The Coordination Frontier

The Coordination Frontier: Sequence Intro

Sometimes, groups of humans disagree about what to do.

We also sometimes disagree about how to decide what to do.

Sometimes we even disagree about how to decide how to decide.

Among the philosophically unsophisticated, there is a sad, frustrating way this can play out: People resolve "how to decide" with yelling, or bloodshed, or, (if you're lucky), charismatic leaders assembling coalitions. This can leave lots of value on the table, or actively destroy value.

Among the *extremely* philosophically sophisticated, there are different sad, frustrating ways this can play out: People have very well thought out principles informing their sense of "how to coordinate well." But, their principles are not the same, and they don't have good meta-principles on when/how to compromise. They spend hours (or years) arguing about how to decide. Or they burn a lot of energy in conflict. Or they end up walking away from what *could* have been a good deal, if only people were a bit better at communicating.

I've gone through multiple iterations on this sequence intro, some optimistic, some pessimistic.

Optimistic takes include: "I think rationalists are in a rare position to *actually figure out good coordination meta-principles,* because we are smart, and care, and are in positions where good coordination actually matters. This is exciting, because coordination is basically the most important thing [citation needed]. Anyone with a shot at pushing humanity's coordination theory and capacity forward should do that."

Pessimistic takes include: "Geez louise, rationalists are all philosophical contrarians with weird, extreme, self-architected psychology who are a pain to work with", as well as "Actually, the most important facets of coordination to improve are maybe more like 'slightly better markets' than like 'figuring out how to help oddly specific rationalists get along'."

I started writing this post several years ago because I was annoyed at, like, 6 particular people, many of them smarter and more competent than me, many of whom were explicitly interested in coordination theory, who nonetheless seemed to despair at coordinating with rationalists-in-particular (including each other). The post grew into a sequence. The sequence grew into a sprawling research project. My goal was "provide a good foundation to get rationalists through the Valley of Bad Coordination". I feel like we're *so close* to being able to punch above our weight at coordination and general competence.

I think my actual motivations were sort of unhealthy. "If only I could think better and write really good blogposts, these particular people I'm frustrated with could get along."

I'm currently in a bit of a pessimistic swing, and do not expect that writing sufficiently good blogposts will fix the things I was originally frustrated by. The people in question

(probably) have decent reasons for having different coordination strategies.

Nonetheless, I think “mild irritation at something not quite working” is pretty good as motivations go. I’ve spent the past few years trying to reconcile the weirdly-specific APIs of different rationalists who each were trying to solve pretty real problems, and who had developed rich, complex worldviews along the way that point towards something important. I feel like I can almost taste the center of some deeper set of principles that unite them.

Since getting invested in this, I’ve come to suspect “If you want to succeed at coordination, ‘incremental improvements on things like markets’ is more promising than ‘reconcile weird rationalist APIs.” But, frustration with weird rationalist APIs was the thing that got me on this path, and I think I’m just going to see that through to the end.

So.

Here is this sequence, and here is what the deal is:

Deep Inside Views, and the Coordination Frontier

A common strength of rationalists is having deep inside-view models. Rich, gears-based inside views are often a source of insight, but are hard to communicate about because they are many inferential steps away from common knowledge.

Normally, that’s kinda fine. If you’re not specifically building a product together, it’s okay if you mostly go off in different directions, think hard-to-explain-thoughts, and only occasionally try to distill your thoughts down into something the median LessWronger can understand.

But it’s trickier when your rich, nuanced worldview is *specifically about coordinating with other people.*

The **Coordination Frontier** is my term for “the cutting edge of coordination techniques, which are not obvious to most people.” I think it's a useful concept for us to collectively have as we navigate complex new domains in the coming years.

Sometimes you are on the coordination frontier, and unfortunately that means it's either your job to explain a principle to other people, or you have to sadly watch value get destroyed. Often, this is in the middle of a heated conflict, where noticing-what’s-going-on is particularly hard.

Other times, you might *think* you are on the coordination frontier, but actually you're wrong – your principles are missing something important and aren’t actually an improvement. Maybe you’re just rationalizing things that are convenient for you.

Sometimes, Alice and Bob disagree on principles, but are importantly *both somewhat right*, and would benefit from somehow integrating their different principles into a coherent decision framework.

