Tensions in Truthseeking

Tensions in Truthseeking

Epistemic Effort: I've thought about this for several weeks and discussed with several people who different viewpoints. Still only moderately *confident though.*

So, I notice that people involved with the rationalsphere have three major classes of motivations:

Truthseeking (how to think clearly and understand the world)

Human/Personal (how to improve your life and that of your friends/family)

Impact (how to change/improve the the world at large)

All three motivations can involve rationality. Many people who end up involved care about all three areas to some degree, and have at least some interest in both epistemic and instrumental rationality. And at least within the rationalsphere, the Personal and the Impact motivations are generally rooted in Truth.

But people vary in whether these motivations are terminal, or instrumental. They also have different intuitions about which are most important - or about how to pursue a given goal. This sometimes results in confusion, annoyance, distrust, and exasperated people working at cross purposes.

Terminal vs Instrumental Truth

For some, truthseeking is important because the world is confusing. Whether you're focused on your personal life or on changing the world, there's a lot of ways you might screw up because something seems right but doesn't work or has a lot of negative externalities. It's necessary to do research, to think clearly, and to be constantly on the lookout for new facts that might weigh on your decisions.

For others, truth-seeking seems more like a fundamental part of who they are. Even if it didn't seem necessary, they'd do it anyway because because it just seems like the right thing to do.

I think there's a couple layers of conflict here. The first is that instrumental-truthseekers tend to have an intuition that lots of other things matter as much or more than truth.

> It's more important to be able to launch a startup confidently than to have an accurate perception of its chance of success. It's important not to immediately criticize people because that disincentivizes them from trying new things. Interpersonal relationships seem to need 5x as many compliments as criticisms to flourish. It's important to be able to run a successful marketing campaign, and even if you're trying to run an honest marketing campaign, comprehensive honesty requires effort that might be better spent on actually building your product or something.
>
> It may even be necessary to schmooze with people you don't respect because they have power and you need their help if you're going to make a dent in the universe.

Then, there are people (who tend to be terminal-truthseekers, although not always), who counter:

> Earnest truthseeking is incredibly rare and precious. In almost every movement, it gets sacrificed on the altar of practicality and in-group solidarity. Can't we just once have a movement where truthseeking is the primary thing that never gets sacrificed?
>
> This doesn't just seem worth trying for the novelty of it: the world seems so deeply confusing, the problems it faces seem so immense and so entwined with tribal-politics that distort the truth, that we probably need a movement of impact-oriented truthseekers who never compromise their intellectual integrity no matter what.

I find this argument fairly compelling (at least for a deeper delve into the concept). But what's interesting is that even if it's an overriding concern, *it doesn't really clarify what to do next.*

The Trouble With Truthseeking While Human

On the one hand, social reality is a thing.

Most cultures involve social pressure to cheer for your ingroup's ideas, to refrain from criticizing your authority figures. They often involve social pressure to say "no, that outfit doesn't make you look fat" whether or not that's true. They often involve having overt, stated goals for an organization (lofty and moral sounding) that seem at odds with what the organization ends up doing - and if you try to mention the disconnect between what people are saying and what they are doing, they get upset and angry at you challenging their self-conception.

The pressure to conform to social reality is both powerful and subtle. Even if you're trying to just think clearly, privately for yourself, you may find your eyes, ears and brain conforming to social reality anyway - an instinctive impulse to earnestly believe the things that are in your best interest, so you peers never notice that you are doubting the tribe. I have noticed myself doing this, and it is scary.

In the face of that pressure, many people in the rationality community (and similar groups of contrarians), have come to prize criticism, and willingness to be rude. And beyond that - the ability to see through social reality, to actively distance themselves from it to reduce its power over them (or simply due to aesthetic disgust).

I earnestly believe those are important things to be able to do especially in the context of a truthseeking community. But I see many people's attempts as akin Stage 2 of Sarah Constantin's "Hierarchy of Requests":

> Let's say you're exhausted; you want to excuse yourself from the group and take a nap.
>
> In stage 1, you don't dare ask. Or you don't understand why you feel shitty, you don't recognize it as fatigue. You just get more and more upset until you collapse in a heap. In stage 2, you rudely interrupt people in the middle of something important and announce that you're tired and you're leaving. In stage 3, you find a convenient moment, apologize for cutting things short, but explain that you've got

to get some rest. In stage 4, you manage to subtly wrap things up so you can get some rest, without making anyone feel rushed.

It's better to be able to rudely criticize than not at all. And for some people, a culture of biting, witty criticism is fun and maybe an important end in-and-of-itself. (Or: a culture of being able to talk about things normally considered taboo can be freeing and be valuable both for the insight and for the human need for freedom/agency). I've gotten value out of both of those sorts of cultures.

But if you're unable to challenge social reality *without* brusquely confronting it - or if that is the manner in which you usually do - I think there's a lot of net-truth you're leaving on the table.

There are people who don't feel safe sharing things when they fear brusque criticism. I think Robby Bensinger summarized the issue compactly: "My own experience is that 'sharp culture' makes it more OK to be open about certain things (e.g., anger, disgust, power disparities, disagreements), but less OK to be open about other things (e.g., weakness, pain, fear, loneliness, things that are true but not funny or provocative or badass)."

Brusque confrontation leads to people buckling down to defend their initial positions because they feel under attack. This can mean less truth gets uncovered and shared.

Collaboration vs Criticism For The Sake Of It

The job of the critic is much easier than the job of the builder.

I think that there's a deeper level productive discussion to be had when people have a shared sense that they are *collaboratively building something*, as opposed to a dynamic where "one person posts an idea, and then other people post criticisms that tear it down and hopefully the idea is strong enough to survive." Criticism is an important part of the building process, but I (personally) feel a palpable difference when criticized by someone who shows a clear interest in *making sure that something good happens as a result of the conversation.*

Help Brainstorm Solutions - If you think someone's goals are good but their approach is wrong, you can put some effort into coming with alternate approaches that you think are more likely to work. If you can't think of any ways to make it work (and it seems like it's better to do nothing than to try something that'll make a situation worse), maybe you can at least talk about some other approaches you considered but still feel inadequate.

Active Listening / Ideological Turning Tests - If you disagree with a person's *goals,* you can try to understand why they have those goals, and showcase to them that you at least get where they're coming from. In my experience people are more willing to listen when they feel they're being listened to.

Accompanying criticism with brainstorming and active listening acts as a costly signal, that helps create an atmosphere where it's a) worth putting in the effort to develop new ideas, and b) easier to realize (and admit) that you're wrong.

Truth As Impact

If you constantly water down your truth to make it palatable for the masses, you'll lose the spark that made that truth valuable. There are downsides to being constantly guarded, worried that a misstep could ruin you. Jeff Kaufman writes:

> There are a lot of benefits to unguarded communication: you can move faster, you can open up your tentative thoughts to friendly consideration and criticism, you don't have the mental or process overhead of needing to get every statement as perfect as possible. You might say something that you don't mean to, but in a friendly environment you can correct yourself or accept someone else's correction.
>
> Despite these benefits, it seems to me that things generally move in the more guarded direction, at least publicly, as they become more successful.

Daniel in the comments notes:

> And I think one cost of guardedness that seems missing from the post is that guardedness can bias thinking in favor of more easily palatable and defensible ideas, both in discussions between people as well as one's own thoughts.
>
> Unfortunately, I think it's a natural consequence of growing large and powerful enough to actually affect the big picture: If you're communicating, not to a few trusted friends but to the entire world, then a verbal misstep will *not* be something you can easily correct, and the cost may grow from "a few minutes of clarification" to "millions of dollars worth of value lost".

I'm not sure how to handle that paradox (Less Wrong is hardly the first group of people to note that PR-speak turns dull and lifeless as organizations grow larger and more established - it seems like an unsolved problem).

But there's a difference between watering things down for the masses and speaking guardedly... and learning to communicate in a way that uses other people's language, that starts from their starting point.

If you want your clear insights to *matter anywhere outside a narrow cluster of contrarians*, then at some point you need to figure out how to communicate them so that the rest of the world will listen. Friends who are less contrarian. Customers. Political bodies. The Board of Directors at the company you've taken public.

How to approach this depends on the situation. In some cases, there's a specific bit of information you want people to have, and if you can successfully communicate that bit then you're won. In other cases, the one bit doesn't doing anything in isolation - it only matters if you successfully get people to think clearly about a complex set of ideas.

Consider Reversing All Advice You Hear

One problem writing this is that there's a lot of people here, with different goals, methods and styles of communication. Some of them could probably use advice more like:

- "Learn to criticize more kindly/constructively."
- "Communicate more clearly."
- "Keep in mind the layers of signals you're sending when you try to state 1st-order-true-things."

And some some could probably use advice more like:

- "Make sure in your efforts to avoid conflict you don't gloss over important disagreements."
- "Don't get so wrapped up in what other people think that you lose the ability to think clearly for yourself."

I started writing this post four months ago, as part of the Hufflepuff Sequence. Since then, I've become much less certain about which elements here are most important to emphasize, and what the risks are of communicating half-baked versions of each of those ideas to different sorts of people.

But I do still believe that the end-goal for a "true" truth-oriented conversation will need to bear all these elements in mind, one way or another.

The Archipelago Model of Community Standards

Epistemic Status: My best guess. I don't know if this will work but it seems like the obvious experiment to try more of.

Epistemic Effort: Spent several months thinking casually, 25ish minutes consolidating earlier memories and concerns, and maybe 10ish minutes thinking about potential predictions. See comment.

Building off:

- Open Problems in Group Rationality [Conor Moreton]
- Archipelago and Atomic Commutarianism [Scott Alexander]

Claim 1 - If you are dissatisfied with the norms/standards in a vaguely defined community, a good first step is to **refactor that community into sub-groups with clearly defined goals and leadership.**

Claim 2 - People have different goals, and you may be wrong about what norms are important even given a certain goal. So, also consider ***proactively cooperating* with other people forming alternate subgroups** out of the same parent group, with the goal of learning from each other.

Refactoring Into Subcommunities

Building groups that accomplish anything is hard. Building groups that prioritize independent thinking to solve novel problems is harder. But when faced with a hard problem, a useful technique is to refactor it into something simpler.

In "Open Problems in Group Rationality", Conor lists several common tensions. I include them here for reference (although *any* combination of difficult group rationality problems would suffice to motivate this post).

1. Buy-in and retention.
2. Defection and discontent.
3. Safety versus standards.
4. Productivity versus relevance.
5. Sovereignty versus cooperation.
6. Moloch and the problem of distributed moral action.

These problems *don't go away* when you have clearly defined goals. A corporation with a clearcut mission and strategy (i.e maximize profit by selling widgets) still has to navigate the balance of "hold their employees to a high standards to increase performance" and "make sure employees feel safe enough to do good work without getting wracked with anxiety" (or, just quit).

Such a corporation might make different tradeoffs in different situations - if there's a labor surplus, they might be less worried about employees quitting because they can just find more. If the job involves creative knowledge work, anxiety might have greater costs to productivity. Or maybe they're *not* just profit-maximizing: maybe the CEO cares about employee mental health for its own sake.

But well defined goals, with leaders who can enforce them, at least makes it *possible* to figure out what tradeoffs to make and actually make them.

Whereas if you live in a loosely defined community where people show up and leave whenever they want, *and nobody can even precisely agree on what the community is,* you'll have a lot more trouble.

People who care a lot about, say, personal sovereighty, will constantly push for norms that maximize freedom. People that care about cooperation will push for norms encouraging everyone to work harder and be more reliabl at personal freedom's expense.

