A
C
D
B
?

Keep your beliefs cruxy and your frames explicit

What product are you building?

Epistemic status: It appears like it works for me. Not meant to be a hard/fast rule.

A frame I frequently examine conversations through is "are we building a product together, or not?"

Many conversations (online or in person) are "just sorta hanging out." Just-sorta-hanging-out can be quite important, both because it's fun, and as part of building up friendships, etc. But if one or more participants *isn't* having fun, it's more likely that the conversation should end, or change.

"Building a product" conversations have a goal, and the goal is "create something that *someone* is going to use someday." Here's a few types of products you might be building:

- A literal product
    - i.e. programming a website or designing a widget
- A relationship
    - you and a friend, or a romantic partner, are trying to improve the foundations of trust or communication
- A party
    - you and your roommates are planning a surprise birthday bash for a friend
- A new set of norms
    - you and your friends, or your coworkers, all agree that something about the current social equilibrium is off, and should be replaced by something new
- A fun idea
    - you're doing random brainstorming on something kinda crazy, but you still want it to be the *best* something-kinda-crazy that it can be. Maybe it'll later turn into a literal product, or maybe it'll turn into a fun story to tell.
- A felt sense
    - Maybe the object level discussion isn't about the product. Maybe you're exploring a philosophical idea. But your goal isn't to output a useful philosophical idea. Your goal might instead be:
        - develop a sense of what a particular flavor of confusion feels like
        - develop some shared intuitions that *underlie* the philosophical idea

In some of those cases, there are "just hanging out" conversations you might also have about it. You can talk about "a new set of norms" that's less goal directed and more just meandering and exploratory. A fun idea can certainly be nothing more than "a fun idea."

But, I find product-building conversations way more interesting, most of the time. In particular, I'm willing to invest a lot more effort into a conversation if it seems like it's about building something.

Product-building-conversations don't have to be *rushed,* or laser-focused. Sometimes the best way to design a product is to have a long, meandering conversation that gives you time to soak in each facet of the design constraints, or share intuitions about it.

Building a product imposes constraints on the conversation

There are lots of different conversational styles you can be building a product in, depending on your culture. Gruff auto-mechanics might brusquely swear at each other when they screw something up. Some companies might have weird politeness norms, and those norms might be different in the US vs Japan while designing their latest widget.

I discussed some examples recently of (what I saw as) product-building in a collaborative frame. My comments then were highlighting my own preferred norms. But in this post now I'm trying to make a broader point.

I have best guesses about what sort of norms make for productive product building. But I have a much more important meta-level belief: If you're building a product together, there are *some* kinds of constraints on "what sort of conversation is actually worth having?". And one of the most important constraints is:

If you're *not building the same product*, you're going to have a bad time.

When a conversation seems to be getting confused, and people talking past each other, some questions worth asking might be:

- Are we building a product, or just sorta hanging out?
- If we're just sorta hanging out, are we having fun, or otherwise getting value? If not, stop, or change something about the conversation.
- If we *are* building a product, do we have agreement about what product we're building? Do we have meta-level agreement that, if we *don't* agree on what product we're building, at least one of our goals should be to figure that out?

For me, at least, that last question dictates whether this is a conversation that I'm going to put a particular kind of serious effort into.

Doublecrux is for Building Products

Previously:

- What product are you building?

2 years ago, CFAR's doublecrux technique seemed "probably good" to me, but I hadn't really stress tested it. And it was particularly hard to learn in isolation without a "real" disagreement to work on.

Meanwhile, some people seemed skeptical about it, and I wasn't sure what to say to them other than "I dunno man this just seems obviously good? Of *course* you want to treat disagreements like an opportunity to find truth together, share models, and look for empirical tests you can run?"

But for people who didn't share that "of course", that wasn't very helpful.

For the past two years I've worked on a team where big disagreements come up pretty frequently, and where doublecrux has been more demonstrably helpful. I have a clearer sense of where and when the technique is important.

Intractable Disagreements

Some intractable disagreements are fine

If you disagree with someone on the internet, or a random coworker or something, often the disagreement doesn't matter. You and your colleague will go about their lives, one way or another. If you and your friends are fighting over "Who would win, Batman or Superman?", coming to a clear resolution just isn't the point.

It might also be that you and your colleague are doing some sort of coalition-politics fight over the overton window, and most of the debate might be for the purpose of influencing the public. Or, you might be arguing about the Blue Tribe vs Red tribe as a way of signaling group affiliation, and earnestly understanding people isn't the point.

This makes me sad, but I think it's understandable and sometimes it's even actually important.

Such conversations are don't *need* to be doublecrux shaped, unless both participants want them to be.

Some disagreements are *not* fine

When you're building a product together, it actually matters that you figure out how to resolve intractable disagreements.

I mean "product" here pretty broadly – anything that somebody is actually going to use. It could be a literal app or widget, or an event, or a set of community norms, or a philosophical idea. You might literally sell it or just use it yourself. But I think there is something helpful about the "what if we were coworkers, how would we resolve this?" frame.

The important thing is "there is a group of people collaborating on it" and "there is a stakeholder who cares about it getting built."

If you're building a website, and one person thinks it should present all information very densely, and another person thinks it should be sleek and minimalist... *somehow* you need to actually decide what design philosophy to pursue. Options include (not necessarily limited to)

- Anarchy
- One person is in charge
- Two or more people come to consensus
- People have domain specializations in which one person is in charge (or gets veto power).

Anarchy

To start with, what's wrong with the "everyone just builds what seems right to them and you hope it works out" option? Sometimes you're building a bazaar, not a cathedral, and this is actually fine. But it often results in different teams building different tools at cross purpose, wasting motion.

One person in charge?

In a hierarchical company, maybe there's a boss. If the decision is about whether to paint a bikeshed red or blue, the boss can just say "red", and things move on.

This is less straightforward in the case of "minimalism" vs "high information density."

First, is the boss even doing any design work? What if the boss and the lead designer disagree about aesthetics? If the lead designer hates minimalism they're gonna have a bad time.

Maybe the boss trusts the lead designer enough to differ to them on aesthetics. Now the lead designer is the decision maker. This is an improvement, but just punts the problem down one level. If the lead designer is Just In Charge, a few things can still go wrong:

Other workers don't actually understand minimalism

"Minimalist websites" and "information dense websites" are designed very differently. This filters into lots of small design decisions. Sometimes you can solve this with a comprehensive style guide. But those are a lot of work to create. And if you're a small startup (or a small team within a larger company), you may not have have the resources for that. It'd be nice if your employees just *actually understood* minimalism so they could build good minimalist components.

The lead designer is wrong

Sometimes the boss's aesthetic isn't locally optimal, and this actually needs to be pointed out. If lead-designer Alice says "we're building a minimalist website" it might be important for another engineer or designer to say "Alice, you're making weird tradeoffs for minimalism that are harming the user experience."

Alice might think "Nah, you're wrong about those tradeoffs. Minimalism is great and history will bear me out on this." But Alice might also respect Bob's opinion enough to *want* to come to some kind of principled resolution. If Bob's been right about similar things before, what should Alice and Bob do, if Alice *wants* to find out she's wrong – *if and only if she's actually wrong*, and that her minimalist aesthetic is harming the user experience.

The lead designer is right, but other major stakeholders think she's wrong

Alternately, maybe Bob thinks Alice is making bad design calls, but Alice is actually just making the right calls. Bob has rare preferences that don't overlap much with the average user, that *shouldn't* necessitate a major design overhaul.

Initially, this will look the same to both parties as the previous option.

If Alice has listened to Bob's complaints a bunch, and Alice generally respects Bob but thinks he's wrong here, at some point she needs to say "Look Bob, we just need to actually build the damn product now, we can't rehash the minimalism argument every time we build a new widget."

I think it's useful for Bob to gain the skill of saying "Okay. fine." Let go of his frustration and embrace the design paradigm.

But that's a tough skill. And meanwhile, Bob is probably going to spend a fair amount of time and energy *being annoyed* about having to build a product they're less excited about. And sometimes, Bob's work is less efficient because he doesn't understand minimalism and keeps building site-components subtly incompatible with it.

What if there was a process by which *either* Alice would update *or* Bob would update, that both Alice and Bob considered fair?

You might just call that process "regular debate." But the problem is that regular debate just *often doesn't work.* Alice says "We need X, because Y". Bob says "No, we need A, because B", and they somehow both repeat those points over and over without ever changing each other's mind.

This wastes loads of time, which could have been better spent building new site features if they were able to do it faster.

Even if Alice is in charge and gets final say, it's still suboptimal for Bob to have lower morale and keep making subtly wrong widgets.

And even if Bob understands that Alice is in charge, it might still be suboptimal for Bob to feel like Alice never *really* understood exactly what Bob's concerns were.

What if there's no boss?

Maybe your "company" is just two friends in a basement doing a project together, and there isn't really a boss. In this case, the problem is much sharper – somehow you need to actually make a call.

You might solve this by deciding to appoint a decision-maker – change the situation from a "no boss" to "boss" problem. But if you were just two friends making a game together in their spare time, for fun, this might kinda suck. (If the whole point was to make it together as friends, a hierarchical system may be fundamentally un-fun and defeat the point)

You might be doing a more serious project, where you agree that it's important to have clear coordination protocols and hierarchy. But it nonetheless feels premature to commit to "Alice is always in charge of design decisions." Especially if Bob and Alice both have reasonable design skills. And especially if it's early on in the project and they haven't yet decided what their product's design philosophy should be.

