in your social network. We should instead asses another person's beliefs on merits.

It's also largely contradicted by Beware Other-Optimizing. The advice-giver successfully following the advice themselves is *not* much evidence in favor of the advice. So, checking for hypocrisy is not a very good test of advice quality.

Anti-hypocrisy norms *do* provide a safeguard against using knowledge of cognitive biases to become more effective at motivated arguing, *if* you can successfully stop yourself from calling out biases in others when you haven't conquered them yourself. However: this, too, is a very weak heuristic. Conquering a bias in yourself should not give you free license to use that particular bias as a claim against others. Nor should *failing* to conquer a bias stop you from trying to *help* others conquer it.

One reason for the anti-hypocrisy norm which I *do* find more concerning is that an inconsistency between words and actions is a good indicator of dishonesty, so a norm against it may be a significant safeguard against liars. Again, though, the heuristic seems over-emphasized. There are discrepancies between words and actions which suggest that someone is lying, and there are those which don't. What I see isn't people considering whether hypocritical advice is coming from liars. What I see is an unthinking knee-jerk reaction which discounts hypocritical advice.

Putting your money where your mouth is, overcoming akrasia, doing the best thing you can figure out how to do, seeking to understand and honestly state the motives behind your actions: there are all things which we value, which point toward minimizing the discrepancy between words and actions. Anti-hypocrisy norms don't help us work toward these things, however. If anything, they Goodhart on minimizing the discrepancy, by encouraging us to align words with actions in cases where it prevents maximal honesty and truth-seeking.

So, what do you think? Have I failed to Chesterton-fence this one? ***Why do people flinch away from hypocrisy*?** Is hypocrisy bad? Am I failing to see something? Is my hufflepuff cynicism blocking me from seeing the advantages of holding people to their words / holding people to higher standards?

Hat tip: The seed of this post was planted long ago by my friend Thomas Kroll, who once said that the one idea he would have people remember him for was that hypocrisy isn't bad.

ETA:

Thanks to a number of discussions in the comments, several important distinctions have been called out, which I was conflating.

Most importantly, I was ignoring the question of norms and status claims, and arguing as if all advice is about the epistemic question of whether it would be useful to do some thing. Hypocrisy about norms is problematic, because it is easy for a hypocrite to set up unfair (self-serving) norms which can't be called out as unfair/unjust in any other way than to call out the hypocrisy itself (due in part to limitations on what can be publicly called out -- other objections to the rule may simply not be clear/defensible). There's a case to be made that this in itself justifies the anti-hypocrisy flinch, because it is too difficult in practice to detangle norm-setting claims from non-norm-setting ones.

I also conflated my point with whether "hypocrisy is bad". It should be noted that hypocrisy absolutely implies that *something* unfortunate is going on: either the hypocrite is lying, or mistaken, or taking suboptimal actions which they have enough information to improve. The issue, metaphorically, is that this is just like saying that if you're travelling, then you must either not be where you want to be or headed in the wrong direction. Avoiding travel doesn't necessarily make you where you need to be; avoiding hypocrisy doesn't necessarily help you say the right words or take the right actions.

Other important distinctions:

Joachim Bartosik points out that I'm not clear on which of these questions I'm asking:

1. Are there good reasons to be suspicious of advice that advice giver doesn't follow themselves?

2. Is there a good reason to support social norms against hypocrisy?

3. Are there good reasons to avoid giving advice that I don't follow myself?

To this, I might add "are there good reasons to avoid taking actions which go against something I said?"

Along similar lines, my discussion with Said resulted in a number of possible rules around hypocrisy which we might follow or question:

1. hypocrisy => what hypocrite said is wrong

2. hypocrisy => what hypocrite said is not evidence

3. hypocrisy => hypocrite is blameworthy

4. hypocrisy => hypocrite is to be viewed with high suspicion, on priors

5. hypocrisy => EITHER what hypocrite said was wrong, OR they are blameworthy

6. hypocrisy => hypocrite is to be treated as a hostile agent for the purpose of evaluating their words in this context

7. hypocrisy => call out hypocrisy

8. it's a conversation about norms, and the hypocrite is making an implicit status claim with their words => hypocrisy should counter-indicate the norm fairly strongly and also detract from the status of the hypocrite

My primary intention in the post was to argue against #1 and #2. I still think they are bad rules; hypocrisy provides a small amount of evidence against a claim, but I think it's highly over-emphasized in practice. I also disagree with all of the others, except for #8, which I think is true because of the earlier-mentioned point about hypocrisy and norms.