Maybe one group can win - possibly by persuading everyone they are right, or simply by being more numerous.

But,

A) You probably can't win every cultural battle.

B) Even if you could, you'd spend a lot of time and energy fighting that might be better spent actually accomplishing whatever these norms are actually *for.*

So if you can manage to *avoid* infighting while still accomplishing your goals, all things being equal that's preferable.

Considering Archipelago

Once this thought occured to me, I was immediately reminded of Scott Alexander's Archipelago concept. A quick recap:

Imagine a bunch of factions fighting for political control over a country. They've agreed upon the strict principle of harm (no physically hurting or stealing from each other). But they still disagree on things like "does pornography harm people", "do cigarette ads harm people", "does homosexuality harm the institution of marriage which in turn harms people?", "does soda harm people", etc.

And this is bad not just because everyone wastes all this time fighting over norms, but because the nature of their disagreement incentivizes them to fight over *what harm even* ***is****.*

And this in turn incentivizes them to fight over both definitions of words (distracting and time-wasting) and *what counts as evidence or good reasoning* through a politically motivated lens. (Which makes it harder to *ever* use evidence and reasoning to resolve issues, even uncontroversial ones)

Then...

Imagine someone discovers an archipelago of empty islands. And instead of continuing to fight, the people who want to live in Sciencetopia go off to found an island-state based on ideal scientific processes, and the people who want to live in Libertopia go off and found a society based on the strict principle of harm, and the people who want to live in Christiantopia go found a fundamentalist Christian commune.

They agree on an overarching set of rules, paying some taxes to a central authority that handles things like "dumping pollutants into the oceans/air that would affect other islands" and "making sure children are well educated enough to have the opportunity to understand why they might consider moving to other islands."

Practical Applications

There's a bunch of reasons the Archipelago concept doesn't work as well in practice. There are no magical empty islands we can just take over. Leaving a place if you're unhappy is harder than it sounds. Resolving the "think of the children" issue will be very contentious.

But, we don't need perfect-idealized-archipelago to make use of the general concept. We don't even need a broad critical mass of change.

You, personally, could just do something with it, right now.

If you have an event you're running, or an online space that you control, or an organization you run, you can set the norms. Rather than opting-by-default into the generic average norms of your peers, you can say "This is a space specifically for X. If you want to participate, you will need to hold yourself to Y particular standard."

Some features and considerations:

You Can Test More Interesting Ideas. If a hundred people have to agree on something, you'll only get to try things that you can can 50+ people on board with (due to crowd inertia, regardless of whether you have a formal democracy)

But maybe you can get 10 people to try a more extreme experiment. (And if you share knowledge, both about experiments that work and ones that don't, you can build the overall body of community-knowledge in your social world)

I would rather have a world where 100 people try 10 different experiments, even if I *disagree* with most of those experiments and wouldn't want to participate myself.

You Can Simplify the Problem and Isolate Experimental Variables. "Good" science tests a single variable at the time so you can learn more about what-causes-what.

In practice, if you're building an organization, you may not have time to do "proper science" - you may need to get a group working ASAP, and you may need to test a few ideas at once to have a chance at success.

But, all things being equal it's still convenient to isolate factors as much as possible. One benefit to refactoring a community into smaller pieces is you can pick more specific goals. Instead of reinventing every single wheel at once, pick a few specific axes you're trying to learn about.

This will both make the problem easier, as well as make it easier to *learn from.*

You Can 'Timeshare Islands'. Maybe you don't have an entire space that you can control. But maybe you and some other friends have a shared space. (Say, a weekly meetup).

Instead of having the meetup be a generic thing catering to the average common denominator of members, you can collectively agree to use it for experiments (at least sometimes). Make it easier for one person to say 'Okay, this week I'd like to run an activity that'll require different norms than we're used to. Please come prepared for things to be a bit different.'

This comes with some complications - one of the benefits of a recurring event is people roughly know what to expect, so it may not be good to do this all the time. But generally, giving the person running a given event the authority to try some different norms out can get you some of the benefits of the Archipelago concept.

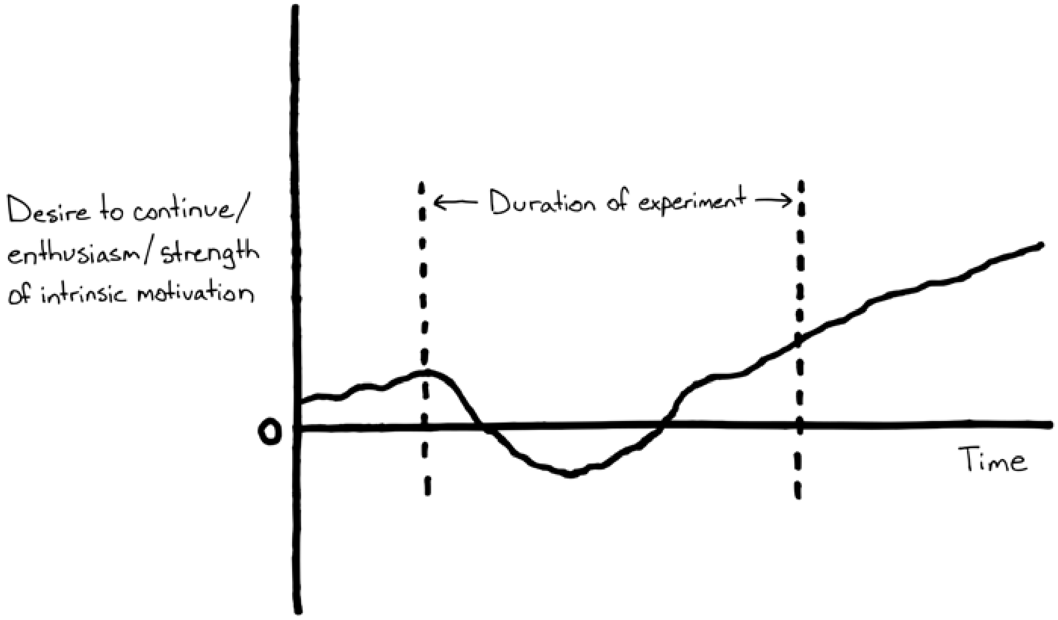
You Can Start With Just *One* Meetup

Viliam in the comments made a note I wanted to include here:

> It is important to notice that the "island" doesn't have to be fully built from start. "Let's start a new subgroup" sounds scary; too much responsibility and possibly not enough status. "Let's have *one meeting* where we try the norm X and see how it works" sounds much easier; and if it works, people would be more willing to have another meeting like that, possibly leading to the creation of a new community.

Making It Through the 'Unpleasant Valley' of *Group* Experimentation.

I think this graph was underappreciated in its original post. When people try new things (a new diet or exercise program, studying a new skill, etc), the new thing involves effort and challenges that in some ways make it seem worse than whatever their default behavior was.



Some experiments are just duds. But oftentimes it *feels* like it'll turn out to be a dud, when you're in the Unpleasant Valley, and in fact you just haven't stuck with it long enough for it to bear fruit.

This is hard enough for *solo* experiments. For group experiments, where not just one but *many* people must all try a thing at once and *get good at it*, all it takes is a little defection to spiral into a mass exodus.

Refactoring communities into smaller groups with clear subgoals can make it *possible* for a group to make it through the Valley of Unpleasantness together.

Overlapping Social Spheres

Sharing Islands and Cross Pollination

In the end, I don't think "Islands" is quite the right metaphor here. One of the things that makes *social* archipelago different from the canonical example is that the islands overlap. People may be a member of multiple groups and sub-groups.

A benefit of this is cross pollination - it's easier to share information and grow if you have people who exist in multiple subcultures (sub-subcultures?) and can translate ideas between them.

How much benefit this yields depends on how mindfully people are approaching the concept, and how much of their ideas they are sharing (making both the object-level-idea and the underlying reasons accessible to others).

This post is primarily intended as reference - I have more specific ideas on what kinds of communities *I* want to participate in, and thoughts on "underexplored social niches" that I think others might consider experimenting with. Some of those thoughts will be on the LessWrong front page, others on my private profile or the Meta section.

But meanwhile, I hope to see more groups of people in my filter bubble self organizing, carving out spaces to try novel concepts.

Musings on Double Crux (and "Productive Disagreement")

Epistemic Status: Thinking out loud, not necessarily endorsed, more of a brainstorm and hopefully discussion-prompt.

Double Crux has been making the rounds lately (mostly on Facebook but I hope for this to change). It seems like the technique has failed to take root as well as it should. What's up with that?

(If you aren't yet familiar with Double Crux I recommend checking out Duncan's post on it in full. There's a lot of nuance that might be missed with a simple description.)

Observations So Far

- Double Crux hasn't percolated *beyond* circles directly adjacent to CFAR (it seems to be learned mostly be word of mouth). This might be evidence that it's too confusing or nuanced a concept to teach without word of mouth and lots of examples. It might be evidence that we have not yet taught it very well.
- "Double Crux" seems to refer to two things: the specific action of "finding the crux(es) you both agree the debate hinges on" and "the overall pattern of behavior surrounding using Official Doublecrux Technique". (I'll be using the phrase "productive disagreement" to refer to the second, broader usage)

Double Crux seems hard to practice, for a few reasons.

Filtering Effects

- In local meetups where rationality-folk attempt to practice productive disagreement on purpose, they often have trouble finding things to disagree about. Instead they either:
  - are already filtered to have similar beliefs,
  - quickly realize their beliefs shouldn't be that strong (i.e. they disagree on Open Borders, but soon as they start talking they admit that neither of them really have that strong an opinion)
  - they have wildly different intuitions about deep moral sentiments that are hard to make headway on in a reasonable amount of time - often untethered to anything empirical. (i.e. what's more important? Preventing suffering? Material Freedom? Accomplishing interesting things?)

Insufficient Shared Trust

- Meanwhile in many online spaces, people disagree all the time. And even if they're both nominally rationalists, they have an (arguably justified) distrust of people on the internet who don't seem to be arguing in good faith. So there isn't enough foundation to do a productive disagreement at all.
- One failure mode of Double Crux is when people disagree on what frame to even be using to evaluate truth, in which case the debate recurses all the way to the level of basic epistemology. It often doesn't seem to be worth the effort to resolve that.
- Perhaps most frustratingly: it seems to me that there are many longstanding disagreements between *people who should totally be able to communicate*

clearly, update rationally, and make useful progress together, and those disagreements don't go away, people just eventually start ignoring each other or leaving the dispute as unresolved. (An example I feel safe bringing up publicly is the argument between Hanson and Yudkowsky, although this may be a case of the 'what frame are we even using' issue above.)

That last point is one of the biggest motivators of this post. If the people I most respect can't productively disagree in a way that leads to *clear progress, recognizable from both sides*, then what is the rationality community even doing? (Whether you consider the primary goal to "raise the sanity waterline" or "build a small intellectual community that can solve particular hard problems", this bodes poorly).

Possible Pre-Requisites for Progress

There's a large number of sub-skills you need to productively disagree. To have *public norms* surrounding disagreement, you not only need individuals to have those skills - they need to trust that each other have those skills as well.