In that case, you can start with straightforward debate, or making a pros/cons list, or exploring the space a bit and hoping you come to agreement. But if you're not coming to agreement... well, you need to do *something.*

If "regular debate" is working for you, cool.

If "just talking about the problem" is working, obviously you don't have an issue. Sometimes the boss actually just says "we're doing it this way" and it doesn't require any extensive model sharing.

If you've never run into the problem of intractable-disagreement while collaborating on something important, this blogpost is not for you. (But, maybe keep it in the back of your mind in case you *do* run into such an issue)

But working on the LessWrong team for about 1.5 years, I've run into numerous deep disagreements, and my impression is that such disagreements are common – especially in domains where you're solving a novel problem. We've literally argued a bunch about minimalism, which isn't an especially unique design decision. We've also had much weirder disagreements about integrity and intellectual progress and AI timelines and more.

We've resolved many (although not all) of those disagreements. In many cases, doublecrux has been helpful as a framework.

What's Doublecrux again?

If you've made it this far, presumably it seems useful to have *some* kind of process-for-consensus that works better than whatever you and your colleagues were doing by default.

Desiderata that I personally have for such a process:

- Both parties can agree that it's worth doing
- It should save more time than it costs (or produce value commensurate with the time you put in)
- It works even when both parties have different frames or values
- If necessary, it untangles confused questions, and replaces them with better ones
- If necessary, it untangles confused goals, and replaces them with better ones
- If people are disagreeing because of aesthetic differences like "what it beautiful/good/obviously-right", it provides a framework wherein people can actually change their mind about "what is beautiful and good and right."

- Ultimately, it lets you "get back to work", and actually build the damn product, confident that you are going about it the right way.

[Many of these goals were not assumptions I started with. They're listed here because I kept running into failures relating to each one. Over the past 2 years I've had some success with each of those points]

Importantly, it's not necessarily needed for such a process to answer the original question you asked. In the context of building a product, what's important is that you figure out a model of the world which you both agree on, which informs which actions to take.

Doublecrux is a framework that I've found helpful for the above concerns. But I think I'd consider it a win for this essay if I've at least clarified why it's desirable to have *some* such system. I share Duncan's belief that it's more promising to repair or improve doublecrux than to start from scratch. But if you'd rather start from scratch, that's cool.

Components of Doublecrux – Cognitive Motions vs Attitudes

There are two core concepts behind the doublecrux framework:

- *A set of cognitive motions:*
    - Looking for the cruxes of your beliefs, and asking what empirical observations would change your mind about them. (Recursing until you find a crux you and your partner both share, the "doublecrux")
- *A set of attitudes*
    - Epistemic humility
        - "*maybe I'm the wrong one*"
    - Good faith
        - "*I trust my partner to be cooperating with me*"
    - Belief that objective reality is real
        - "*there's an actual right answer here, and it's better for each of us if we've both found it*"
    - Earnest curiosity

Of those, I think the set of attitudes is more important than the cognitive motions. If the "search for cruxes and empirical tests" thing isn't working, but you have the four attitudes, you can probably find other ways to make progress. Meanwhile, if you don't each have those four attitudes, you don't have the foundations necessary to doublecrux.

Using language for truthseeking, not politics

But I think the cognitive motions are helpful, for this reason: much of human language is by default *politics* rather than *truthseeking*. "Regular debate" often reinforces the use of language-as-politics, which actives brain modules that are optimizing to win, which involves strategic blindness. (I mean something a bit nuanced by "politics" here, beyond scope of this post. But basically, optimizing beliefs and words for how you fit into the social landscape, rather than optimizing for what corresponds to objective reality).

The "search for empirical tests is and cruxes-of-beliefs" motion is designed to keep each participant's brain in a "language-as-truthseeking" mode. If you're asking yourself "why would *I* change my mind?", it's more natural to be honest to yourself and your partner than if you're asking "how can I change *their* mind?"

Meanwhile, the focus on mutual, opposing cruxes keeps things *fruitful*. Disagreement is more interesting and useful than agreement – it provides an opportunity to actually learn. If people are doing language-as-politics, then disagreement is a red flag that you are on opposing sides and might be threatening each other (which might either prompt you to fight, or prompt you to "agree to disagree", preserving the social fabric by sweeping the problem under the rug).

But *if* you can both trust that everyone's truthseeking, then you can drill directly into disagreements without worrying about that, optimizing for learning, and then for building a shared model that lets you actually make progress on your product.

Trigger Action Plans

Knowing this is all well and good, but what might this translate into in terms of actions?

If happen to have a live disagreement *right now*, maybe you can try doublecrux. But if not, what circumstances should prompt

I've found the "Trigger Action Plan" framework useful for this sort of thing, as a basic rationality building-block skill. If you notice an unhelpful conversational pattern, you can build an association where you take some particular action that seems useful in that circumstance. (Sometimes, the generic trigger-action of "notice something unhelpful is happening ----> stop and *think*" is good enough)

In this case, a trigger-action I've found useful is:

TRIGGER: Notice that we've been arguing awhile, and someone has just repeated the same argument they said a little while ago (for the second, or especially third time)

ACTION: Say something like: "Hey, I notice that we've been repeating ourselves a bit. I feel like conversation is kinda going in circles...." followed by either "Would you be up for trying to formally doublecrux about this?" or following Duncan's vaguer suggestions about how to unilaterally improve a conversation (depending on how much shared context you and your partner have).

Summary

- Intractable disagreements don't always matter. But if you're trying to build something together, and disagreeing substantially about how to go about it, you will need some way to resolve that disagreement.
- Hierarchy can obviate the need for resolution *if* the disagreement is simple, and if everyone agrees to respect the boss's decision.
- If the disagreement has persisted awhile and it's still wasting motion, at the very least it's probably useful to do *something* differently. In particular, if you've been repeating the
- Doublecrux is a particular framework I've found helpful for resolving intractable disagreements (when they are important enough to invest serious energy and time into). It focuses the conversation into "truthseeking" mode, and in particular strives to avoid "political mode"

Keep Your Beliefs Cruxy

Previously in sequence: Doublecrux is for building products.

To recap – you might want to doublecrux if either:

- You're building a product, metaphorical or literal, and you disagree about how to proceed.
- You want to make your beliefs more accurate, and you think a particular person you disagree with is likely to have useful information for you.
- You just... enjoy resolving disagreements in a way that mutually pursues truth for whatever reason.

Regardless, you might find yourself with the problem:

Doublecruxing takes a lot of time.

For a 'serious' disagreement, it frequently takes a least an hour, and often much longer. Habryka and I once took 12 hours over the course of 3 days to make any kind of progress on a particularly gnarly disagreement. And sometimes disagreements can persist for years despite significant mutual effort.

Now, doublecruxing is *faster* than many other forms of truth-aligned-disagreement resolution. I actually it's helpful to think of doublecrux as "the fastest way for two disagreeing-but-honest-people to converge locally towards the truth", and if someone came up with a faster method, I'd recommend deprecating doublecrux in favor it that. (Meanwhile, doublecrux is not guaranteed to be faster for 3+ people to converge but I still expect it to be faster for smallish groups with particularly confusing disagreements)

Regardless, multiple hours is a long time. Can we do better?

I think the answer is yes, and it basically comes in the form of:

- Practice finding your own cruxes
- Practice helping other people find their cruxes
- Develop metacognitive skills that make cruxfinding *natural and intuitive*
- Caching the results into a clearer belief-network

Those are all things you can do unilaterally. If you get buy-in from your colleagues, you might also try something like "develop culture that encourages people to do those four things, and help each other to do so."

I'd summarize all of that as "develop the skill and practice of *keeping your beliefs cruxy.*"

By default, humans form beliefs for all kinds of reasons, without regard for how falsifiable they are. The result is a tangled, impenetrable web. Productive disagreement takes a long time because people are starting from the position of "impenetrable web."

If you make a habit of asking yourself "what observations *would* change my mind about this?", then you gain a few benefits.

First, your beliefs should (hopefully?) be more entangled with reality, period. You'll gain the skill of noticing how your beliefs should constrain your anticipations, and then if they fail to do so, you can maybe update your beliefs.

Second, if you've cultivated that skill, then during a doublecrux discussion, you'll have an easier time engaging with the core doublecrux loop. (So, a conversation that might have taken an hour takes 45 minutes – your conversation partner might still take a long time to figure out their cruxes, but maybe you can do your own much faster)

Third, once you gotten into this habit, this will help your beliefs form in a cleaner, more reality-entangled fashion *in the first place.* Instead of building an impenetrable morass, you'll be building a clear, legible network. (So, you might have all your cruxes full accessible from the beginning of the conversation, and then it's just a matter of stating them, and then helping your partner to do so)

[Note: I don't think you should *optimize directly* for your beliefs being legible. This is a recipe for burying illegible parts of your psyche and missing important information. But rather, if you try to actually understand your beliefs and what causes them, the legibility will come naturally as a side benefit**]**

Finally, if everyone around you is doing, this radically lowers the cost of productive-disagreement. Instead of taking an hour (or three days), as soon as you bump into an important disagreement you can quickly navigate through your respective belief networks, find the cruxes, and skip to the part where you actually Do Empiricism.

I think (and here we get into speculation), that this can result in a phase shift in how disagreement works, enabling much more powerful discourse systems than we currently have.