When you are trying to innovate along the coordination frontier, there aren't purely right-or-wrong answers. There are different things you can optimize for. But, I think there are *righter* and *wronger* answers. There are principles that constrain what types of coordination solutions are appropriate, given a particular goal. There are failure modes you can fall into, or, notice and avoid.

And, if you are a particular agent with a particular set of skills and cognitive bandwidth and time and goals, interacting with other agents with particular goals and resources...

...then I think there (might) be a fairly narrow range of theoretically-best-answers to the question "how do I coordinate with these people."

A rationalist failure mode is to get overly attached to the belief that you've found "the right answer." One of the more important meta-coordination principles is "We don't really have *time* to agree on which of our weird philosophical positions is right, and we need to coordinate *anyway"*.

Nonetheless, I *do* think there is something important about the fact that "righter answers exist."

My overall preferred approach is a mixture of pragmatism in the day-to-day, and curious, lawful thinking about the theoretical ideal.

Distinctions near the Frontier

A few people read an earlier draft of this post and were like "Cool, but, I don't know that I could use 'Coordination Frontier' in a sentence." I think it's easiest to describe it by contrasting a few neighboring concepts:

- The Coordination Baseline
- Coordination Pioneers
- The Coordination Frontier
- The Coordination Limit

The Coordination Baseline

AKA "mainstream civilization"

The Coordination Baseline is what most people around you are doing. In your particular city or culture, what principles do people take as obvious? Which norms do they follow? Which systems do they employ? Does a shopkeeper charge everyone a standardized price for an item, or do they haggle with each individual? Do people vote? Can you generally expect people to be honest? When people communicate, does it tend to be Ask Culture or Guess Culture?

Who exactly this is referring to depends on the context of a discussion. It might refer to an entire country, a city, or a particular subculture. But there is at least some critical mass of people who interact with each other, who have baseline expectations for how coordination works.

Coordination Pioneers

Some people explore novel ways of coordinating, beyond the baseline. They develop new systems and schemes and norms – voting systems, auctions, leadership styles, etc. They are Coordination Pioneers.

Sometimes they are solving fully novel problems that have never been solved before – such as inventing a completely new voting system.

Sometimes they are following the footsteps of others who have already blazed a trail. Perhaps they are reinventing approval voting, not realizing it's already been discovered. Or, perhaps they read about it, and then get excited about it, and join a political movement to get the new voting system adopted.

The Coordination Frontier

The upper limit of human knowledge of how to coordinate well.

The coordination frontier is the pareto frontier of "what coordination strategies we are theoretically capable of implementing."

The frontier changes over time. Once upon a time, our best coordination tools were "physical might makes right, and/or vaguely defined exchange of social status." Then we invented money, and norms like "don't lie".

During the cold war, the United States and Soviet Union were suddenly thrown into a novel, dangerous situation where either could lay devastating waste to the other. Game theorists like Thomas Schelling had to develop strategies that incorporated the possibility of mutually assured destruction, where in some ways it was *better* if both sides had the ability to reliably, inevitably counterattack.

Most people in the world probably didn't understand the principles underlying MAD at the time, but, somewhere in the world were people who did. (Hopefully, ranking generals and diplomats in the US and Soviet Union).

The Coordination Limit

The upper limit of what is theoretically possible.

For any given set of agents, in a given situation, with a given amount of time to think and communicate, there are limits on what the best joint decisions they could reach. The Coordination Limit is the theoretical upper bound of how much value they could jointly optimize for.

There will be different points along a curve, optimizing for different things. There might be multiple "right answers", for any given optimization target. But I think the set of options for "perfect-ish play" are relatively constrained.

I think it's useful to track separately "what would N fully informed agents do, if they are perfectly skilled at communicating and decisionmaking", as well as "given a set of agents who *aren't* fully knowledgeable of coordination theory, with limited communication or decisionmaking skills and some muddled history of interaction,

what is the space of possible optimization targets they can hit given their starting point?"

Where is this going?

The thing I am excited about is pushing the coordination frontier forward, towards the limit.

This sequence covers a mixture of meta-coordination principles, and object-level coordination tools. As I post this, I haven't finished the sequence, nor have I settled on the single-most-important takeaways.