Here's a rough list of those skills. (Note: this is long, and it's less important that you read the whole list than that the list is long, which is why Double Cruxing is hard)

- **Background beliefs** (listed in Duncan's original post)
    - Epistemic humility ("I could be the wrong person here")
    - Good Faith ("I trust the other person to be believing things that make sense to them, which I'd have ended up believing if I were exposed to the same stimuli, and that they are generally trying to find the the truth")
    - Confidence in the existence of objective truth
    - Curiosity / Desire to uncover truth
- **Building-Block and Meta Skills**
- (Necessary or at least very helpful to learn everything else)
    - Ability to gain habits (see Trigger Action Plans, Reflex/Routines, Habits 101)
    - Ability to introspect and notice your internal states (Focusing and Noticing can help)
    - Ability to induce a mental state or reframe
    - Habit of gaining habits
- **Notice you are in a failure mode, and step out.** Examples:
    - You are fighting to make sure an side/argument wins
    - You are fighting to make another side/argument lose (potentially jumping on something that seems *allied* to something/someone you consider bad/dangerous)
    - You are incentivized to believe something, or not to notice something, because of social or financial rewards,
    - You're incentivized not to notice something or think it's important because it'd be physically inconvenient/annoying
    - You are offended/angered/defensive/agitated
    - You're afraid you'll lose something important if you lose a belief (possibly 'bucket errors')
    - You're rounding a person's statement off to the nearest stereotype instead of trying to actually understand and response to what they're saying
    - You're arguing about definitions of words instead of ideas
    - Notice "freudian slip" ish things that hint that you're thinking about something in an unhelpful way. (for example, while writing this, I typed out

"your opponent" to refer to the person you're Double Cruxing with, which is a holdover from treating it like an adversarial debate)

(The "Step Out" part can be pretty hard and would be a long series of blogposts, but hopefully this at least gets across the ideas to shoot for)

- **Social Skills** (i.e. not feeding into negative spirals, noticing what emotional state or patterns other people are in [*without* accidentaly rounding them off to a stereotype])
  - Ability to tactfully disagree in a way that arouses curiosity rather than defensiveness
  - Leaving your colleague a line of retreat (i.e. not making them lose face if they change their mind)
  - Socially reward people who change their mind (in general, frequently, so that your colleague trusts that you'll do so for them)
  - Ability to listen (in a way that makes someone feel listened to) so they feel like they got to actually talk, which makes them inclined to listen as well
  - Ability to notice if someone else seems to be in one of the above failure modes (and then, ability to point it out gently)
  - Cultivate empathy and curiosity about other people so the other social skills come more naturally, and so that even if you don't expect them to be right, you can see them as helpful to at least understand their reasoning (fleshing out your model of how other people might think)
  - Ability to communicate in (and to listen to) a variety of styles of conversation, "code switching", learning another person's jargon or explaining yours without getting frustrated
  - Habit asking clarifying questions, that help your partner find the Crux of their beliefs.
- Actually Thinking About Things
  - Understanding when and how to apply math, statistics, etc
  - Practice thinking causally
  - Practice various creativity related things that help you brainstorm ideas, notice implications of things, etc
  - Operationalize vague beliefs into concrete predictions
- **Actually Changing Your Mind**
  - Notice when you are confused or surprised and treat this as a red flag that something about your models is wrong (either you have the wrong model or no model)
  - Ability to identify what the actual Crux of your beliefs are.
  - Ability to track bits of small bits of evidence that are accumulating. If enough bits of evidence have accumulated that you should at least be taking an idea *seriously* (even if not changing your mind yet), go through motions of thinking through what the implications WOULD be, to help future updates happen more easily.
  - If enough evidence has accumulated that you should change your mind about a thing... like, actually do that. See the list of failure modes above that may prevent this. (That said, if you have a vague nagging sense that something isn't right even if you can't articulate it, try to focus on that and flesh it out rather than trying to steamroll over it)
  - Explore Implications: When you change your mind on a thing, don't just acknowledge, actually think about what other concepts in your worldview should change. Do this
    - because it *should* have other implications, and it's useful to know what they are....

        - because it'll help you actually retain the update (instead of letting it slide away when it becomes socially/politically/emotionally/physically inconvenient to believe it, or just forgetting)
    - If you notice your emotions are not in line with what you now believe the truth to be (in a system-2 level), figure out why that is.
- **Noticing Disagreement and Confusion, and then *putting in the work* to resolve it**
- If you have all the above skills, and your partner does too, and you both trust that this is the case, you can still fail to make progress if you don't actually follow up, and schedule the time to talk through the issues thoroughly. For deep disagreement this can take years. It may or may not be worth it. But if there are longstanding disagreements that continuously cause strife, it may be worthwhile.

Building Towards Shared Norms

When smart, insightful people disagree, at least one of them is doing something wrong, and it seems like we should be trying harder to notice and resolve it.

A rough sketch of a norm I'd like to see.

Trigger: You've gotten into a heated dispute where at least one person feels the other is arguing in bad faith (especially in public/online settings)

Action: Before arguing further:

- stop to figure out if the argument is even worth it
- if so, each person runs through some basic checks (i.e. "am *I* being overly tribal/emotional?)
- instead of continuing to argue in public where there's a lot more pressure to not lose face, or steer social norms, they continue the discussion privately, in whatever the most human-centric way is practical.
- they talk until *at least* they succeed at Step 1 Double Crux (i.e. agree on where they disagree, and *hopefully* figure out a possible empirical test for it). Ideally, they also come to as much agreement as they can.
- Regardless of how far they get, they write up a short post (maybe just a paragraph, maybe longer depending on context) on what they *did* end up agreeing on or figuring out. (The post should be something they both sign off on)

Common vs Expert Jargon

tldr: Jargon always has a complexity cost, but you can put effort into making a concept more accessible, and it's especially valuable to put that effort in for terms that you'd like to be used by layfolk, or that you expect to be used a lot in spaces where you expect lots of layfolk to be reading/participating.

I. Lessons from Game Design

Magic the Gathering deals a lot with complexity. Each year, new abilities and rules are added to the game. This gives experienced players the chance to constantly discover new things, but it comes with some issues.

First, it makes the game harder for new players (the game kept growing more complex over time, raising the amount of information a new player had to process at once)

And second, even for experienced players: each instance of complexity is a cost. Players (both new and old) can only handle so much, and some forms of complexity are less fun than others. (For example, forcing players to do a lot of book-keeping, rather than letting them make interesting strategic decisions)

Six years ago, their creative director wrote about a new paradigm of Magic design. One of their solutions was to pay careful attention to how they spent complexity points in ways that affected new players.

Three examples:

1. Common Cards

In Magic, when you buy a new pack, 11 cards are "common", 3 are "uncommon" and one is "rare". Experienced players buy lots of cards and can have access to lots of rares, but new players generally just buy a few cards, so most of their cards are common. Therefore, the complexity of the cards at common determines how much complexity newcomers have to deal with.

2. Keywords

One way to reduce "effective complexity" is to bundle concepts together in a keyword. Instead of saying "this creature deals damage to each of the creatures blocking it and then deals the remainder of its damage to the player", it just says "Trample". There's an initial cost of learning what Trample means, but afterwards, every time you see the word "Trample" on a creature it works the same way.

Trample has some neat things going for it: it sounds evocative, and gets to build off of existing ideas in your brain. You already know what a big animal looks like. You can imagine a small creature getting in the way of the elephant, and it slowing the elephant down slightly but not really stopping it, and the elephant continuing on, trampling over it, and then going on to attack some bigger target.

This imagery is helpful for intuiting what the rules mean, even if the wording is somewhat confusing.

The problem comes when you introduce too many keywords at once. It gets overwhelming. Which brings to a final concept:

3. Evergreen keywords

Every 3 months, new magic cards are released to keep things fresh. New keywords are introduced (usually 3-5).

But there are some keywords (like Trample) that are *always* in season. There are about 16 evergreen keywords. Many of them are pretty intuitive (such as flying creatures only be able to be blocked by other flying creatures) so they aren't hard to learn.

A new player has an implicit goal of "learn all the evergreen keywords", which is a manageable task.

II. Building a high level conversation

I think some of this applies to the rationalsphere, where a lot of important concepts have been built up, or, combined together from neighboring disciplines. (See Anna Salmon's Single Conversational Locus)

Jargon is *useful*. They let you summarize a complex concept in a single word, and then have deeper conversations where each word packs a lot more meaning.

I have a lot of thoughts about how to do jargon *right*, which are beyond the scope of this post. But to summarize, I think good jargon:

- encapusulates an idea that's important to build off of
- lets you distinguish between *similar* concepts that have importantly-different-nuances. (viral infection vs bacterial infection)
- provides some context clues that help you learn it (the way Trample does), while...
- ...not *also* resulting in people confusing what it means (a bad example perhaps being "negative reinforcement", which is not actually the same thing as "punishment")

Some considerations:

1) Sometimes you want a 101 space where you're either introducing ideas to a broader audience. Sometimes you want a 201 space where you're building on those ideas (either helping somewhat-less-newcomers build up a more advanced understanding, or literally developing new content at the cutting edge)

2) Different venues of conversation can have both different expectations of who-is-participating, and different social norms of what kind of participation is encouraged. (i.e an academic journal, a semi-formal internet forum, a facebook post)

3) Some concepts are pretty standalone: layfolk can learn them and use them immediately without having to fit them into a big edifice of theory

4) Furthermore, some concepts make good "gateway" terminology. They're useful standalone, but then they open up a world of ideas to you that you can then further explore.

So my thought is basically: if you are developing jargon, pay extra attention to whether this is Common or Expert Level jargon. There's not a clear dividing line between them, but roughly:

Common Jargon means you're expecting it to be a useful enough idea for layfolk to use regularly (or, you'd like to be able to have conversations with layfolk, or write popularization articles, that rely on the term already percolating into the mainstream, or, use it as a gateway term)

Consequently, it's much more important to put a lot of effort into choosing a term that:

- resonates easily, is memorable...
- ...but avoids people latching onto the wrong aspect of it and misinterpreting it
- doesn't sound like a weird insider term...
- ... but maybe ideally hints at a broader ecosystem of ideas

Expert Jargon is only really useful if you're buying into a broader ecosystem of ideas that build on each other. Accessibility and avoiding misunderstanding is still important if possible but being precise and build-on-able is more valuable.

Further Reading

This post was inspired by and builds upon:

Complexity is bad (Zvi Mowshowitz)
The Purple Sparkly Ball Thing (Malcolm Ocean)
New World Order (Mark Rosewater)

Writing That Provokes Comments

Epistemic Effort: Thought about it for a year. Solicited feedback. Checked my last few posts' comment count to make sure I wasn't *obviously* *wrong.*

A thing that happens to me, and perhaps to you:

Someone writes a beautiful essay that I agree with, that sheds new light on something important.

I don't have anything really to say about it. I don't want to just say "I agree!". So instead of commenting, I give it an upvote and move on.

This feels bad for a few reasons:

- I like commenting.
- I like *getting* comments when I write things that (I hope!) are insightful, beautiful and true. It's a stronger signal that people care.
- Comments correlate with something *staying in the public sphere of attention*. A highly upvoted post eventually fades behind newer upvoted posts. But a post with lots of comments keeps people paying attention (with new people constantly checking in to see what the hubbub is about)
- I don't trust (as a reader or a writer) that people who read a post, give it an upvote, and move on, are really *learning* anything. I think that talking through an a new concept and figuring out how to apply is where much of the learning happens.

I've been impressed with how much quality writing has been going on on LW2.0 so far. There has been *some* but not *as much* commenting as I'd like.

I've gotten a sense of what inspires interesting, meaty discussion.

Unfortunately, most of it seems... kinda bad?

Things That Get People To Comment

1. Be Wrong - It has been said: if google fails you, the fastest way to get a question answered is to post a *wrong answer* on reddit. This will result in a lot of flood of people explaining things to you.