I think keeping beliefs cruxy is a good example of a practice that is *both* a valuable "Rabbit Strategy", as well as something worth Stag Hunting Together on. (In Rabbit/Stag parlance, a Rabbit strategy is something you can *just do* without relying on anyone else to help you, that will provide benefit to you. A Stag strategy is something with a larger payoff, but only actually works if you've coordinated with people to do so)

If you have an organization, community, or circle of friends where many people have practiced keeping-beliefs-cruxy, I predict you will find individuals benefiting, as well as creating a truthseeking culture more powerful than the sum of its parts.

Hopefully obvious addenda, operationalizing:

My crux for all this is that I think this is locally tractable.

My recommendation is not that you go out of your way to practice this all the time – rather, that you spend some time on-the-margin practicing the "look-for-cruxes" whenever you actually have a noteworthy disagreement.

If you practice crux-finding whenever you have such a disagreement, and find that it isn't locally useful, I probably wouldn't recommend continuing.

If 10 people tried this and 6 of them told me it didn't feel useful, even superficially, I'd probably change my mind significantly. If they tried it continuously for a year and there weren't at least 3 of them that could point to transparently useful outcomes, (such as disagreements taking much less time than they'd naively predict), I'd also significantly change my beliefs here.

This operationalization has some important caveats, such as "I don't necessarily expect this to work for people who haven't read the sequences or some equivalent. I also think it might require some skills that I haven't written up explanations for yet – which I hope to do soon as I continue this sequence."

Noticing Frame Differences

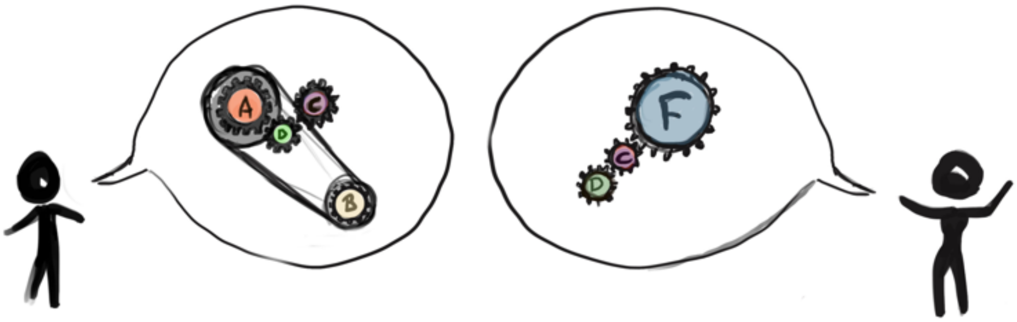
Previously: Keeping Beliefs Cruxy

When disagreements persist despite lengthy good-faith communication, it may not just be about factual disagreements – it could be due to people operating in entirely different frames — different ways of *seeing*, *thinking* and/or *communicating*.

If you can’t notice when this is happening, or you don’t have the skills to navigate it, you may waste a lot of time.

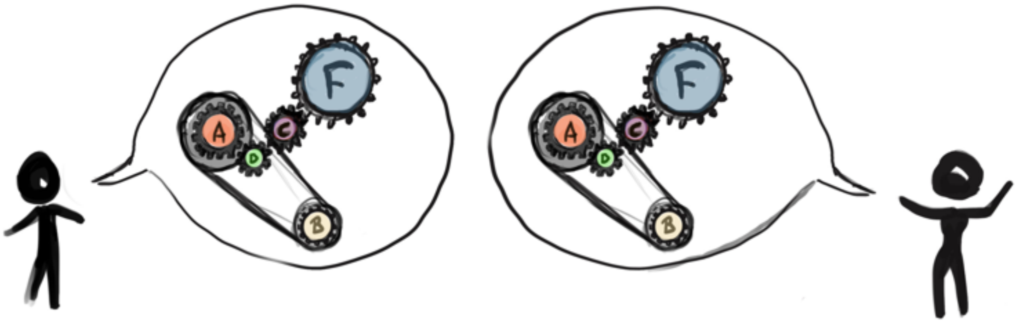
Examples of Broad Frames

Gears-oriented Frames



Bob and Alice’s conversation is about cause and effect. Neither of them are planning to take direct actions based on their conversation, they’re each just interested in understanding a particular domain better.

Bob has a model of the domain that includes gears A, B, C and D. Alice has a model that includes gears C, D and F. They’re able to exchange information, and their information is compatible,and they each end up with a shared model of how something works.



There are other ways this could have gone. Ben Pace covered some of them in a sketch of good communication:

- Maybe they discover their models don't fit, and one of them is wrong
- Maybe combining their models results in a surprising, counterintuitive outcome that takes them awhile to accept.
- Maybe they fail to integrate their models, because they were working at different levels of abstraction and didn't realize it.

Sometimes they might fall into subtler traps.

Maybe the thing Alice is calling "Gear C" is actually different from Bob's "Gear C". It turns out that they were using the same words to mean different things, and even though they'd both read blogposts warning them about that they didn't notice.

So Bob tries to slot Alice's gear F into his gear C and it doesn't fit. If he doesn't already have reason to trust Alice's epistemics, he may conclude Alice is crazy (instead of them referring to subtly different concepts).

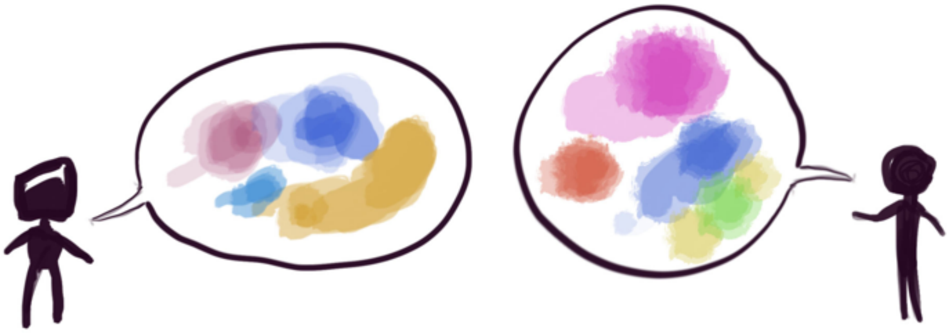
This may cause confusion and distrust.

But, the point of this blogpost is that *Alice and Bob have it easy.*

They're actually trying to have the same conversation. They're both trying to exchange explicit models of cause-and-effect, and come away with a clearer understanding of the world through a reductionist lens.

There are many other frames for a conversation though.

Feelings-Oriented Frames



Clark and Dwight are exploring how they feel and relate to each other.

The focus of the conversation might be navigating their particular relationship, or helping Clark understand why he's been feeling frustrated lately

When the Language of Feelings justifies itself to the Language of Gears, it might say things like: "Feelings are important information, even if it's fuzzy and hard to pin down or build explicit models out of. If you don't have a way to listen and make sense of that information, your model of the world is going to be impoverished. This involves sometimes looking at things through lenses *other* than what you can explicitly verbalize."

I think this is true, and important. The people who do their thinking through a gear-centric frame *should* be paying attention to feelings-centric frames for this reason. (And meanwhile, *feelings themselves* totally have gears that can be understood through a mechanistic framework)

But for many people that's *not actually the point* when looking through a feelings-centric frame. And not understanding this may lead to further disconnect if a Gearsy person and a Feelingsy person are trying to talk.

"Yeah feelings are information, but, also, like, man, you're a human being with all kinds of fascinating emotions that are an important part of who you are. This is super interesting! And there's a way of making sense of it that's necessarily *experiential* rather than about explicit, communicable knowledge."

Frames of Power and Negotiation

Dominance and Threat

Erica is Frank's boss. They're discussing whether the project Frank has been leading should continue, or whether it should stop and all the people on Frank's team reassigned.

Frank argues there's a bunch of reasons his project is important to the company (i.e. it provides financial value). He also argues that it's good for morale, and that cancelling the project would make his team feel alienated and disrespected.

Erica argues back that there are other projects that are more financially valuable, and that his team's feelings aren't important to the company.

It so happens that Frank had been up for a promotion soon, and that would put him (going forward) on more even footing with Erica, rather than her being his superior.

It's not (necessarily) about the facts, or feelings.

If Alice and Bob wandered by, they might notice Erica or Frank seeming to make somewhat basic reasoning mistakes about how much money the project would make or why it was valuable. Naively, Alice might point out that they seem to be engaging in motivated reasoning.

If Clark or Dwight wandered by, they might notice that Erica doesn't seem to really be engaging with Frank's worries about team morale. Naively, Clark might say something like "Hey, you don't seem to really be paying attention to what Frank's team is experiencing, and this is probably relevant to actually having the company be successful."

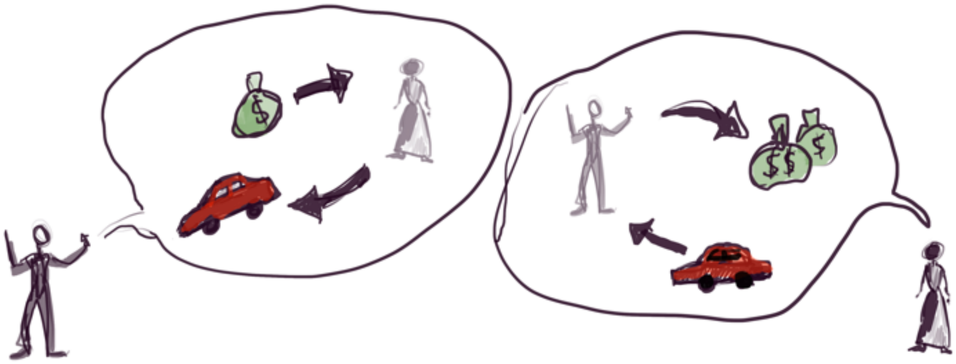
But the conversation is not about sharing models, and it's not about understanding feelings. It's not even necessarily about "what's best for the company."