But here are my current guesses for where this is going:

1. **Most of the value of coordination-experimentation lives in the future**. Locally, novel coordination usually costs more than it gains. This has implications on what to optimize for when you're experimenting. Optimize for longterm learning, and for building up coordination-bubbles where you'll get to continue reaping the benefits.

2. Complex coordination requires some combination of **Shared-Understanding-And-Skills**, or, **Simplifying UI.**

3. **Misjudging inferential distance, and failing to model theory of mind properly, are particularly common failure modes**. People are usually not coordinating based on the same principles as you. This is more true the more you've thought about your principles. Adjust your expectations accordingly.

4. Lack of **reliable reputation systems** is a major bottleneck, at multiple scales. (Open Problem #1)

5. Another bottleneck is **ability to quickly converge on coordination-frame**. This is tricky because "which coordination frame we use" is a negotiation, often with winners and losers. But I think rationalists often spend more time negotiating over coordination-frame that it's worth. (Open Problem #2)

6. **Coordination is very important during a crisis,** but it's hard to apply *new* principles or depend-on-particular-skills during high stakes crises. This means it's valuable to establish good policies during non-crisis times (and, make sure to learn from crises that do happen)

Coordination Schemes Are Capital Investments

Second post in the Coordination Frontier sequence. Intro is here .

Once upon a time, we didn't have money, or auctions, or markets. We didn't have the concept of voting (let alone distinctions between first-past-the-post, or approval voting). We didn't have banks or stock trading. We didn't have Kickstarter.

~~We had to haggle over everything, which cost time. And it fundamentally limited the scale at which we could accomplish things. If a transaction cost exceeds the value of a trade, that trade can't happen.~~ [edit: arguably false, see discussion in comments, but fortunately not that cruxy for the rest of the post]

Once upon a time, someone had to be the *first person to invent* each of these concepts. The people around them were probably vaguely annoyed at having to learn a new thing. Many of the concepts required multiple building blocks and took thousands of years to reach saturation.

If you are happen to be around a lot of people around who:

1. geek out about coordination schemes, and
2. have an entrepreneurial bent

Then a really valuable thing you can do is explore new coordination schemes, while making an effort to distill them down into something the general population might actually be able to use.

This is hard. There are lots of failure modes. But, coordination schemes are a key thing that makes humanity powerful. Pushing the state-of-the-art forward is valuable.

With that in mind, let's explore some examples and concepts.

Examples: Negotiation Technology

Second Price Auctions

I remember the first time I got the results of a second-price auction. It felt like magic. Instead of arguing for hours about who got which room in a new apartment, I just wrote down my true preference for how much I was willing to pay for each room. Then I automagically got assigned a room that was *cheaper* than I had been willing to pay for it. (In a second price auction, each participant submits a sealed bid, and then person who bid highest pays the *second-highest* bid)

At previous apartments, we had just eyeballed the size of each room, and then came up with rent-allocations for each room that had vaguely round, fair-sounding numbers, and then picked rooms. It was chaotic, and didn't have a good way to account for some people valuing particular rooms for subtle, personal reasons.

Second price auction was an important, conceptual advance. But it's not very popular in broader society. Why?

At my first second-price auction, someone had to patiently explain it to me before I trusted it. I'm not sure I actually *did* trust it until after everything had been resolved. Beforehand, it felt overcomplicated and I was sort of annoyed that we weren't just eyeballing the rooms and winging it with nice round numbers. The explanation process took awhile, and I'm not 100% sure the marginal improvements in fairness/efficiency were worth it.

Thinking about how much I valued each room was actually pretty hard. "How much do I value something" is a skill that I had never really used – I never needed to. I just needed to eyeball a price and think "Worth it? or not worth it."

Still, I think this was clearly worthwhile. Because *later*, I got to use second price auctions in other situations. I went in having a clear understanding of how it worked. I had developed some skill of evaluating how much things were worth to me. I got all of the magical "it just works" feeling and none of the stress.

Other Formats

More recently, I was having a different sort of negotiation (over how much to sell someone some used exercise equipment for). They proposed a negotiation scheme intended to sidestep the haggling.

The premise was: we each privately note down our true value for the object, then reveal simultaneously. If their value was higher than mine, we'd make the trade for the *midpoint* between those two values. Otherwise, there is no deal.