2. Be Controversial - Even better, post something that *some* people think are wrong. Then you get a bunch of people commenting to correct you, and then other people who disagree correcting *them!* The arguments perpetuate themselves from there. You won't even have to do any commenting work yourself to keep it going!

[*BTW, these are observations, not recommendations. This list is optimized to answer the question* "*what causes comments*" *not* "*how to make the world better.*"*]*

3. Write About Things People Feel Qualified to Have Opinions On - If you write a post on machine learning, and post it somewhere where nobody really understands machine learning, it doesn't matter if you're wrong or controversial! Nobody will understand enough to care, or feel confident enough to argue. Some considerations:

- It's not necessary for people to be qualified. They just need to feel like they are.
- If you write more informally (or in informal forums), people feel more entitled to respond.
- You can either tailor your topic to an existing audience, or p*roactively try to get an* existing audience who understands your weird niche topic to read your post.

4. Invoke Social Reality - People pay more attention when you're talking about social norms, or about changing coalitions of people, or arguing that some people are Bad and Wrong. This is for two reasons:

- Social Reality is powerful and scary. A person's sense of social safety is one of the most important things to them. People like to know who is Bad and Wrong so that they can be on the other side. People like making sure that if social norms changing, they are changing in ways they understand and like (so that nobody later decides they are Bad and Wrong).
- Social Reality almost always has something confusing and dumb going on that needs fixing, that people think is worth thinking about.
- People *understand* Social Reality. Or, they think they do. (See #3)
- Social Reality is often controversial! (See #2)

5. Be So Inspiring That People Create Entire Fandoms of Your Work - This worked for Eliezer and arguably Scott. It can probably be broken down into smaller steps. It's pretty hard though. And a bunch of people *trying* but failing to do this can be annoying. (*I've* tried/failed to do this sometimes)

...

And then there's...

6. Leave People With An Unsolved Problem That They Care About - This is related to "they feel qualified to have opinions", with the followup step of "there is actual useful thinking they can contribute to, either to solve *your* problem, or to apply your idea to solve *their* problems."

Things I've Noticed Myself Doing

Since comments are socially validating, I've noticed a tendency for me to end up writing:

- Facebook posts, where people feel a lower barrier to entry. (If the shortform section of LessWrong were up, I might do that instead)
- Unfinished thoughts, where there's a good chance that I'm wrong about a few things (but not all things, and not wrong *on purpose to be provocative* which would feel skeezy*),* and where there's still an unsolved problem that people will feel qualified to help out figure out.
- Posts engaging with social norms (which people feel excited to weigh in on and/or afraid not to)
- Posts engaging with personal habits that people can easily apply to their own life.

This doesn't all seem *bad,* necessarily. But I've noticed other people that seem to be doing similar things. I've also noticed some people who tried to get people to talk about important things, and failed, and gradually resorted to writing more provocative things to get people to pay attention (which succeeded!).

It seems like a rationality community warped by those incentives isn't going to accomplish the things it needs to.

So, some open problems I'm thinking about, which maybe are relevant to you:

- I'd like feel **incentivized to research things I don't understand** as much (which I don't expect other people to understand as much either), to expand my (and our collective) domains of expertise.
- Insofar as people do end up writing the sorts of posts listed above, I think it'd be good if people **thought more consciously and carefully** about which tools they're employing. #6 at the very least seemed fine, and some of the others seem fine in some contexts.
- I'd like to **learn how to be a better commenter**, on posts that *don't* go out of their way to make it easy to comment. I have a sense that if I took the step of *actually stopping to think for a half-hour* about possible ramifications of a given post, I could probably think of something worth saying, and that it might get easier with time. (I've been thinking about that for the past week or two, but keep end up spending that time mostly writing my own posts, or engaging with other commenters who did more heavy lifting of initiating discussion)
- I'd like *people who have important things to say* to **be able to trust that people will listen**, without falling into an attentional arms race that leads inevitably to BuzzFeed. But right now *I* have trouble paying attention to things that are important but non-drama-laden, so I can't reasonably expect people to trust in that.

That's all I got for now.

Writing Down Conversations

Epistemic Status: Didn't think through exactly how I worded things.

tldr: When you have insightful conversation, write it down and share it so people can build on it (instead of just sharing in person). Most of humanity's power comes from being able to build complex thoughts out of other thoughts and transmit them across the world.

This is a rehash/re-examining post. Related:

- Write Down Your Process (Zvi Mowshowitz)
- Some Thoughts on Public Discourse (Holden Karnofsky)
- Why and How to Name Things (Conor Moreton)
- Single Locus of Discussion (Anna Salamon)
- Return to Discussion (Sarah Constantin)
- [Edit: Turning Discussions into Blogposts by Brian Tomasik apparently covers this very topic. Still needed to be said again]

This is part 2 of N of my "Ray writes down conversations he had with people" series. It's also the most adorably meta of them.

A month ago I was talking with Oliver Habryka about why Less Wrong was important. (Call us biased if you will). One thing we both noted: in the days of yore, it seemed like a lot of prominent scholars/thinkers wrote down their insights and research on Less Wrong. Then, eventually they turned professional and joined official organizations whose job was to think fulltime.

Also over the past few years, those organizations (including but not limited to MIRI, CFAR, Givewell/OpenPhil) shifted from being younger-with-nothing-to-lose to older-with-reputations-to-safeguard, and their public facing tone seems to have shifted from "earnestly sharing thoughts as they come up" to "carefully crafted PR statements."

Thirdly, a lot of people moved to major geographic hubs, where it became easier to have in person conversations than to communicate via written blogpost. So... that's what people have tended to do.

I sympathize with the notion that people are busy and writing things up is a) time consuming and b) potentially risky. But I think the consequences of this are at least underweighted.

At *least*, I think people having informal conversations should make more of an effort to write those up in accessible form when there *aren't* reasons to be cautious.

Consequences of *not* writing stuff down include:

1. If you're not plugged into the personal-conversation-network, it's hard to keep up with a lot of collective insights relating to both rationality, effective altruism and x-risk. (This results in weird filtering effects which aren't the *worst* - being able to network your way is a credible signal of *something*. But I don't think it's the best possible filter)
2. It's actually fairly time consuming to propogate ideas via in-person conversation. Like, one of the *major* advantages of humanity is ideas being able to efficiently

spread via writing.

3. Another major advantage of writing is being able to *increase your collective working memory and build on ideas.* When much of your insights are developed via conversation, not only are you preventing far-away-people from building on your ideas, but you are hampering the ability of *yourself and your immediate colleagues* to build on those ideas. Writing things down (and then turning them into essays with titles that can be easily referenced) makes it easier to build complex models.

I think people feel a lot of pressure to write things up *well*, and as a result don't write things up at all. So my current take is, if you have a conversation that seems to contain important insights, err on the side of getting it written up *quickly* without as much regard for being timeless.

Later on if you think it deserves to be written up in a more timeless, scholarly form, you (or someone with more time) can still do that.

Demon Threads

tldr: a Demon Thread is a discussion where everything is subtly warping towards aggression and confusion (i.e. as if people are under demonic influence), even if people are well intentioned and on the same 'side.' You can see a demon thread coming in advance, but it's still hard to do anything about.

("Flame Wars" are similar but I felt the connotation was more like "everything has already *gone to hell, and people aren't even pretending to be on the same side")*

I kept wanting to reference this post when discussing internet discussion policy, and I kept forgetting that nobody has written it yet. So here it is.

Suggested Background Reading:

- Politics is Hard Mode (Rob Bensinger)
- Civility is Never Neutral (Ozy)
- Writing that Provokes Comments (me)
- Musings on Double Crux and Productive Disagreement (me)

If someone in the future linked you to this post, it's probably because a giant sprawling mess of angry, confused comments is happening - or is about to happen - and it's going to waste a lot of time, make people upset, and probably *less* likely to listen to each other about whatever the conversation ostensibly is about.

I have some ideas on what to do *instead*, which I discuss in this followup post.

But for now, this post is meant to open a discussion, explore the mechanics of how demon threads work, and then in the comments brainstorm solutions about how to handle them.

Wrong On the Internet

I find “Someone Is Wrong On the Internet” to be a weird, specific feeling.

It's distinct from someone being *factually wrong* - people can be wrong, point it out, and hash out their disagreements without a problem. But a common pattern I’ve witnessed (and experienced) is to notice someone being wrong in a way that *feels distinctly bad*, like if you don’t correct them, something precious will get trampled over.

This is when people seem most prone to jump into the comments, and it’s when I think people *should be most careful*.

Sometimes there actually *is* an important thing at stake.

There usually isn’t.

It often *feels* like there is, because our social intuitions were honed for tribes of a hundred or two, instead of a world of 7 billion. We live in a different world now. If you actually want to have an impact on society, yelling at each other on the internet is almost certainly not the best way to do so.

When there *actually is* something important at stake, I think there are usually better plans than “get into a giant internet argument.” Think about what your goals are. Devise a plan that actually seems like it might help.

Different situations call for different plans. For now, I want to talk about the common anti-pattern that often happens instead.

Demon Threads are explosive, frustrating, many-tentacled conversations that draw people in regardless of how important they are. They come in two forms:

- **Benign Demon Threads** are mostly *time wasting.* Nobody gets *that* angry, it's just a frustrated mess of "you're wrong" "no *you're* wrong" and then people spend loads of digital ink arguing about something that doesn't matter much.
- **Malignant Demon Threads** feed upon emotions of defensiveness, anger, tribal affiliation and righteousness - and inflame those emotions, drawing more people into the fire.

(A malignant demon thread is cousin to the *flame war* - people hurling pure insults at each other. What makes a malignant demon thread insideous is the way it can warp discussion even among people who are earnestly trying to communicate, seek truth and solve problems)

If you find yourself in a malignant demon thread, I think it's likely you are not only *not helping*, but are actually hurting your cause.

The Demon Seed

*How to write so that people will comment [disclaimer: not necessarily **good** advice]*

> *1. Be wrong*
>
> *2. Be controversial*
>
> *3. Write things people feel qualified to have opinions on.*
>
> *4. Invoke social reality.*
>
> - *Writing That Provokes Comments*

In the comments on YouTube, or the worst parts of Facebook or tumblr, demon threads are not surprising. People write comments that inflame ideological warfare all the time. Internets be internets. People be people. What can you do?

The surprising thing is how this works in places where everyone should really know better. The powers of demons are devious and subtle.

There's an experiment — insert obligatory replication crisis disclaimer — where one participant is told to gently poke another participant. The second participant is told to poke the first participant the same amount the first person poked them.

It turns out people tend to poke back slightly harder than they were first poked.

Repeat.

A few iterations later, they are striking each other really hard.

I think something like this is at work in the mechanics of demon threads.

The **Demon Seed** is the first comment in what will soon become a demon thread. It might look pretty innocuous. Maybe it feels *slightly* rude, or slightly oblivious, or pushing a conversation that should be about concrete empirical facts slightly towards being about social consensus (or, vice versa?).

It feels 1% outside the bound of a reasonable comment.

And then someone **waters the demon seed**. They don't want to let the point stand, so they respond with what seems like a fair rebuke.

Maybe they're self-aware that they're feeling annoyed, so they intentionally dial back the aggression of their response. "Ah, actually this probably comes across as too hostile, so I'll tweak my wording to reduce hostility by 4%." But, actually, the words were 6% more hostile than they thought, and now they've escalated 2%.