Their conversation is a negotiation. For Erica and Frank, most of what's at stake are their own financial interests, and their social status within the company.

The discussion is a chess board. Financial models, worker morale, and explicit verbal arguments are more like game pieces than anything to be taken at face value.

This might be fully transparent to both Erica and Frank (such that neither even considers the other deceptive). Or, they might both earnestly believe what they're saying – but nonetheless, if you try to interpret the conversation as a practical decision about what's best for the company, you'll come away confused.

The Language of Trade



George and Hannah are negotiating a trade.

Like Erica and Frank, this is ultimately a conversation about what George and Hannah want.

A potential difference is that Erica and Frank might think of their situation as zero-sum, and therefore most of the resolution has more to do with figuring out “who would win in a political fight?”, and then having the counterfactual loser back down.

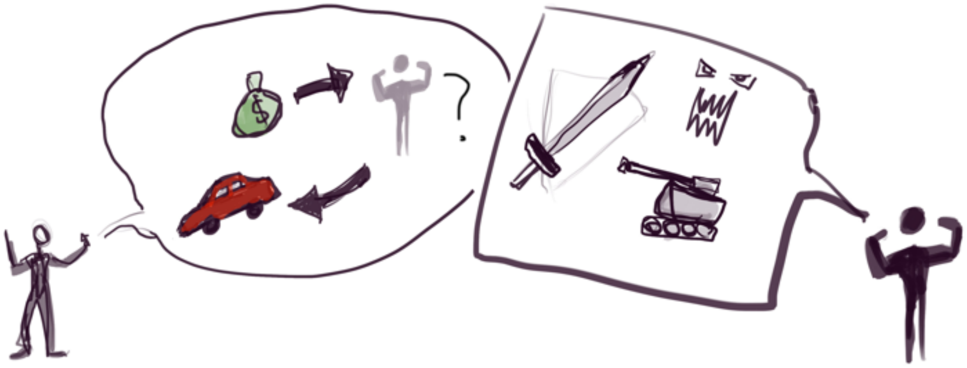
Whereas George/Hannah might be actively looking for positive sum trades, and in the event that they can’t find one, they just go about their lives without getting in each other’s way.

(Erica and Frank might also look for opportunities to trade, but doing so honestly might first require them to establish the degree to which their desires are mutually incompatible and who would win a dominance contest. Then, having established their respective positions, they might speak plainly about what they have to offer each other)

Noticing obvious frame differences

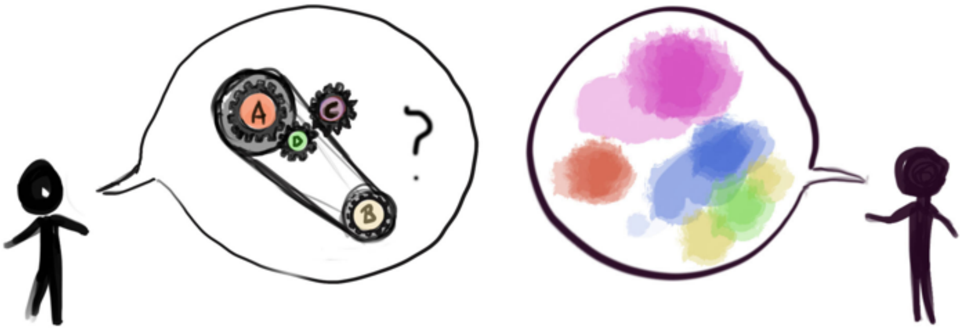
So the first skill here, is noticing when you’re having wildly different expectations about what sort of conversation you’re having.

If George is looking for a trade and Frank is looking for a fight, George might find himself suddenly bruised in ways he wasn’t prepared for. And/or, Frank might have randomly destroyed resources when there’d been an opportunity for positive sum interaction.



Or: If Dwight says “I’m feeling so frustrated at work. My boss is constantly belittling me”, and then Bob leaps in with an explanation of why his boss is doing that and maybe trying to fix it…

Well, this one is at least a stereotypical relationship failure mode you’ve probably heard of before (where Dwight might just want validation).



Untangling Emotions, Beliefs and Goals

A more interesting example of Gears-and-Feelings might be something like:

Alice and Dwight are talking about what career options Dwight should consider. (Dwight is currently an artist, not making much money, and has decided they want to try something else)

Alice says “Have you considered becoming a programmer? I hear they make a lot of money and you can get started with a 3 month bootcamp.”

Dwight says “Gah, don’t talk to me about programming.”

It turns out that Dwight’s dad always pushed him to learn programming, in a fairly authoritarian way. Now Dwight feels a bunch of ughiness around programming, with a mixture of “You’re not the boss of me! I’mma be an artist instead!”

In this situation, perhaps the best option might be to say: “okay, seems like programming isn’t a good fit for Dwight,” and move on.

But it might also be that programming is actually a good option for Dwight to consider… it’s just that the conversation can’t proceed in the straightforward cost/benefit analysis frame that Alice was exploring.

Dwight making meaningful updates on whether programming is good for him depends on untangling his emotions, and/or exploring the relationship between his explicit models and his messier internals. It might require making piece with some longstanding issues with his father, or learning to detach them from the “should I be a programmer” question.

It might be that the most useful thing Alice can do is give him the space to work through that on his own.

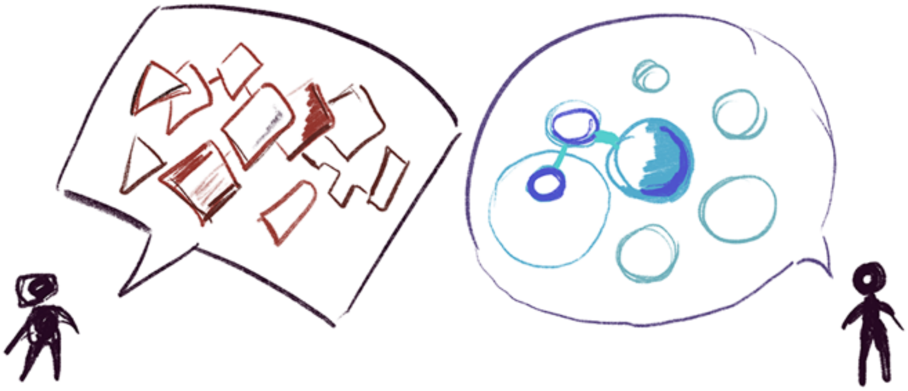
If Dwight trusts Alice to shift into a feelings-oriented framework (or a framework that at least includes feeling), Alice might be able to directly help him with the process.

It may also be that this prerequisite trust doesn't exist, or that Dwight just doesn't want to have this conversation, in which case it's probably just best to move on to another topic.

Subtle differences between frames

This gets much more complicated when you observe that a) there’s lots of slight variations on frames, and b) many people and conversations involve a mixture of frames.

It’s not *that* hard to notice that one person is in a feelings-centric frame while another person is in a gears-centric frame. But things can actually get even more confusing if two people share a broad frame (and so *think* they should be speaking the same language), but actually they’re communicating in two different subframes.



Example differences between gears-frames

Consider variations of Alice and Bob – both focused on causal models – who are coming from these different vantage points:

Goal-oriented vs Curiosity-driven conversation

Alice is trying to solve a specific problem (say, get a particular car engine fixed), and Bob thinks they're just, like, having a freewheeling conversation about car engines and how neat they are (and if their curiosity took them in a different direction they might shift the conversation towards something that had nothing to do with car engines).

Debate vs Doublecrux

Alice is trying to present arguments for her side, and expects Bob to refute those arguments or present different arguments. The burden of presenting a good case is on Bob.

Whereas Bob thinks they're trying to mutually converge on true beliefs (which might mean adopting totally new positions, and might involve each person focusing on how to change their *own* mind rather than their partner's)

Specific ontologies

If one person is, say, really into economics, then they might naturally frame everything in terms of transactions. Someone else might be really into programming and see everything as abstracted functions that call each other.

They might keep phrasing things in terms that fit their preferred ontology, and have a hard time parsing statements from a different ontology.

Example differences between feelings-frames

"Mutual Connection" vs "Turn Based Sharing"

Clark might be trying to share feelings for the sake of building connection (sharing back and forth, getting into a flow, getting resonance).

Whereas Dwight might think the point is more for each of them to fully share their own experience, while the other one listens and takes up as little space as possible.

"I Am My Feelings" vs "My Feelings are Objects"

Clark might highly self identify with his feelings (in a sort of Romantic framework). Dwight might care a lot about understanding his feelings but see them as temporary objects in his experience (sort of Buddhist)

Concrete example: The FOOM Debate

One of my original motivations for this post was the Yudkowsky/Hanson Foom Debate, where much ink was spilled but AFAICT neither Yudkowsky nor Hanson changed their mind much.