At the time, I didn't *quite* understand the principles that were underlying this. I was somewhat stressed out about unrelated things. I wasn't sure what my true value was, or how to figure it out. I had a vague expectation that I was supposed to do *something* strategic but I wasn't sure what. (It turns out this was in fact false)

I felt slightly annoyed at the proposed system, and thought to myself "Well, I'll just try this out and if it doesn't work we can haggle normally afterwards." I didn't put much effort into figuring out my true value of the item.

I ended up choosing a number that was too high.

I tried to haggle more, and then my partner said "no, sorry the whole point was to replace the haggling process."

So, the deal didn't go through.

Shortly afterwards I realized that the whole scheme depended on precommitment not to make further deals, to incentivize us to give reasonable numbers. Alas, it was too late for realizing it to matter.

I later sold the item on craigslist, costing me a bit of time I might not have needed to spend.

Non-Auctions

Once, I moved into a house with three friends. We were all veterans of second-price auctions. But, in this case, it was overdetermined who would need each room:

- One room was huge, and it only made sense for me and my girlfriend to live there.
- One room was teeny, and it only made sense for the poorest roommate to live there.
- One room was medium, and obviously made sense for our fourth roommate to live there.

I'm not sure if there was an ideal, mathematical solution here. But we spent a couple hours talking about what to do. It seemed to be a zero-sum-negotiation where everything was predetermined except exactly how much rent we could extract from each other, and we didn't know how to do that fairly. We each cared about not screwing over each other, but didn't know how.

After an hour of arguing, I proposed "Wait, can we each just write down on a napkin how much we think each room should cost, and then reveal simultaneously? Maybe we all just secretly agree on the obvious distribution of rent and we don't have to debate".

People were skeptical. But then we did that. And then it turned out, everyone wrote a *higher* number for their room than the other three people had written. Meaning everyone got to pay a *lower* price for their room than they thought was fair. (And meanwhile, our numbers for each room were indeed pretty similar to each other)

I'm not 100% sure what happened next – probably we eyeballed the numbers and then picked some round numbers that were cheaper than everyone had been expecting to pay.

But it felt magical. It depended on a lot of trust in the group. I think it might also have somehow been subtly informed by our previous experience with second-price-auctions.

Life Lessons

There's not a single-definitive-moral here. But, possible morals include:

- **Coordination schemes are capital investments**. It's more work to use a new one, but once everyone in your ingroup has paid the upfront cost of understanding it, you get to use it over and over again.
- Being able to ***use* a scheme**, ***explain* a scheme**, and ***convince* people of a scheme**, are separate skills.
- **The Skill-Of-Understanding-New-Coordination Schemes is *also* a capital investment.** If you get good at understanding new systems when they're presented to you, you can reap the efficiency and fairness benefits immediately. (Corollary: gears-level understanding of coordination schemes are particularly valuable capital investments because they help you understand other nearby concepts)
- **Coordination is cognition-constrained.** It's often bottlenecked on the least sophisticated person. If one person doesn't get it, and needs to have it patiently explained to them, the increased efficiency might not be worth it in that

instance. (But it might be worth it for future iterations, if you expect to negotiate with those people again in the future).

- **Some skills generalize across coordination schemes.** In particular, the skill of "figure out how much you actually value something so you can make intelligent choices about how to bid for it" is going to come up in lots of different negotiation systems.
- **If you're a game theory nerd, you may continuously find yourself on the coordination frontier.** You may be forever having to patiently explain to people why they are leaving value on the table (for them, and/or for you). Or, sadly watching that value disappear. Alas. I recommend learning to grieve for it and accept it to some degree. But, also:
- **The skill of *Explaining and Leading* new coordination schemes is a *very valuable capital investment.*** If you constantly are noticing that there are more efficient ways people could be doing things, you should probably invest in communication skills and leadership skills to actually help enact those schemes. (Related: Coordination as a Scarce Resource)
- **Sometimes, you can get most of the value just by eyeballing it and going with your gut**. Shrug. Knowing when to do that is one of the relevant coordination skills. If you're trying to learn coordination-leadership, make sure not to lose sight of this. I've known many a nerd who was too attached to the dream of perfect efficiency and didn't notice it was coming at the expense of *actually being more efficient* in the current case.