Repeat 2-3 times. The demon seed is watered. Latent underlying disagreements about how to think properly... or ideal social norms... or which coalitions should be highest status... or pure, simple *you're insulting me and I'm angry*…

They have festered and now they are ready to explode.

Then someone makes a comment that pushes things over the edge, and a demon thread is born.

(It is, of course, possible to skip steps 1-4 and just write a blatantly rude, incendiary comment. I'm trying to describe how this happens *even when everyone is well intentioned and mostly trusts each other)*

From there, if you're *lucky* it's contained to two people. But often, well meaning bystanders will wander by and think "Ah! People are being wrong on the internet! Wrong about things I am qualified to have opinions on! I can help!"

And it grows.

Then people start linking it from elsewhere, or FB algorithms start sharing it because people are *commenting* so the thread must be *important*.

It grows further.

And it consumes days of people's attention and emotional energy. More importantly, it often entrenches people's current opinions, and it burns people's *good will* that they might have been willing to spend on honest, cooperative discourse.

Why Demon Threads are Bad

I think demon threads are not just a bad plan - I think they are often *net negative* plan.

The reason is best expressed in Conor Moreton's Idea Innoculation and Inferential Distance. [Edit: the full article is no longer available]

> **Inferential distance** is the gap between [your hypotheses and world model], and [my hypotheses and world model]. It's *just how far out* we have to reach to one another in order to understand each other.
>
> If you share political and intellectual and cultural foundations, it's (relatively) easy. If you have completely different values and assumptions, (say you get dropped off in the 15th century and need to argue with Christopher Columbus) it may be nigh impossible.
>
> It's right in the name—*inferential* distance. It's not about the "what" so much as it is about the "how"—how you infer new conclusions from a given set of information. When there's a large inferential distance between you and someone else, you don't just disagree on the object level, you also often disagree about ***what counts as evidence, what counts as logic,*** and ***what counts as self-evident truth.***
>
> What makes this really bad is **idea inoculation.**
>
> When a person is exposed to a weak, badly-argued, or uncanny-valley version of an idea, they afterwards are *inoculated* against stronger, better versions of that idea. The analogy to vaccines is extremely apt—your brain is attempting to conserve energy and distill patterns of inference, and once it gets the shape of an idea and attaches the flag "bullshit" to it, it's *ever after* going to lean toward attaching that same flag to any idea with a similar shape.
>
> When you combine idea inoculation with inferential distance, you get a recipe for disaster—if your first attempt to bridge the gap fails, your second attempt will *also* have to overcome the person's rapidly developing resistance.
>
> You might think that each successive attempt will bring you closer to the day that you finally establish common ground and start communicating, but alas—often, each attempt is just increasing their resistance to the core concept, as they build up a library of all the times they saw something like this defeated, proven wrong, made to look silly and naive.

A demon thread is a *recipe for bad attempts at communicating*. Lots of people are yelling at once. Their defenses are raised. There's a sense that if you give in, you or your people look like losers or villains.

This'll make people worse at listening *and* communicating.

Why the Internet Worse

"Demon threads" can happen in person, but they're worse online.

One obvious reason is that **the internet is more anonymous**. This reduces consequences to the person writing a comment, and makes the target of the comment easier to round off to a bad stereotype or an abstract representation of The Enemy.

Other things people do:

A. People end up writing long winded monologues without anyone interrupting them to correct basic, wrong assumptions.

i.e. "you're just wrong because you think X, therefore... [complicated argument]", without providing opportunity for someone to respond "no I don't actually think X at all". And then, having written out [complicated argument] you're already *invested* in it, despite it being built on faulty premises.

B. Lots of people are writing. Especially as the demon thread grows. After 24 hours of its existence, the thread will have so much content it's a huge investment to actually read everything that's been said.

C. The comments aren't necessarily displayed in order. Or, if they are, people aren't reading them in order, they're reading whatever it's largest or most interesting.

D. The internet is full of lots of other content competing for attention.

This all means that:

E. People are *skimming*. This is most true when lots of people are writing lengthy monologues, but even when the thread first begins, people's eyes may be bouncing around to different tabs or different threads within a page so they *aren't even reading what's being said*, not with the intentionality and empathy they would when confronted with a real person in front of them.

And they might *first* be reading the most explosive, recent parts of a thread rather than piecing together the actual order of escalation, which may make people look less reasonable than they were.

This all adds up to giant threads being a *uniquely bad way to resolve nuanced, emotionally fraught issues.*

Containment?

Demon threads are like wildfires. *Maybe* you can put them out, with coordinated effort. You can also try to ignore them and hope they burn themselves out.

But if you wanted to actually stop it, the best bet is to do so is before they're erupted in the first place.

I've developed a sense of what seeds look like. I'll see a comment, think "god, this is going to become a demon thread in like two hours", and then sure enough, two hours later people are yelling at each other and everything is awful and everyone involved seems *really sure* that they are helping somehow.

Some flags that a demon thread might be about to happen:

Flags Regarding: Tension and Latent Hostility

- When you look a comment and want to respond, you feel a visceral sense of "you're *wrong*", or "ugh, those people [from group that annoy me]" or "important principle must be defended!" or “I am literally under attack.”

- You feel physiological defensiveness or anger - you notice the hairs on the back of your neck or arms standing on end, or a tightness in your chest, or however those emotions manifest in your body.
- People in the thread seem to be talking past each other.
- For whatever reason, tensions seem to be escalating.

Flags Regarding: Social Stakes

- The argument seems like it's about who should be high or low status, which people or groups are virtuous and which are not, etc.
- The argument is about social norms (in particular if at stake is whether some people will end up feeling unwelcome or uncomfortable in a given community/space that is important to them - this is *extremely* threatening)
- More generally - the argument touches *in some way* on social reality, in ways that might have ramifications beyond the immediate conversation (or that people are *afraid* might have such ramifications).

If some of the above seem true (in particular, at least one of the first group and at least one of the second), then I think it's worth stepping back and being *very careful* about how you engage, even if no comment seems especially bad yet.

Potential Solutions

The first line of defense is to notice what’s happening - recognize if you’re feeling defensive or angry or talking past each other. Brienne's Noticing Sequence is pretty good for this (as well as her particular posts on training the skills of Empathy and handling Defensiveness - these may not work for everyone but I found the underlying thought process useful).

But while noticing is necessary, it’s not *sufficient.*

Rather than list my first guesses here, I’ll be discussing them in the comments and following this up with a “best-seeming of the potential solutions” post.

Meanwhile, some factors to consider as you decide what to do:

- How are you involved?
    - Are you one of the people initially arguing, or a bystander?
    - How much do you normally trust the people involved?
    - Is it possible to take the conversation private?
- Are we on the demon *seed* or demon *thread* stage? Is there common knowledge about either?
- What are the actual stakes?
- What are the moderation tools available to you?
- Are you in a venue where you have the ability to shape conversational norms?
    - Do you directly control them (i.e. personal blog or feed?)
    - Does *anyone* have direct ownership of the venue? (either technically, or culturally)
    - Is there anything you can do *unilaterally* to make the conversation better, or will it require help from others?
- Are you building a site where you get to develop entire new tools to deal with this class of problem?

With that in mind…

In whatever venues you most find yourself demon-thread-prone, what sort of plans can you actually think of that *might actually help?*

Note: I have since written a followup post with a working example of what I think people should usually do instead of demon threads.

Taking it Private: Short Circuiting Demon Threads (working example)

This post is intended as a working example of how I think Demon Threads should be resolved. The gist of my suggestion is:

> **Step 1.** Make it easy and common to *take a conversation private* if someone is feeling annoyed/threatened/angry/etc (*if* it seems like the conversation is actually important. Meanwhile, *also* make it easier to tap out if the conversation doesn't seem like the best use of your time)
>
> **Step 2.** In private chat, two people do their best to communicate honestly, to notice when they are defensive, to productively find the truth as best they can. (I think this is much easier 1-on-1 than in public)
>
> **Step 3.** Someone writes a short summary of whatever progress they were able to make (and any major outstanding disagreements that remain), focusing primarily on *what they learned* and rather than "who's right."
>
> The summary should be something both parties endorse. Ideally they'd both sign off on it. If that trivial inconvenience would prevent you from actually writing the post, and you both generally trust each other, I think it's fine to make a good-faith effort to summarize and then correct each other if they missed some points.
>
> Writing such a summary needs to get you as much kudos / feel-good as winning an argument does.
>
> **Step 4.** The public conversation continues, with the benefit of whatever progress they made in private.
>
> Ideally, this means the *public* conversation gets to progress, without being as emotionally fraught, and every time something comes up that *does* feel fraught, you recurse to steps 1-3 again.

Qiaochu had a criticism of the Demon Thread article. I had said:

> Demon Threads are explosive, frustrating, many-tentacled conversations that feel important but aren’t.

He responded:

> I want to object to this framing, particularly the "but aren't." It's far from clear to me that demon threads are unimportant. It may seem like nothing much happened afterwards, but that could be due to everyone in the thread successfully canceling out everyone else's damage. If that's true it means that no one side can unilaterally back down in a demon thread without the thing they're protecting potentially getting damaged, even while the actual observed outcome of demon threads is that nobody apparently benefited.

I initially responded publicly. (I think the details are important in their own right, <linked here>, but aren't the main point of this post)

We still disagreed, and the nature of the disagreement hinged on past threads full of social drama. This was exactly the sort of thing I didn't want to discuss publicly on the internet. Yes, the details mattered, but public discussion would have lots of bystanders showing up with opinions about the object-level-details about the social drama itself.

In this case, Qiaochu and I were able to discuss it privately, which:

- helped my own demon thread model
- was a useful working example of what should come out of steps 1-4.

So I've written this up as a post instead of a comment. I haven't run this by Qiaochu yet (I think getting formal permission/endorsement adds an "significant trivial inconvenience" that might disrupt the process too much), but I expect him to endorse the following, and I'll update/clarify if I got anything wrong.

Things I learned

i. Mattering-ness is orthogonal to Demon-Thread-ness

The most important update on my part. Qiaochu provided a few examples where it felt right to call a thing a demon thread, which the thing-in-question mattered in some sense - either because the tribal affiliation and status mattered, or because the actual ideas getting discussed mattered.

"Is it a demon thread?" is more about "there is some weird force compelling more and more people to argue, raising tensions" than it's about "the argument is counterproductive or going in circles" (although I think the latter is common).

Since matter-ness isn't part of the central definition, I've removed it from the description at the beginning of the post.

ii. Avoid bundling normative claims with descriptive claims.

One reason I think Demon Threads (often) don't matter is a *normative claim* about what people should value, and it is unfair to bundle this with claims about what people should do given what they *currently* value*,* even if I think they're being silly.

Conflating descriptive and normative claims can be a useful (but deceptive) rhetorical trick, and part of the point of LessWrong is to avoid doing that so we can think clearly about things.

Empirically, people care about what groups and ideas have relative status among their peers.

My point was more like: *Arguing on the internet about the relative status of things is* ***not effective altruism.*** People care about things other than accomplishing the greatest good for the least effort. It is perhaps *most* of what most people care about. And that's fine.

I think this claim is still relevant, because people often seem to *think* the thing they're doing is helping a much larger amount than it is, as well as accomplishing *different* things than they think it is. (i.e. you think talking about the president is having an impact on national policy, but it's mostly having an impact on what opinions are acceptable to express in your local peer group).