I recently re-read through some portions of it. The debate seemed to feature several of the "differences within gears-orientation" listed above:

Specific ontologies: Hanson is steeped in economics and sees it as the obvious lens to look at AI, evolution and other major historical forces. Yudkowsky instead sees things through the lens of optimization, and how to develop a causal understanding of what recursive optimization means and where/whether we've seen it historically.

Goal vs Curiosity: I have an overall sense that Yudkowsky is more action oriented – he's specifically setting out to figure out the most important things to do to influence the far future. Whereas Hanson mostly seems to see his job as "be a professional economist, who looks at various situations through an economic lens and see if that leads to interesting insights."

Discussion format: Throughout the discussion, Hanson and Yudkowsky are articulating their points using very different styles. On my recent read-through, I was impressed with the degree and manner to which they discussed this explicitly:

Eliezer notes:

> I think we ran into this same clash of styles last time (i.e., back at Oxford). I try to go through things systematically, locate any possible points of disagreement, resolve them, and continue. You seem to want to jump directly to the disagreement and then work backward to find the differing premises. I worry that this puts things in a more disagreeable state of mind, as it were—conducive to feed-backward reasoning (rationalization) instead of feed-forward reasoning.
>
> It's probably also worth bearing in mind that these kinds of metadiscussions are important, since this is something of a trailblazing case here. And that if we really want to set up conditions where we can't agree to disagree, that might imply setting up things in a different fashion than the usual Internet debates.

Hanson responds:

> When I attend a talk, I don't immediately jump on anything a speaker says that sounds questionable. I wait until they actually make a main point of their talk, and then I only jump on points that seem to matter for that main point. Since most things people say actually don't matter for their main point, I find this to be a very useful strategy. I will be very surprised indeed if everything you've said mattered regarding our main point of disagreement.

I found it interesting that I find both these points quite important – I've run into each failure mode before. I'm unsure how to navigate between this rock and hard place.

My main goal with this essay was to establish frame-differences as an important thing to look out for, and to describe the concept from enough different angles to (hopefully) give you a *general* sense of what to look for, rather than a single failure mode.

What to *do* once you notice a frame-difference depends a lot on context, and unfortunately I'm often unsure what the best approach is. The next few posts will approach "what has sometimes worked for me", and (perhaps more sadly) "what hasn't."

Picture Frames, Window Frames and Frameworks

Some commenters on Noticing Frames were confused about what I actually meant by "frame." I defined it briefly as "different ways of seeing, thinking and communicating", but this was a bit vague.

I'm definitely using the word "frame" as a broad metaphor, rather than a concrete specific phenomenon. I'm not at a point where *I'm* sure I have a more concrete phenemonon to explicitly describe. I recognize this as an important red flag – specificity is good, and I'll try to get more concrete in future posts. For now, though, I think the broad metaphor is useful, and want to explain it slightly better.

After reflecting on the feedback, I realized it was actually three different metaphors:

- **Picture Frames** (ways of communicating)
- **Window Frames** (ways of seeing)
- **Frameworks** (ways of thinking)

Upon reflection, I endorse using all three metaphors fuzzily rolled into one.

The past few years, I kept thinking I had figured out why resolving disagreements was hard, and kept being surprised by *new* ways to subtly miss each other. At first I thought the main reasons to disagree were "*different beliefs*", "*different values*", and "*confusion over beliefs/values.*" This isn't wrong, but it wasn't nuanced enough for me to notice all the disconnects.

I've come to believe there's an important general skill, which goes something like:

> **Trigger:** Notice when a conversation is going nowhere, especially over the timescale of months/years.
>
> **Action:** Look for frame differences *outside of the ones you know to make sense of,* and figure out how to make sense of them.

That's a hell of a vague action, I admit. It's composed of smaller, more specific actions (which are probably easier to learn). I was worried that if I named the most specific actions I could point at, people would focus on that and lose sight of the more general problem. I'm not even certain that the "different ways of seeing, thinking, and communicating" schema is complete – it's meant to be illustrative rather than comprehensive.

Someday, ideally, there'll be a "how to" doublecrux sequence that teaches concrete skills and exercises that are each immediately useful. (I think Eli might be working on something like this). But this sequence is more about "Why" than "How", much of what I want to talk about is the overall mindset, and how various skills fit together.

I'm using "frame" to mean "the broadest possible way for two people can miss each other." I deliberately didn't use words like "ontology" or "outlook", that misleadingly sound like they mean something specific.

Attention conservation notice: if this all makes intuitive sense and you're pretty sure you understand what I mean by 'frame' you can probably skip the rest of this article.

Holistic Frame Evaluation

Why focus on all three ways-of-*seeing/thinking/communicating* at once? My experience is that they often come all muddled together. It's useful to be able to separate them, but their default state is often intertwined.

In the previous essay, I listed some possible frames, including:

- *Gears-Oriented-Frames,* focused on physical processes in the world and how they interact
- *Feelings-Oriented-Frames,* focused on what emotions and inner experiences the conversation participants are having
- *Power-Oriented-Frames*, focused on the relative status and power of the participants and how that fits into the broader world

(Again, these are not at all meant to be comprehensive. These are just a few examples, like pointing at a rabbit, a bumblebee and a human and saying "Hey, I'm talking about animals. Here's a few examples of animals to get the idea across of what an animal is.")

This time I'll try to cross-reference those frames with the Seeing / Thinking / Communicating schema.

Ways of Communicating

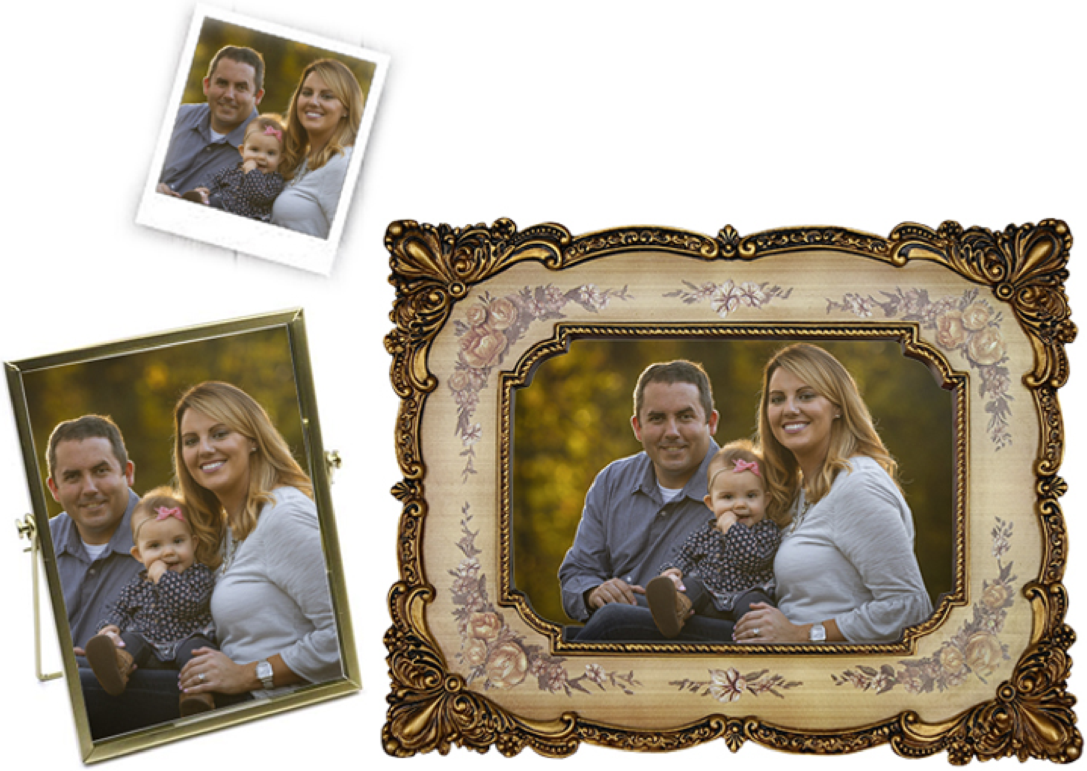
Taymon on Facebook remarked, regarding my previous post:

> The conflicts you describe are when each participant has a false belief about the other participants' objectives, and so the things they say, that they expect to help accomplish those objectives, don't.
>
> The whole thing doesn't seem that conceptually difficult.
>
> You seems to have had something bigger and harder to grasp in mind, but I still don't feel like I have much of a clue as to what it is.

I have at least a bit more in mind here. But, "people have different conversational goals" is at least one major source of frame conflict, and is perhaps the most obvious one. So let's start there.

Picture Frames, and Context

These pictures are identical, but have different frames.



(Image by [Ken Rementer](#))

The frame conveys information about the intent of the photo. The first is a polaroid that was probably taken off-the-cuff. The second and third look more like deliberate family portraits. But the third's elaborate frame conveys some additional "specialness" and pride.

Similarly, the same conversation might have very different contexts depending on who's participating, how they relate to each other, and what they think they're talking about.

Picture Frame Examples

[note: [political examples can be unnecessarily mindkilling](#) , but it seemed useful to have a concrete example that readers would likely be familiar with. Most of the examples here aren't too dependent on which position is right or wrong]

Alice and Bob are in a newly formed relationship, arguing about the politics of minimum wage.

There are some "obvious" frame disconnects that can happen here, if, say, Alice, thinks that the primary focus of the conversation is to figure out optimal policy, and Bob thinks it's to avoid relationship conflict.