High Level Takeaways

I think this all distills into two high level lessons, for two different target audiences:

For people vaguely annoyed at new complicated-seeming coordination plans:

Even if a scheme seems more complex-than-it's-worth right now, consider whether you'll get to keep using it again in the future. See if you can get excited about that future value, and channel that into willingness to learn.

The engineers pitching complicated schemes at you may not be very good at explaining them. Try to ask questions that help you understand the underlying principles, so you can learn similar systems in the future.

Perhaps also: try to ask questions that *help the engineers get better at explaining things*. (Especially if the engineer is your friend or colleague that you expect to work with for awhile)

For people who love designing coordination systems:

Be aware: you are not your audience.

The curse of many-an-engineer is to have poor insight into what their users actually want. Engineers overestimate how much complexity they can handle. If you spend a lot of time thinking about coordination theory, you may take a lot of skills and concepts for granted that others barely understand.

Your brilliant system might easily take more time to explain than it saves in generating value (in any given instance). So, if you want to actually be adding value

to the world, it either needs to be the case that…

1. You're actually going to use it multiple times with a given group, and therefore it's worth the upfront cost of everyone learning it.
2. The current use-case is a pretty big deal. i.e. if thousands of dollars are at stake, it's worth a few hours to optimize them. If a hundred dollars are at stake, it may not be.
3. You're going to use *similar* schemes with the same group, such that teaching people the underlying principles/skills is a worthwhile investment. (In this case, it's maybe worth more time to distill out those key skills, or reflect on what the future situations might be like so you can optimize for something more generalizable)
4. You personally (or, the broader world) will benefit from implementing similar schemes in the future. Or, you expect to learn from the process which results in better schemes in the future, even if this particular group isn't going to use them. In this case, recognize that you're expecting to gain value from other people's efforts, and it might be worth thinking of it more as a trade (where you offer them something in return for them putting extra time into learning the system)

A particular failure mode to avoid is when a system is easier for you, but not actually easier for others, and to not mistake "it's easier for me" with "actually net valuable for people who don't have my background knowledge."

A particular *success* mode to be in is to use novel coordination schemes as a way to build up the skills of people around you and make them better off longterm. But I think this requires a particular outlook, which doesn't come automatically.

Norm Innovation and Theory of Mind

Disclaimer: this was the first concept that led to me thinking about the coordination frontier. But I think something on the frame here feels subtly off. I decided to go ahead and post it – I'm pretty sure I believe all the words here. But not 100% sure this is the best way to think about the norm-negotiation problems.

Last post was about coordination schemes. Today's post is about a subset of coordination schemes: norms, and norm enforcement.

The internet is full of people unilateral enforcing new norms on each other, often based on completely different worldviews. Many people have (rightly, IMO) developed a defensiveness to getting accused of things they don't think are wrong.

Nonetheless, if society shall improve, it may be useful to invent (and enforce) new norms. What's a good way to go about that?

Ideally, I think people discuss new norms with each other *before* starting to enforce them. Bring them up at town hall. Write a thoughtful essay and get people to critique it or discuss potential improvements.

But often, norm-conflict comes up suddenly and confusingly. Someone violates what *you* thought was a foundational norm of your social circle, and you casually say "hey, you just did X". And they're like "yeah?" and you're flabbergasted that they're just casually violating what you assumed was an obvious pillar of society.

This is tricky even in the best of circumstances. You thought you could rely on a group following Norm X, and then it turns out if you want Norm X you have to advocate it yourself.

It's even more tricky when multiple people are trying to introduce new norms at once.

Multiplayer Norm Innovation

Imagine you have Alice, Bob, Charlie and Doofus, who all agree that you shouldn't steal from or lie to the ingroup, and you shouldn't murder anyone, ingroup or outgroup.

(Note the distinction between ingroups and outgroups, which matters quite a bit).

Alice, Bob, and Charlie *also* all agree that you should (ideally) aim to have a robust set of coordination meta-principles. But, they don't know much about what that means. (Doofus has no such aspirations. Sorry about your name, Doofus, this essay is opinionated)

One day Alice comes to believe: "Not only should you not *lie* to the ingroup, you also shouldn't use misleading arguments or cherry picked statistics to manipulate the ingroup."

Around the same time, Bob comes to believe: “Not only should you not steal from the ingroup, you also shouldn’t steal from the *outgroup*.” Trade is much more valuable than stealing cattle. Bob begins trying to convince people of this using misleading arguments and bad statistics.