I do think, in most social/political-drama-laden threads, if people took a step back and thought about it, they would *either* realize an internet debate wasn't the best way to accomplish their goals, *or* they'd realize their goals were different than they thought they were.

iii. Maybe something didn't matter *before* the demon thread, but *after* a giant explosion of arguments happens, it may matter a lot (at least to the people involved).

I cited an example where, in a local community, people started arguing about [internet drama from several years ago]. Prior to the argument, it hadn't mattered what your opinions about that particular political drama was. But suddenly, everyone knew what many prominent community-member's opinions were, and people disagreed strongly, and there was a risk that if the thread went the wrong way, *having one set of opinions might no longer be okay.*

Qiaochu and I agreed it would have been better if the argument never happened, and that the political drama wasn't objectively important. But he argued, once it *had* exploded, it *became* relevant to the people involved. So the pressure to add your 2 cents was real and important.

This seems true, but also feeds back into my central claim, which is that it's best to stop malignant demon threads before they begin.

Outstanding Disagreements

Often, when people are coming from very different intuitions, they can argue a lot about factual claims, and agree that each other made good points... and still go back to basically holding their original position.

This can be frustrating, but understandable: people have a lot of background experience that feeds into whether something makes sense. Explicit arguments often can't fully address that background experience.

While we agreed with many of each other's claims in principle, each claim of Qiaochu and I included lots of words like "usually" and "sometimes", that were doing a lot of work, and our respective takeaways rounded those words in directions closer to our original positions.

Qiaochu's current overall position as I understand it is:

> People are constantly tracking the relative status of groups and ideas, our intuitions about this are actually pretty good - both at detecting what's going on, and whether it is relevant to our goals.

(I originally interpret this to basically be arguing *against* the "We're adapted for Dunbar Number tribes, therefore our intuitions for the modern world are useless" hypothesis, which seemed confusing. In the comments below Qiaochu clarified among other things that although our society is bigger, so are our tools for broadcasting signals. See his comment for more clarity)

My current position is:

People's intuitions for tracking the relative status of groups or ideas is doing s*omething,* but it's not really doing what they think it is, and it's not well adapted for the modern world. Lots of things matter as much, or more, than political goals, but we have a much easier time understanding (or thinking we understand) political goals, so we spend disproportionate time on them.

Meanwhile, because the modern world *is* different, accomplishing political goals that are relevant outside of your immediate social circle usually requires doing things that feel counterintuitive.

Meta-tations on Moderation: Towards Public Archipelago

The recent moderation tools announcement represents a fairly major shift in how the site admins are approaching LessWrong. Several people noted important concerns about transparency and trust.

Those concerns deserve an explicit, thorough answer.

Summary of Concepts

1. **The Problem of Private Discussion** – Why much intellectual progress in the rationalsphere has happened in hard-to-find places
2. **Public Discussion vs Intellectual Progress** – Two subtly conflicting priorities for LessWrong.
3. **Healthy Disagreement** – How to give authors tools to have the kinds of conversations they want, without degenerating into echo chambers.
4. **High Trust vs Functioning Low Trust environments** – Different modes of feeling safe, with different costs and risks.
5. **Overton Windows, Personal Criticism** – Two common conversational attractors. Tempting. Sometimes important. But rarely what an author is interested in talking about.
6. **Public Archipelago** - A model that takes all of the above into account, giving people the tools to create person spaces that give them freedom to explore, while keeping all discussion public, so that it can be built upon, criticized, or refined.

i. The Problem

The issue with LessWrong that worries me the most:

In the past 5 years or so, there's been a lot of progress – on theoretical rationality, on practical epistemic and instrumental rationality, on AI alignment, on effective altruism. But much of this progress has been on some combination of:

- On various private blogs you need to keep track of.
- On facebook – where discussions are often private, where searching for old comments is painful, and some people have blocked each other so it's hard to tell what was actually said and who was able to read it.
- On tumblr, whose interface for following a conversation is the most confusing thing I've ever seen.
- On various google docs, circulated privately.
- In person, not written down at all.

People have complained about this. I think a common assumption is something like "if we just got all the good people back on LessWrong at the same time you'd have a critical mass that could reboot the system." That might *help,* but doesn't seem sufficient to me.

I think LW2.0 has roughly succeeded at becoming "the happening place" again. But I still know several people who I intellectually respect, who find LessWrong an actively inhospitable place and don't post here, or do so only grudgingly.

More Than One Way For Discussion To Die

I realize that there’s a very *salient* pathway for moderators to abuse their power. It’s easy to imagine how echo chambers could form and how reign-of-terror style moderation could lead to, well, reigns of terror.

It may be less salient to imagine a site subtly driving intelligent people away due to being boring, pedantic, or frustrating, but I think the *latter is in fact more common, and a bigger threat to intellectual progress.*

The current LessWrong selects somewhat for people who are thick skinned and conflict prone. Being thick-skinned is good, all being equal. Being conflict prone is not. And neither of these are the same as *being able to generate useful ideas and think clearly*, the most important qualities to cultivate in LessWrong participants.

The site admins don’t just have to think about the people currently here. We have to think about people who have things to contribute, but don’t find the site rewarding.

Facebook vs LessWrong

When I personally have a new idea to flesh out... well...

...I'd prefer a LessWrong post over a Facebook post. LW posts are more easily linkable, they have reasonable formatting options over FB's plain text, and it's easier to be sure a lot of people have seen it.

But to *discuss* those ideas…

In my heart of hearts, if I weren't actively working on the LessWrong team, with a clear vision of where this project is going... I would prefer a Facebook comment thread to a LessWrong discussion.

There are certain blogs – Sarah, Zvi, Ben stick out in my mind, that are comparably good. But not many – the most common pattern is "post idea on blog, and the good discussion happens on FB, and individual comment insights only make it into the broader zeitgeist if someone mentions them in a high profile blogpost."

On the right sort of Facebook comment thread, at least in my personal filter bubble, I can expect:

- People I intellectually respect to show up and hash out ideas.
- A collaborative attitude. "Let's figure out and build a thing together."
- People who show up will share enough assumptions that we can talk about refining the idea to a usable state, rather than "is this idea even worth talking about?"

Beyond that, more subtle: even if I don't know everyone, an intellectual discussion on FB usually feels like, well, *we're friends*. Or at least *allies.*

Relatedly: the *number* of commenters is manageable. The comments on Slatestarcodex are reasonably good these days, but… I'm just not going to sift through hundreds or thousands of comments to find the gems. It feels like a firehose, not a conversation.

Meanwhile, the comments on LessWrong often feel... nitpicky and pointless.

If an idea isn't presented maximally defensibly, people will focus on tearing holes in the *non-loading-bearing* parts of the idea, rather than help refine the idea into something more robust. And there'll be people who disagree with or don't understand foundational elements that the idea is supposed to be *building off of*, and the discussion ends up being about rehashing 101-level things instead of building 201-level knowledge.

Filter Bubbles

An obvious response to the above might be "of *course* you prefer Facebook over LessWrong. Facebook heavily filter bubbles you so that you don't have to face disagreement. It's *good* to force your ideas to intense scrutiny."

And there's important truth to that. But my two points are that:

1. I think a case can be made that, during *idea formation*, the *kind* of disagreement I find on Facebook, Google Docs and in-person is actually better from the standpoint of intellectual progress.
2. Whether or not #1 turns out to be true, if people prefer private conversations over public discussions (because they're easier/more-fun/safer), then much discussion will tend to continue taking place in mostly private places, and *no matter how suboptimal this is, it won't change*.

My experience is that my filter bubbles (whether on FB, Google Docs or in-person) *do* involve a lot of disagreement, and the disagreement is *higher quality.* When someone tells me I'm wrong, it's often accompanied by an attempt to understand what my goals are, or what the core of a new idea was, which either lets me fix an idea, or abandon it but find something better to accomplish my original intent.

(On FB, this isn't because the *average* commenter is that great, but because of a smallish number of people I deeply respect, who have different paradigms of thinking, at least 1-2 of whom will reliably show up)

There seems to be a sense that good ideas form fully polished, without any work to refine them. Or that until an idea is ready for peer review, you should keep it to yourself? Or be willing to have people poke at it with no regard how hedonically rewarding that experience is? I'm not sure what the assumption is but it's contrary to how everyone I personally know generates insights.

The early stages work best when *playful* and collaborative.

Peer review is important, but so is idea formation. Idea formation often involves running with assumptions, crashing them into things and seeing if it makes sense.

You could keep idea-formation private and then share things when they're 'publicly presentable', but I think this leads to people tending to keep conversation in "safe, private" zones longer than necessary. And meanwhile, it's valuable to be able to *see* the generation process among respected thinkers.

Public Discussion vs Knowledge Building

Some people have a vision of Less Wrong as a *public discussion.* You put your idea out there. A conversation happens. Anyone is free to respond to that conversation as long as they aren't being actively abusive. The best ideas rise to the top.

And this is a fine model, that should (and does) exist in some places. But:

1. It's never actually been the model or ethos LessWrong runs on. Eliezer wrote Well Kept Gardens Die By Pacifism years ago, and has always employed a Reign-of-Terror-esque moderation style. You may disagree with this approach, but it's not new.
2. A public discussion is not *necessarily* the same as the ethos Habryka is orienting around, which is to *make intellectual progress.*

These might seem like the same goal. And I share an aesthetic sense that *in the 'should' world,* where things are fair, public discussion and knowledge-building are somehow the same goal.

But we don't live in the 'should' world.

We live in the world where *you get what you incentivize.*

Yes, there's a chilling effect when authors are free to delete comments that annoy them. But there is a different chilling effect when *authors* aren't free to have the sort of conversation they're actually interested in having. The conversation won't happen at all, or it'll happen somewhere else (where you can't comment on their stuff *anyway)*.

A space cannot be universally inclusive. So the question is: is LessWrong *one space*, tailored for only the types of people who enjoy that space? Or do we give people tools to make their own spaces?

If the former, who is that space for, and what rules do we set? What level of knowledge do we assume people must have? We’ve long since agreed “if you show up arguing for creationism, this just isn’t the space for you.” We’ve generally agreed that if you are missing concepts in the sequences, it’s your job to educate yourself before trying to debate (although veterans should politely point you in the right direction).

What about posts written since the sequences ended?

What *skills* and/or *responsibilities* do we assume people must have? Do we assume people have the ability to notice and speak up about their needs a la Sarah Constantin’s Hierarchy of Requests? Do we require them to be able to express those needs ‘politely’? Whose definition of polite do we use?

No matter which answer you choose for any of these questions, some people are going to find the resulting space inhospitable, and take their conversation elsewhere.

I’d much rather sidestep the question entirely.

A Public Archipelago Solution

Last year I explored applying Scott Alexander's Archipelago idea towards managing community norms. Another quick recap:

> Imagine a bunch of factions fighting for political control over a country. They've agreed upon the strict principle of harm (no physically hurting or stealing from each other). But they still disagree on things like "does pornography harm people", "do cigarette ads harm people", "does homosexuality harm the institution of marriage which in turn harms people?", "does soda harm people", etc.
>
> And this is bad not just because everyone wastes all this time fighting over norms, but because the nature of their disagreement incentivizes them to fight over *what harm even is.*
>
> And this in turn incentivizes them to fight over both definitions of words (distracting and time-wasting) and *what counts as evidence or good reasoning* through a politically motivated lens. (Which makes it harder to *ever* use evidence and reasoning to resolve issues, even uncontroversial ones)

And then...