(Or: if one of them thinks the goal is to figure out who is smarter and more socially dominant in their relationship)

There could be subtler mismatches:

Maybe they both think the conversation is more about their relationship than their preferred policies. (Apart from voting once a year, neither are major political activists).

But Alice enjoys barbed wit and thinks of the argument as a playful banter, which is a sign of comfort-with-each-other. (She thinks they are both comfortable enough with each others positions that the relationship is not at risk if they disagree too hard). She thinks it's fine to make straw-man-ish-jokes about Bob's position.

Bob thinks that political disagreement is evidence of longterm incompatibility, and thinks that Alice's banter is signaling disrespect, and strawmanny jokes make him feel particularly "not seen."

Ways of Seeing

Windows, Lenses, and "Tuning In"

Take a look at this street:

There's a lot going on here. But, imagine that you live in a house with only one window facing this landscape. Your impression of the landscape might depend a lot on where that window is placed.

One window showcases a city skyline. Another primarily focuses on trees and sunset. The third reveals a pile of construction rubble.

Sometimes people have the exact same goals for a conversation, and yet nonetheless face weird disconnects because they *aren't looking at the same parts of reality.*

It's compounded if they have different goals for the conversation *and also* aren't looking at the same parts of reality. Even if they say "oh, I just realized we were coming at this conversation with two different goals", they'll keep talking past each other.

Another metaphor here is the *lens,* where two people are looking at the exact same scene, but they perceive it differently.

Window Frame / Lens Example

Alice and Bob are still arguing about minimum wage. Here are few different sets of "windows" they could be looking through.

Different priors and past experience

Alice has read some econblogs and the first question she's asking is "what do the laws of supply and demand say about this? Is it marginally profitable to hire an additional person?". Bob has read some socialist blogs and his first questions have more to do with "what is fair, what do people deserve, who has power and how much surplus do they have?"

Bob has a bunch of friends struggling to get by on minimum wage who can't pay their bills.

Alice has friends trying to run small businesses, constantly frustrated by the difficulty of navigating regulations.

They might believe each other's points are real and relevant, but feel frustrated by each other's missing moods, which leads them to be suspicious of each other.

Different needs in the moment

What if Alice and Bob are having *both* a "Picture Frame" and "Window Frame" mismatch?

Maybe it's a situation where Alice thinks the point is figuring out optimal policy, and Bob thinks the point is to figure out longterm relationship compatibility, ideally demonstrating that they care about each other. So at the beginning of the conversation, Alice is casually dismissing arguments she thinks are bad as "stupid".

In this scenario, both Alice and Bob have enough relationship/conversational intelligence to notice that that's happening. They have a meta conversation about their respective needs, and then continue in the conversation to *both* figure out optimal policy while attending to Bob's feelings and their overall relationship.

But...

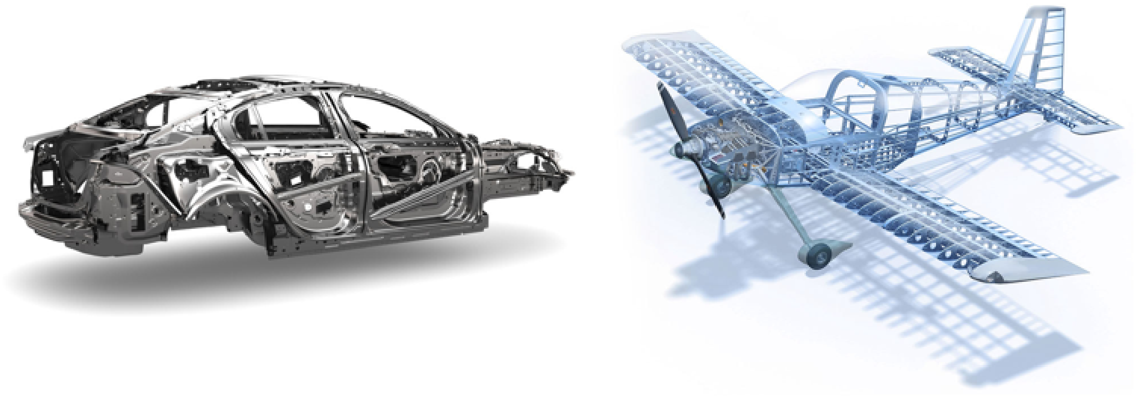
...it's still the case that Alice is *very practiced* at noticing how policy positions fit together, and *not* practiced at tracking Bob's facial expressions, emotional state, or how her comments might be coming across. If part of what Bob wants is to *feel seen*, then Alice might still not do a good job of doing that because Alice is still used to looking through the "policy cause-and-effect" window than the "feelings and relationships" window — *even when that's explicitly her goal.*

Part of navigating frame differences is learning to tune into different aspects of a conversation.

Ways of Thinking

Frameworks, and how ideas fit together

Then, there's how ideas fit together. The framework of a car can fit different sorts of pieces into it than the framework of a bridge. There are different ways of conceptually connecting things.



Different ways of thinking afford different ways of combining evidence, and permit different sorts of answers.

In many social environments, "authority" is treated as a legitimate source of truth, and "the boss/priest/god says so" is a legitimate argument to bring to the table.

In an empirical/rationalist framework, authority often still matters, but ideally argument screens off authority.

Even among people trying to reason carefully, there's a lot of disagreement about how to think. Frequentist vs Bayesian? Probabilistic vs Proof Based? How do you evaluate a study? In what circumstances (if ever) are personal experiences better evidence than a study?

Do certain types of evidence have a track record of being biased or harmful to you? (Or: do they have a track record of... making you uncomfortable and annoyed and you don't like to think about it?)

All of this can interface with Picture Frames or Window Frames. You might have different ways of thinking you employ in different domains. Your way of thinking may bias which sort of evidence you look for and how seriously you take it.

Framework Example

Alice and Bob are both arguing about minimum wage policy through a materialist cause-and-effect lens. But Alice's opinion is shaped by econ 101 models (i.e. raising prices on a thing should decrease the amount of it), and Bob's opinion is shaped by having read some studies saying that raising minimum wage hasn't caused more unemployment.

Bob thinks empiricism is the ultimate judge, more important than theory. Alice thinks the theory is robust enough that she defies the data, and argues Bob's studies are either cherry-picked or the result of a messy, underpowered world.

At least part of their discussion is going to need to address the question "what counts as good evidence." (Exactly how to go about that is, in this post, left as an exercise to the reader)

Composite and Unconscious Frames

What makes this all quite hard, in my experience, is the way this is all intertwined. Often there's multiple layers of Window-Frame/Picture-Frame/Framework at play, and it's not obvious how important each of them respectively are.

There also can be multiple "contents of frame", involved in each layer. A person's window-frame might be focused on *both* economic theory and also power dynamics (or subtle subsets of each other those).

And some of this may be totally unconscious: sometimes Alice is quite confident that her goal is simply to point out good policy, but she's unconsciously employing a strategy optimized to make her feel smart and in control, and Bob may be more tuned into the later than the former.

(Or, Bob may think he's noticing important social moves that Alice is pulling, but Bob has become hypersensitized to social nuances and is noticing things that honestly aren't there, or aren't nearly as causally relevant as he thinks)

What exactly to do with all this is tricky, and context dependent, and requires building up a few different skills (many of which are hard to convey via blogpost). But it seems useful to at least have a rough sense of the landscape.

Karate Kid and Realistic Expectations for Disagreement Resolution

There’s an essay that periodically feels deeply relevant to a situation:

> Someday I want to write a self-help book titled “F*k The Karate Kid: Why Life is So Much Harder Than We Think”.
>
> Look at any movie with a training montage: The main character is very bad at something, then there is a sequence in the middle of the film set to upbeat music that shows him practicing. When it's done, he's an expert.
>
> It seems so obvious that it actually feels insulting to point it out. But it's not obvious. Every adult I know--or at least the ones who are depressed--continually suffers from something like sticker shock (that is, when you go shopping for something for the first time and are shocked to find it costs way, way more than you thought). Only it's with effort. It's Effort Shock.
>
> We have a vague idea in our head of the "price" of certain accomplishments, how difficult it should be to get a degree, or succeed at a job, or stay in shape, or raise a kid, or build a house. And that vague idea is almost always catastrophically wrong.
>
> Accomplishing worthwhile things isn't just a little harder than people think; it's 10 or 20 times harder. Like losing weight. You make yourself miserable for six months and find yourself down a whopping four pounds. Let yourself go at a single all-you-can-eat buffet and you've gained it all back.
>
> So, people bail on diets. Not just because they're harder than they expected, but because they're so much harder it seems unfair, almost criminally unjust. You can't shake the bitter thought that, "This amount of effort should result in me looking like a panty model."
>
> It applies to everything. [The world] is full of frustrated, broken, baffled people because so many of us think, "If I work this hard, this many hours a week, I should have (a great job, a nice house, a nice car, etc). I don't have that thing, therefore something has corrupted the system and kept me from getting what I deserve."

Last time I brought this up it was in the context of realistic expectations for self improvement.

This time it’s in the context of productive disagreement.

Intuitively, it feels like when you see someone being wrong, and you have a simple explanation for why they’re wrong, it should take you, like, 5 minutes of saying “Hey, you’re wrong, here’s why.”

Instead, Bob and Alice people might debate and doublecrux for 20 hours, making serious effort to understand each other’s viewpoint… and the end result is a conversation that *still* feels like moving through molasses, with both Alice and Bob feeling like the other is missing the point.

And if 20 hours seems long, try *years*.

AFAICT the Yudkowsky/Hanson Foom Debate didn’t really resolve. But, the general debate over “should we expect a sudden leap in AI abilities that leaves us with a single victor, or a multipolar scenario?" has actually progressed over time. Paul Christiano's Arguments About Fast Takeoff seemed most influential of reframing the debate in a way that helped some people stop talking past each other, and focus on the actual different strategic approaches that the different models would predict.

Holden Karnofsky initially had some skepticism about some of MIRI's (then SIAI's) approach to AI Alignment. Those views changed over the course of years.

On the LessWrong team, we have a lot of disagreements about how to make various UI tradeoffs, which we still haven't resolved. But after a year or so of periodic chatting about I think we at least have better models of each other's reasoning, and in some cases we've found third-solutions that resolved the issue.

I have observed myself taking years to really assimilate the worldviews of others.

When you have deep frame disagreements, I think "years" is actually just a fairly common timeframe for processing a debate. I don't think this is a *necessary fact* about the universe, but it seems to be the status quo.

Why?

The reasons a disagreement might take years to resolve vary, but a few include:

i. Complex Beliefs, or Frame Differences, that take time to communicate.

Where the blocker is just "dedicating enough time to actually explaining things." Maybe the total process only takes 30 hours but you have to actually do the 30 hours, and people rarely dedicate more than 4 at a time, and then don't prioritize finishing it that highly.

ii. Complex Beliefs, or Frame Differences, that take time to absorb

Sometimes it only takes an hour to explain a concept explicitly, but it takes awhile for that concept to propagate through your implicit beliefs. (Maybe someone explains a pattern in social dynamics, and you nod along and say "okay, I could see that happening sometimes", but then over the next year you start to see it happening, and you don't "really" believe in it until you've seen it a few times.)

Sometimes it's an even vaguer thing like "I dunno man I just needed to relax and not think about this for awhile for it to subconsciously sink in somehow"

iii. Idea Innoculation + Inferential Distance

Sometimes the first few people explaining a thing to you suck at it, and give you an impression that anyone advocating the thing is an idiot, and causes you to subsequently dismiss people who pattern match to those bad arguments. Then it takes someone who puts a lot of effort into an explanation that counteracts that initial bad taste.

iv. Hitting the right explanation / circumstances

Sometimes it just takes a specific combination of "the right explanation" and "being in the right circumstances to hear that explanation" to get a magical click, and unfortunately you'll need to try several times before the right one lands. (And, like reason #1 above, this doesn't necessarily take *that* much time, but nonetheless takes years of intermittent attempts before it works)

v. Social pressure might take time to shift

Sometimes it just has nothing to do with good arguments and rational updates – it turns out you're a monkey who's window-of-possible beliefs depends a lot on what other monkeys around you are willing to talk about. In this case it takes years for enough people around you to change their mind first.

Hopefully you can take actions to improve your social resilience, so you don't have to wait for that, but I bet it's a frequent cause.

Optimism and Pessimism

You can look at this glass half-empty or half-full.

Certainly, if you're expecting to convince people of your viewpoint within a matter of hours, you may sometimes have to come to terms with that not always happening. If your plans depend on it happening, you may need to re-plan. (Not always: I've *also* seen major disagreements get resolved in hours, and sometimes even 5 minutes. But, "years" might be an outcome you need to plan around. If it *is* taking years it may not be worthwhile unless you're actually building a product together.)

On the plus side... I've now gotten to see several deep disagreements actually progress. I'm not sure I've seen a years-long disagreement resolve *completely*, but have definitely seen people change their minds in important ways. So I now have existence proof that this is even possible to address.

Many of the reasons listed above seem addressable. I think we can do better.

Propagating Facts into Aesthetics

Epistemic status: Tentative. I've been practicing this on-and-off for a year and it's seemed valuable, but it's the sort of thing I might look back on and say "hmm, that wasn't really the right frame to approach it from."

In doublecrux, the focus is on "what observations would change my mind?"

In some cases this is (relatively) straightforward. If you believe minimum wage helps workers, or harms them, there are some fairly obvious experiments you might run. "Which places have instituted minimum wage laws? What happened to wages? What happened to unemployment? What happened to worker migration?"

The details will matter a lot. The results of the experiment might be weird and confusing. If I ran the experiment myself I'd probably get a lot of things wrong, misuse statistics and forget to account for some confounding factors. But I don't feel *confused* about how to learn better statistics, account for more confounders, etc.

But there's a problem that seems harder to me, which is *how to change my mind about aesthetics.* Sarah Constantin first brought this up in Naming the Nameless, and I've been thinking about it ever since.

I think a lot of deep disagreements have to do with "what is beautiful, and what is ugly?", and inability to directly address this is part of what prevents those disagreements from resolving.

In the case of the minimum wage example, you might run an experiment, and find overwhelming evidence that minimum wage helps or hurts workers. But because there's lots of confounders, the evidence might be mixed and confusing. How you interpret it will depend on how it fits into your existing worldview.

Part of this has to do with your ontological frame. But I think a lot has to do with aesthetics judgments, such as:

- *Is capitalism ugly and/or distasteful?* You might have very salient examples of how capitalism can result in exploitation, pollution, or people becoming trapped in unhealthy power structures.
- *Is capitalism beautiful?* Alternately, it might be salient that capitalism creates supermarkets, gains from trade, and vast surplus. Economic efficiency isn't just pretty numbers on a graph, it's real value being created.

These don't directly bear on the minimum wage question, but might make it harder to resolve.

In some cases, your aesthetic taste might make it harder to update on new information properly. In other cases, your aesthetic taste might help you to notice important patterns more readily.

Why 'aesthetics?'

I’m using the word aesthetic in a nonstandard way. When people do that, I think it’s important to be clear and about what they’re doing and why.

There’s a few different words I might have used here, including “feelings”, “ontologies”, “frameworks”, and “values.”

Most obviously, I could have asked ‘is capitalism *good/bad?*’ instead of ‘is capitalism beautiful or ugly?’.

I’m making a fairly strong claim (weakly held) that “is it beautiful or ugly?” is at least one of the important questions to be asking, in addition to “is capitalism good/bad” and “does raising minimum wage help or harm workers?”. Not because it’s how a flawless AI would think about it, but because it’s how humans seem to often think about it.

What is an aesthetic?

An aesthetic is a mishmash of values, strategies, and ontologies that reinforce each other.

The values reinforce “you want to use strategies that achieve these values.”

The act of using a particular strategy shapes the ontology that you see the world through.

The ontology reinforces what values seem important to you.

Together, this all creates a feedback loop between your metagoals and subgoals, where the process of using this cluster of value/strategy/ontology makes each link in the chain stronger.

In humans (who have messy, entangled brains), this caches out into feelings, felt senses. The original goal and the metagoals blur together. I think “this helps me achieve my [generic] goals” might reinforce “these *particular* subgoals I have are good goals to help with my overall flourishing.”

This might be implemented via evolution over millions of years, or via human brains over decades. A Just So Story I'm not sure I endorse but hopefully gets the point across:

> "A flower is beautiful, you say. Do you think there is no story behind that beauty, or that science does not know the story? Flower pollen is transmitted by bees, so by sexual selection, flowers evolved to attract bees—by imitating certain mating signs of bees, as it happened; the flowers' patterns would look more intricate, if you could see in the ultraviolet. Now healthy flowers are a sign of fertile land, likely to bear fruits and other treasures, and probably prey animals as well; so is it any wonder that humans evolved to be attracted to flowers?

Here are some things that you might find beautiful, or distasteful:

- Mozart
- Punk Rock
- Readable, well-written code
- Clever hacks that got the job done quickly

- Cities built on rectangular grids
- Winding alleyways in villages where nobody has consistent names
- People being physically affectionate in public
- A harsh, barren desert
- A lush valley with a river
- Swamps / wetlands
- Nature in general
- Manicured gardens
- Books
- Throwing away books
- People speaking in languages different from yours
- Dense spreadsheets laden with accurate data
- Minimalism
- Frugalism
- Patriotism
- People going out of their way to be kind to their neighbors
- People going out of their way to solve small-but-common problems using math

(If you’re like me, you might find it distasteful when people make moral arguments that seem rooted in distaste... and then feel kinda self conscious about the contradiction)

Sometimes you’re doublecruxing with someone, and they’ve explained their model. And their model… makes sense. But the conclusion just seems *so damn ugly*. You want to take 5x the time to write beautiful code, and they just want you to get the job done and ship it.

One thing you can do is push aside your aesthetic judgment, shut up and multiply. This may be useful for expediency.

But sometimes, I think the correct thing is for one or both people to backpropagate facts through their aesthetics.

I do *not* think you should rush or "force" this. Your sense of beauty is there for a reason. But I have a sense that figuring out how to do this well is a key open problem in applied rationality.

Examples

Are Swamps Beautiful?

Compare the swamp with a verdant forest.

If you're like me, swamps seem ugly. Forests seem pretty.

My associations with swamps come largely from stories (and perhaps most concretely, from the game "Magic the Gathering"), where they're often presented as places of disease, murky horrors and corrupt magic. In person, swamps are physically hard to walk in (sometimes solid ground turns out to be algae), and full of mosquitos that bite me.