> Imagine someone discovers an archipelago of empty islands. And instead of continuing to fight, the people who want to live in Sciencetopia go off to found an island-state based on ideal scientific processes, and the people who want to live in Libertopia go off and found a society based on the strict principle of harm, and the people who want to live in Christiantopia go found a fundamentalist Christian commune.
>
> **This lets you test more interesting ideas.** If a hundred people have to agree on something, you'll only get to try things that you can can 50+ people on board with (due to crowd inertia, regardless of whether you have a formal democracy)
>
> But maybe you can get 10 people to try a more extreme experiment. (And if you share knowledge, both about experiments that work and ones that don't, you can build the overall body of community-knowledge in your social world)

Taking this a step farther is the idea of *Public Archipelago,* with islands that overlap.

Let people create their own spaces. Let the conversations be *restricted* as need be, but *centralized and public*, so that everyone at least has the opportunity to follow along, learn, respond and build off of each other's ideas, instead of having to network their way into various social/internet circles to keep up with everything.

This necessarily means that not *all* of LessWrong will be a comfortable place to any given person, but it at least means a wider variety of people will be able to use it, which means a wider variety of ideas can be seen, critiqued, and built off of.

Healthy Disagreement

Now, there's an obvious response to my earlier point about "it's frustrating to have to explain 101-level things to people all the time."

Maybe you’re *not* explaining 101-level things. Maybe you’re actually *just wrong* about the foundations of your ideas, and your little walled garden isn’t a 201 space, it’s an echo chamber built on sand.

This is, indeed, quite a problem.

It’s an even harder problem than you might think at first glance. It’s difficult to offer an informed critique of something that’s actually useful. I’m reminded of Holden Karnofsky’s Thoughts on Public Discourse:

> For nearly a decade now, we've been putting a huge amount of work into putting the details of our reasoning out in public, and yet I am hard-pressed to think of cases (especially in more recent years) where a public comment from an unexpected source raised novel important considerations, leading to a change in views.
>
> This isn't because nobody has raised novel important considerations, and it certainly isn't because we haven't changed our views. Rather, it seems to be the case that we get a large amount of valuable and important criticism from a relatively small number of highly engaged, highly informed people. Such people tend to spend a lot of time reading, thinking and writing about relevant topics, to follow our work closely, and to have a great deal of context. They also tend to be people who form relationships of some sort with us beyond public discourse.
>
> The feedback and questions we get from outside of this set of people are often reasonable but familiar, seemingly unreasonable, or difficult for us to make sense of.

The obvious criticisms of an idea may have obvious solutions. If you interrupt a 301 discussion to ask “but have you considered that you might be wrong about everything?”... well, yes. They have probably noticed the skulls. This often feels like 2nd-year undergrads asking post-docs to flesh out everything they’re saying, using concepts only available to the undergrads.

Still, peer review *is* a crucial part of the knowledge-building process. You need high quality critique (and counter-critique, and counter-counter-critique). How do you square that with giving an author control over their conversation?

I hope (and fairly confidently believe) that most authors, even ones employing Reign-of-Terror style moderation policies, will *not* delete comments willy nilly – and the site admins will be proactively having conversations with authors who seem to be abusing the system. But we *do* need safeguards in case this turns out to be worse than we expect.

The answer is pretty straightforward: it’s not at all obvious that the public discussion of a post has to be on *that particular post’s* comment section.

(Among other things, this is not how most science works, AFAICT, although traditional science leaves substantial room for improvement anyhow).

If you disagree with a post, and the author deletes or blocks you from commenting, you are welcome to write another post about your intellectual disagreement.

Yes, this means that people reading the original post may come away with an impression that a controversial idea is more accepted than it really is. But if that person looks at the front page of the site, and the idea *is* controversial, there will be both other posts and recent comments arguing about its merits.

It also means that no, you don’t automatically get the engagement of everyone who read the original post. I see this as a feature, not a bug.

If you want your criticism to be read, *it has to be good and well written.* It *doesn't* have to fit within the overall zeitgeist of what's currently popular or what the locally high-status people think. Holden's critical [Thoughts on Singularity Institute](#) is one of the most highly upvoted posts of all time. (If anything, I think LessWrong folk are *too* eager to show off their willingness to dissent and upvote people just for being contrarian).

It does suck that you must be good at writing and know your audience (which isn't necessarily the same as good at thinking). But this applies just as much to being the original author of an idea, as to being a critic.

The author of a post doesn't owe you their rhetorical strength and audience and platform to give you space to write your counterclaim. We don't want to incentivize people to protest quickly and loudly to gain mindshare in a popular author's comment section. We want people to write good critiques.

Meanwhile, if you're making an effort to understand an author's goals and frame disagreement in a way that doesn't feel like an attack, I don't anticipate this coming up much in the first place.

ii. Expectations and Trust

I think a deep disagreement that underlies a lot of the debate over moderation: *what sort of trust is important to you?*

This is a bit of a digression – almost an essay unto itself – but I think it's important.

Elements of Trust

Defining trust is tricky, but here's a stab at it: "Trust is having expectations of other people, and *not having to worry* about whether those expectations will be met."

This has a few components:

- **Which expectations** do you care about being upheld?
- **How *much* do you trust** people in your environment to uphold them?
- **What strategies do you prefer** to resolve the cognitive load that comes when you *can*'t trust people (or, are not *sure* if you can)?

Which expectations?

You might trust people…

- to keep their promises and/or mean what they say.
- to care about your needs.
- to uphold particular principles (clear thinking, transparency).
- to be able (and willing) to perform a particular skill (including things like *noticing* that when you're *not* saying what you mean).

Trust is a [multiple-place function](#). Maybe you trust Alice to reliably provide all the relevant information even if it makes her look bad. You trust Bob to pay attention to your emotional state and not say triggering things. You can count on Carl to call you on your own bullshit (and listen thoughtfully when you call him on his). Eve will reliably enforce her rules even when it's socially inconvenient to do so.

You may care about different kinds of trust in different contexts.

How much do you trust a person or space?

For the expectations that matter most to you, do you generally expect them to be fulfilled, or do you have to constantly monitor and take action to ensure them?

With a given person, or a particular place, is your guard always up?

In *high trust environments*, you expect other people to care about the same expectations you do, and follow through on them. This might mean looking out for each other's interests. Or, merely that you're focused on the same goals such that "each other's interests" doesn't come into play.

High trust environments require you to either personally know everyone, or to have strong reason to believe in the selection effects on who is present.

Examples:

- *A small group of friends by a campfire* might trust each other to care about each other's needs and try to ensure they are met (but *not* necessarily to have particular skills required to do so).
- *A young ideological startup* might trust each other to have skills, and to care about the vision of the company (but, perhaps *not* to 'have each other's back' as the company grows and money/power becomes up for grabs)
- *A small town*, where families have lived there for generations and share a culture.
- *A larger military battalion***,** where everyone knows that everyone knows that everyone went through the same intense training. They clearly have particular skills, and would suffer punishment if they don't follow the orders from high command.

Low trust environments are where you have no illusions that people are looking out for the things you care about.

The barriers to entry are low. People come and go often. People often represent themselves as if they are aligned with you, but this is poor evidence for whether they are in fact aligned with you. You must constantly have your guard up.

Examples:

- A large corporation where no single person knows everybody
- A large community with no particular barrier to entry beyond showing up and talking as if you understand the culture
- A big city, with many cultures and subcultures constantly interfacing.

Transparent Low Trust, Curated High Trust

Having to watch your back all the time is exhausting, and there's at least two strategy-clusters I can think of to alleviate that.

In a **transparent low trust environment**, you don't *need* to rely on anyone's word or good intentions. Instead, you rely upon transparency and safeguards built into the system.

It's your responsibility to make use of those safeguards to check that things are okay.

A **curated high trust environment** has some kind of strong barrier to entry. The advantage is that things can move faster, be more productive, require less effort and conflict, and focus only on things you care about.

It's the *owner* of the space's responsibility to kick people out if they aren't able to live up to the norms in the space. It's *your* responsibility to decide whether you trust the the space, and leave if you don't.

The current atmosphere at LessWrong is something like “transparent medium trust.” There are rough, site-level filters on what kind of participation is acceptable – much moreso than the average internet hangout. But not much micromanaging on what precise expectations to uphold.

I think some people are expecting the new moderation tools to mean “we took a functioning medium trust environment and made it more dangerous, or just weirdly tweaked it, for the sake of removing a few extra annoying comments or cater to some inexplicable whims.”

But part of the goal here is to create a *fundamental phase shift,* where types of conversations are *possible* that just *weren't* in a medium-trust world.

Why High Trust?

Why take the risk of high trust? Aren’t you just exposing yourself to people who might take advantage of you?

I know some people who’ve been *repeatedly* hurt, by trying to trust, and then having people either trample all over their needs, or actively betray them. Humans are political monkeys that make up convenient stories to make themselves look good all the time. If you aren’t actually aligned with your colleagues, you will probably eventually get burned.

And high trust environments can’t *scale* – too many people show up with too many different goals, and many of them are good at presenting themselves as aligned with you (they may even *think* they’re aligned with you), but… they are not.

LessWrong (most likely) needs to scale, so it’s important for there to be spaces here that are Functioning Low Trust, that don’t rely on load-bearing authority figures.

I do *not* recommend this blindly to everyone.

But. To misquote Umesh – "*If you’re not occasionally getting backstabbed, you’re probably not trusting enough.*"

If you can trust the people around you, all the attention you put into watching your back can go to other things. You can expect other people to look out for your needs, or help you in reliable ways. Your entire body physiologically changes, no longer poised for fight or flight. It’s physically healthier. In some cases it’s better for your epistemics – you’re less defensive when you don’t feel under attack, making it easier to consider opposing points of view.

I live most of my life in high trust environments these days, and… let me tell you *holy shit when it works it is amazing.* I know a couple dozen people who I trust to be honest about their personal needs, to be reasonably attentive to mine, who are aligned with me on how to resolve interpersonal stuff as well as Big Picture How the Universe Should Look Someday.

When we disagree (as we often do), we have a shared understanding of how to resolve that disagreement.

Conversations with those people are smooth, productive, and insightful. When they are not smooth, the process for figuring out how to resolve them is smooth or at least mutually agreed upon.

So when I come to LessWrong, where the comments assume at-most-medium trust… where I’m not able to set a higher or different standard for a discussion beyond the lowest common denominator...

It’s really frustrating and sad, to have to choose between a public-untrusted and private-but-high-trust conversation.

It’s worth noting: I participate in multiple spaces that I trust *differently.* Maybe I wouldn’t recommend particular friends join Alice’s space because, while she’s good stating her clear reasons for things and evaluating evidence clearly and making sure others do the same, she’s *not* good at noticing when you’re triggered and pausing to check in if you’re okay.

And maybe Eve really needs that. That’s usually okay, because Eve can go to Bob’s space, or run her own.

Sometimes, Bob’s space doesn’t exist, and Eve lacks the skills to attract people to a new space. This is really important and sad. I personally expect LessWrong to contain a wide distribution of preferences that can support many needs, but it probably won’t contain something for *everyone.*

Still, I think it’s an overall better strategy to *make it easier to create new subspaces* than to try to accommodate everyone at once.

Getting Burned

I expect to get hurt sometimes.

I expect some friends (or myself) to not always be at our best. Not always self-aware enough to avoid falling into sociopolitical traps that pit us against each other.

I expect that at least some of the people I’m currently aligned with, I may eventually turn out to be unaligned with, and to come into conflict that can’t be easily resolved. I’ve had friendships that turned *weirdly and badly* adversarial and I spent months stressfully dealing with it.

But the benefits of high trust are so great that I don’t regret *for a second* having spent the first few years with those friends in a high-trust relationship.

I acknowledge that I am pretty privileged in having a set of needs and interpersonal preferences that are easier to fit into a high trust environment. There are people who just don’t interface well with the sort of spaces I thrive in, who may never get the benefits of high trust, and that... really sucks.

But the benefit of the Public Archipelago model is that there can be multiple subsections of the site with different norms. You can participate in discussions where you trust the space owner. Some authors may clearly spell out norms and take the time to clearly explain why they moderate comments, and maybe you trust them the most.

Some authors may not be willing to take that time. Maybe you trust them less, or maybe you know them well enough that you trust them anyhow.

In either case, you know what to expect, and if you’re not okay with it, you either don’t participate, or respond elsewhere, or put effort into understanding the author’s goals so that you are able to write critiques that they find helpful.

iii. The Fine Details

Okay, but can’t we at least require *reasons?*

I don’t think many people were resistant to deleting comments – the controversial feature was “delete without trace.”

First, spam bots, and dedicated adversaries with armies of sockpuppets make it at least necessary for this tool to be an *available* (LW2.0 has had posts with hundreds of spam or troll comments we quietly delete and IP ban)

For non-obvious spam…

I do hope delete without trace is used rarely (or that authors send the commenter a private reason when doing so). We plan to implement the moderation log Said Achmiz recommended, so that if someone is deleting a lot of comments without trace you can at least go and check, and notice patterns. (We may change the name to “delete and hide”, since *some* kind of trace will be available).

All things being equal, clear reasons are better than none, and more transparency is better than less.

But all things are not equal.

Moderation is *work.*

And I don’t think everyone understands that the amount of work varies a lot, both by volume, and by personality type.

Some people get energized and excited by reading through confrontational comments and responding.

Some people find it incredibly draining.

Some people get maybe a dozen comments on their articles a day. Some get barely any at all. But some authors get *hundreds*, and even if you’re the sort of person who *is* energized by it, there are only so many hours in a day and there are other things worth doing.

Some comments are not just mean or dumb, but immensely hateful and triggering to the author, and simply glancing at a reminder that it existed is painful – enough to undo the personal benefit they got from having written their article in the first place.

For many people, figuring out how to word a moderation notice is stressful, and I’m not sure whether it’s more intense on average to have to say:

> “Please stop being rude and obnoxiously derailing threads”

vs

> “I’m sorry, I know you’re trying your best, but you’re asking a lot of obvious questions and making subtly bad arguments in ways that soak up the other commenter’s time. The colleagues that I’m trying to attract to these discussion threads are tired of dealing with you.”

Not to mention that moderation often involves people getting angry at you, so you don't just have to come up with the initial posted reason, but also deal with a bunch of followup that can wreck your week. Comments that leave a trace invite people to *argue.*

Moderation can be tedious. Moderation can be stressful. Moderation is generally unpaid. Moderators can burn out or decide “you know what, this just isn’t worth the time and bullshit.”

And this is often the worst deal for the *best* authors, since the best authors attract more comments, and sometimes end up acquiring a sort of celebrity status where commenters don’t quite feel like they’re *people* anymore, and feel justified (or even obligated) to go out of their way to take them down a peg.

If none of this makes sense to you, if you can’t imagine moderating being this big a deal… well... all I can say is *it just really is a god damn big deal*. It really really is.

There is a tradeoff we have to make, one way or another, on whether we want to force our best authors to follow clear, legible procedures, or to write and engage more.

Requiring the former can (and has) ended up punishing the latter.

We prioritized building the delete-and-hide function because Eliezer asked for it and we wanted to get him posting again quickly. But he is not the only author to have asked and expressed appreciation for it.

Incentivizing Good Ideas and Good Criticism

I'll make an even stronger claim here: punishing idea generation is *worse* than punishing criticism.

You certainly need both, but criticism is *easier*. There might be environments where there isn't enough quantity or quality of critics, but I don't think LessWrong is one of them. Insofar as we *don't* have good enough criticism, it's because the critiques are nitpicky and unhelpful instead of trying to deeply understand unfamiliar ideas and collaboratively improve their load-bearing cruxes.

And meanwhile, I think the best critics also tend to be the best idea-generators – the two skills are in fact tightly coupled – so making LessWrong a place they feel excited to participate in seems very important.

It's possible to go too far in this direction. There are reasonable cases for making a different tradeoffs that different corners of the internet might employ. But our decision on LessWrong is that authors are not obligated to put in that work if it's stressful.

Overton Windows, and Personal Criticism

There's a few styles of comments that reliably make me go "ugh, this is going to become a mess and I really don't want to deal with it." Comments whose substance is *"this idea is bad, and should not be something LessWrong talks about."*

In that moment, the conversation stops being about whatever the idea was, and starts being about politics.

A recent example is what I'd call "fuzzy system 1 stuff." The Kensho and Circling threads felt like they were mostly arguing about "is it even okay to talk about fuzzy system 1 intuitions in rational discourse?". If you wanted to talk about the core ideas and how to use them effectively, you had to wade through a giant, sprawling demon thread.

Now, it's actually *pretty important* whether fuzzy system 1 intuitions have a place in rational discourse. It's a conversation that needs to happen, a question that probably has a *right answer* that we can converge on (albeit a nuanced one that depends on circumstances).

But right now, it seems like the *only* discussion that's possible to have about them is "are these in the overton window or not?". There needs to be space to explore ideas that aren't currently in the accepted paradigm.

I'd even claim that doing that productively is one of the things rationality is *for.*

Similar issues abound with critiquing someone's tone, or otherwise critiquing a *person* rather than an *idea*. Comments like that tend to quickly dominate the discussion and make it hard to talk about anything else. In many cases, if the comment were a *private* message, it could have been taken as constructive criticism instead of a personal attack that enflares people's tribal instincts.

For personal criticism, I think the solution is to build tools that make private discussion easier.

For Overton Window political brawls, I think the brawl itself is inevitable (if someone wants to talk about a controversial thing, and other people *don't* want them to talk about the controversial thing, you can't avoid the conflict). But I think it's reasonable for authors to say "if we're going to have the overton discussion, can we have it somewhere else? Right here, I'm trying to talk about the ramifications of X if Y is true."

Meanwhile, if you think X or Y are actively dangerous, you can still downvote their post. Instead of everyone investing endless energy in multiple demon threads, the issue can be resolved via a single thread, and the karma system.

I don't think this would have helped with the most recent thread, but it's an option I'd want available if I ever explored a controversial topic in the future.

iv. Towards Public Archipelago

This *is* a complicated topic, the decision *is* going to affect people. If you're the sort of person for whom the status quo seemed just perfect, your experience is probably going to become worse.

I *do* think that is sad, and it's important to own it, and apologize – I think having a place that felt safe and home and *right* become a place that feels alienating and wrong is in fact among the worst things that can happen to a person.

But the consequences of *not* making some major changes seem too great to ignore.

The previous iteration of LessWrong *died*. It depended on skilled writers continuously posting new content. It dried up as, one by one, as they decided LessWrong wasn't best place for them to publish or brainstorm.

There's a lot of reasons they made that choice. I don't know that our current approach will solve the problem. But I strongly believe that to avoid the same fate for LessWrong 2.0, it will need to be structurally different in some ways.

An Atmosphere of Experimentation

We have some particular tools, and plans, to give authors the same control they'd have over a private blog, to reduce the reasons to move elsewhere. This may or may not help. But beneath the moderation tools and Public Archipelago concept is an underlying approach of *experimentation.*

At a high level, the LessWrong 2.0 team will be experimenting with the site design. We want this to percolate through the site – we want authors to be able to experiment with modalities of discussion. We want to provide useful, flexible tools to help them do so.

Eventually we'd like users to experiment both with their overall moderation policy and culture, as well as the norms for individual posts.

Experiments I'd personally like to see:

- Posts where all commenters are required to fully justify their claims, such that complete strangers with no preconceptions can verify them
- Posts where all commenters are required to take a few ideas as given, to see if they have interesting implications in 201 or 301 concept space

- Discussions where comments must follow particular formats and will be deleted otherwise, such as the r/AskHistorians subreddit or stackoverflow.
- Discussions where NVC is required
- Discussions where NVC is banned
- Personal Blogposts where all commenters are only allowed to speak in poetry.
- Discussions where you need to be familiar with graduate level math to participate.
- Discussions where authors feel free to delete any comment that doesn't seem like it's pulling its attentional weight.
- Discussions where only colleagues the author personally knows and trusts get to participate.

Bubbling Up and Peer Review

Experimentation doesn't mean splintering, or that LessWrong won't have a central ethos connecting it. The reason we're allowing user moderation on Frontpage posts is that we *want* good ideas to bubble up to the top, and we don't want it to feel like a *punishment* if a personal blogpost gets promoted to Frontpage or Curated. If an idea (or discussional experiment) is *successful,* we want people to see it, and build off it.

Still, what sort of experimentation and norms to expect will vary depending on how much exposure a given post has.

On personal blogposts, pretty much anything goes.

On Frontpage posts, we will want to have *some* kind of standard, which I'm not sure we can formally specify. We're restricting moderation tools to users with high karma, so that only people who've already internalized what LessWrong is about have access to them. We want experimentation that *productively explores* rational-discussion-space. (If you're going to ask people to only comment in haiku on a frontpage post, you should have a pretty good reason as to why you think this will foster intellectual progress).

If you're deleting anyone who disagrees with you even slightly, or criticizing other users without letting them respond, we'll be having a talk with you. We may remove your mod privileges or restrict them to your personal blogposts.

Curated posts will (as they already do) involve a lot of judgment calls on the sitewide moderation team.

At some point, we might explore some kind of formal peer review process, for ideas that seem important enough to include in the LessWrong canon. But exploring that in full is beyond the scope of this post.

Norms for *this* comment section

With this post, I'm kinda intentionally summoning a demon thread. That's okay. This is the official "argue about the moderation overton window changing" discussion space.

Still, some types of arguing seem more productive than others. It's especially important for *this particular conversation* to be maximally transparent, so I won't be deleting anything except blatant trolling. Comments that are exceptionally hostile, I might comment-lock, but leave visible with an explicit reason why.

But, if you want your comments or concerns to be useful, some informal suggestions:

Failure modes to watch out for:

- If the Public Archipelago direction seems actively dangerous or otherwise awful, *try to help solve the underlying problem.* Right now, one of the most common concerns

we’ve heard from people who we’d like to be participating on LessWrong is that the comments feel nitpicky, annoying, focused on *unhelpful* criticism, or unsafe. If you’re arguing that the Archipelago approach is fundamentally flawed, you’ll need to address this problem in *some* fashion. Comments that don’t at least acknowledge the magnitude of the tradeoff are unlikely to be persuasive.
- If other commenters seem to have vastly different experiences than you, try to proactively understand them – solutions that don’t take into account diversity of experience are less useful.

Types of comments I expect to be especially useful:

- **Considerations we’ve missed.** This *is* a fairly major experiment. We've tried to be pretty thorough about exploring the considerations here, but there are probably a lot o we haven’t thought of.
- **Pareto Improvements.** I expect there are a lot of opportunities to avoid making tradeoffs, instead finding third-options that get as many different benefits as once.
- **Specific tools you’d like to see.** Ideally, tools that would enable a variety of experiments while ensuring that good content still gets to bubble up.

…

Ok. That was a bit of a journey. But I appreciate you bearing with me, and am looking forward to having a thorough discussion on this.