Are these associations accurate?

Well, the solid ground and mosquito issues are definitely real.

Swamp Thing is not real, life-stealing magic is not real.

Are there additional facts I can learn? My sister evaluates land for construction projects. She says that swamps often serve important roles as a natural way to filter water, and when you naively drain swamps, water quality in an area gets worse. James Scott in Against the Grain claims that early Sumerian civilization developed in swamps, where food and resources were plentiful and life was fairly leisureful – until empires arose and subjected people and forced them to switch to easily-taxable crops instead of the ones that grew naturally. (Speaking of which: is civilization beautiful or ugly?)

I probably find forests beautiful, in part, because they represent a lot of resources that I understand how to make use of. If swamps also supply those resources, maybe I should respect them more?

I also find forests beautiful because my experience stems from a) enchanted forests in fairytales, and b) relatively manicured national parks. If I remind myself that the last time I walked through an *untamed* forest, it was dense with brambles that cut me 'till I bled. It wasn't actually a much nicer experience than the last time I explored a swamp. (It also had non-trivial numbers of mosquitos)

In this example, simply mulling over the facts naturally re-organizes my feelings about them. I still find swamps ugly, but less ugly than before. I expect that, if I reflected on this periodically, over time, it would shift a bit more.

Are Harsh Deserts Beautiful?

I am in fact confused by this. My answer is "yes", and I don't know why. Deserts don't have much in the way of resources. Their stark beauty is more like the way a statue is beautiful than the way a forest is beautiful.

I mulled this one over for a while, am still confused and I note it here because "noticing the limits of a model" seems important.

[Edit: this was discussed more in the comments.]

Is Helping Nearby People Other Beautiful?

The first experience I got with aesthetic doublecrux was debating “hufflepuff virtue” with Oliver Habryka.

I had a strong sense that “people helping each other out” was good and right and virtuous. There was a beauty to the sort of community where everyone notices when someone is hurting (and reaches out to help), or when a space is messy (and cleans it up). There was a cluster of attributes that seemed to fit together in a way that was stronger than the sum of its parts.

And this was visibly lacking in the Berkeley community, and it was resulting in people feeling alienated and distrustful of each other, and many spaces being either messy, or burdening a single person with cleaning up everyone else's mess.

This seemed concretely harmful. But it also just seemed… ugly and bad.

Oliver had a different view, which I summarize as the "systemization and specialization" approach. (previously discussed here)

If everyone has to pay attention to their environment and notice things that need doing, this is a *lot* of cognitive overhead. If people only have seven working memory slots but they're spending one of them on tracking the environment, that's a dramatic cost on their ability to think. For a community that specializes in thinking, this could be quite bad.

Moreover, "everyone pitch in" is just a really inefficient way of getting things done. A better solution is to streamline and automate as much of the work as possible, hire cleaning services, and whatever remaining work needs doing, simply pay one one of the people something commensurate for their time and effort. Specialization is how things get done when you're doing them seriously.

We argued about this over the course of three days.

I still think there are some things habryka was missing here. But eventually my worldview shifted in some significant ways:

- I updated that the "everyone pitch in" way of keeping spaces clean doesn't make sense for longterm organizations with serious funding. Specialization is real, cognitive bandwidth is precious, and it's generally better to just hire a cleaning service if you can afford one.
- I updated a bit (talking with Satvik) that my model that "helping each other out in low-key ways builds trust which later enables more extensive projects" wasn't as strong as I thought. Satvik asked something like "do you think startup cofounders tend to team up because they've helped each other take out the trash? I feel like it's more about sharing a clear vision and principles or something." And I thought back to some experiences and… yeah that seemed maybe more accurate.
- I gained a better understanding of where and why the "everyone pitch in" approach is useful.
    - Cleaning services are expensive, and if you're a fledgling organization or a typical household, it's probably not worth hiring a cleaner more than once a week or so. Meanwhile, people make messes much more frequently than once a week. If you want your space nice, you have to clean it yourself.
    - There's a value that comes from having community spaces use the "everyone pitch in" method, in that it creates a stronger sense of ownership and buy-in for the space. It also is a mechanism by which people can relate to each other more easily. While this might not be that important for a company, it seems important for a community that's aiming to meet community-shaped-needs.

But this all left me with a nagging, frustrated sense that something important and beautiful being lost. I *want* to live in a world where people help each other out in small ways. It's the particular kind of beauty that a small town in a Miyazaki movie embodies. It feels important to me.

Under what circumstances should I change how I feel about that?

There’s a sense in which aesthetics can’t be proven wrong, or at least “trying to prove it wrong” isn’t really the right frame of mind.

But… I have an aesthetic preference for *consistency,* and for *believing true things* (whether this is good is another question, but I’m taking it at face value for now), which informs my other aesthetics. Aesthetics can turn out to be built out of contradictory pieces, and they can turn out to hinge of false beliefs.

“Trying on” another aesthetic

While talking to habryka, I tried to get a sense of what it’s like to live in the world where systemization and specialization are obviously good and right. What was it like to be habryka? How did this fit together with his other beliefs and values?

Then, once I had a good handle on that, I tried to inhabit “what would it be like to be a Raemon who found systemization and specialization good and right?”. Without *actually* adopting the aesthetic, I tried fitting it into my existing model. This was a bit of an aesthetic process of its own – like trying on a new outfit and seeing how I reacted in the mirror.

I’m not sure if habryka endorses considering those as an ‘aesthetic’, per se. But I found this process valuable.

I gained some ability to see systemization as beautiful. My sense of hufflepuff beauty became more nuanced and caveated.

Clean Code vs Quick Hacks

Humans have (an instinctive? Learned? I'm not sure) sense that when you smell fecal matter or rotting flesh, there is probably disease nearby. It's digusting.

Dogs... well, I'm not 100% sure what's going on with dogs but I think it's something like "strong odors that mask my scent are more useful than disease is bad", and for some reason fecal matter is joyful to play around in.

Programmers often learn that spaghetti code is *evidence* of bugs, even if they don't know exactly what the bug is yet. It acquires a *bad code smell.*

Young programmers often do not have this sense of distaste, and it is important for them to acquire it.

On the flipside: there is also a thing where, well, sometimes you're rushing to ship an Minimum Viable Product and you don't have time to do everything right. It can be legitimately hard to figure out how much effort to put into "doing things right." But it seems at least sometimes, experienced coders either need to learn to "hold their nose" and do the quick fix, or to develop alternate aesthetics that they can shift between depending on circumstances.

Knobs to Turn

There are a few different directions this kind of process might go:

- You could shift to find something *more* beautiful than you did before
- You could shift to find something *less* beautiful than you did before.
- You could shift to find something *more* distasteful than you did before.
- You could shift to find something *less* distasteful than you did before.

I have some sense that these are subtly different processes, although not much evidence to back that up. I also feel like in each case, going from Zero to N, or N to Zero, is different than dialing an existing aesthetic response up or down.

Gaining a new appreciation for why something is beautiful feels different than gaining a categorically new form of disgust. In particular, gaining a new form of beauty mostly makes my life feel nicer, whereas gaining a new form of disgust increases the unpleasantness

Why Does this Matter?

In Naming the Nameless, Sarah Constantin references this comment by Scott Alexander:

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

Sarah notes:

> Now, I could say "just don't do that, then" -- but Scott of 2009 would have *also* said he believed in being independent and rational and not succumbing to social pressure. Good intentions aren't enough. [...]
>
> I think it's much better to try to make the implicit explicit, to bring cultural dynamics into the light and understand how they work, rather than to hide from them.

Scott's comment gets at what I mean by "An aesthetic is a mishmash of values, strategies, beliefs, and ontologies that reinforce each other." He starts with a belief, then adopts a strategy for how he relates his communication to that belief, and then ends up with a vague sense that the belief is "cringey", and later collapsing it to "cringey and wrong".

This quite worrying epistemic horror.

I think most of what needed saying, Sarah already said, but it's worth concluding with here:

> If you take something about yourself that's "cringeworthy" and, instead of cringing yourself, try to look at *why* it's cringeworthy, what that's made of, and dialogue honestly with the perspective that disagrees with you -- then there is, in a sense, nothing to fear.
>
> There's an "elucidating" move that I'm trying to point out here, where instead of defending against an allegation, you say "let's back up a second" and bring the entire situation into view. It's what double crux is about -- "hey, let's find out what even is the disagreement between us." Double crux is hard enough with *arguments*, and here I'm trying to advocate something like double-cruxing *aesthetic preferences*, which sounds absurdly ambitious. But: imagine if we could talk about *why* things seem beautiful and appealing, or ugly and unappealing. Where do these preferences come from, in a causal sense? Do we still endorse them when we know their origins? What happens when we *bring tacit things into consciousness*, when we talk carefully about what aesthetics evoke in us, and how that might be the same or different from person to person?
>
> Unless you can think about how cultural messaging works, you're going to be a *mere* consumer of culture, drifting in whatever direction the current takes you.

I'm hoping this post gives some nuts and bolts on how to actually make progress on that goal.

Again, I don't know that the specific techniques I list in this post are the best ones, or how often exactly aesthetic concerns are most relevant. I think it's usually good form to *start* with an attempt to take arguments at face value, and debate about concrete beliefs.

But, if that isn't working, I think digging into aesthetics is one of the tools that's important to have in your toolkit.