Alice tells Bob “Hey, you shouldn’t use misleading arguments to persuade the ingroup of things because it harms our ability to coordinate.”

This argument makes perfect sense to Alice.

The next day, Bob makes another misleading argument at the ingroup.

Alice says “What the hell, Bob?”

The day after that, Bob catches Alice stealing cattle from their rivals across the river, and says “What the hell, Alice, didn’t you read my blogpost on why outgroup-theft is bad?”

Someday, I would like to have a principled answer to the question “What is the best way for all of these characters to interact?” In this post, I’d like to focus on one aspect of why-the-problem is hard.

Disclaimer: This example probably doesn't represent a coherent world. Clean examples be hard, yo.

Theory of Mind

The Sally Anne Marble test is a psychological tool for looking at how children develop theory-of-mind. A child is told a story about Sally and Anne. Sally has a marble. She puts it in her basket, and then leaves. While she’s away, her friend Anne takes the marble and hides it in another basket.

The child is asked “When Sally returns, where does she think her marble is?”

Very young children incorrectly answer “Sally will think the marble is in Anne’s basket.” The child-subject knows that Anne took the marble, and they don’t yet have the ability to model that Sally has different beliefs than they do.

Older children correctly answer the question. They have developed theory of mind.

“What the hell, Bob?”

When Alice says “what the hell, Bob?”, I think she's (sometimes) failing a more advanced theory of mind test.

Alice knows she told Bob “Hey, you shouldn’t use misleading arguments to persuade the ingroup of things because it harms our ability to coordinate.” This seemed like a complete explanation. But she is mismodeling a) how many assumptions she swept under the rug, and b) how hard it is to learn a new concept in the first place.

Sometimes the failure is even worse than that. Maybe Alice told Bob the argument. But then she runs into Bob’s friend, Charlie, who is also making misleading

arguments, and she doesn't even think to check if Charlie has been exposed to the argument *at all*. And she gets mad at Charlie, and then Charlie gets frustrated for getting called out on a behavior he's never even thought of before.

I've personally been the guy getting frustrated that nobody else is following "the obvious norms", when I *never even ever told someone the norm, let alone argued for it.* It just seemed to obviously follow from my background information.

Assuming Logical Omniscience

There are several problems all feeding into each other here. The first several problems are variations on "Inferential distance is a way bigger deal than you think", like:

- Alice expects she can explain something once in 5 minutes and it should basically work. But, if you're introducing a new way of thinking, it might take years to resolve a disagreement, because...
- Alice's claims are obvious to her within her model of the world. But, her frame might have lots of assumptions that aren't obvious to others.
- Alice may have initially explained her idea poorly, and Bob wrote her off as not-worth-listening to. (Idea Inoculation + Inferential Distance)
- Alice has spent tons of time thinking about how bad it is to make misleading arguments, to the point where it feels obviously wrong and distasteful to her. Bob has not done that, and Alice is having a hard time modeling Bob. She keeps expecting that aesthetic distaste to be present, and relying on it to do some rhetorical work that it doesn't do.
- Much of this is also present in the other direction. Bob is really preoccupied with getting people to stop stealing things, it seems obviously really important since right now there's an equilibrium where everyone is getting stolen from all the time. When Alice is arguing about being extra careful with arguments, Bob feels like she has a missing mood, like she doesn't understand why the equilibrium of theft is urgent. And *that* is downstream of Bob similarly underestimating the inferential gulf about why stealing your rival's cattle is limiting economic growth.

This all gets more complex when things have been going on for awhile. Alice and Bob both come to a (plausibly) reasonable belief that "Surely, I have made the case well enough that outgroup-theft/misleading-arguents are bad." They might even have reasonable evidence about this because people are making statements like "Theft is bad!" and "Misleading arguments are bad!".

But, nonetheless, Alice has thought about Misleading Arguments *a lot*. She is very attuned to it, whereas everyone else has just started paying attention. She have begun thinking multiple steps *beyond* that – building entire edifices that take the initial claims as a basic axiom, exploring deep into the coordination frontier, along different directions. Bob is having a similar experience re: Theft.

So they are constantly seeing people take actions that look like straightforward defections to them, and look like defections *they think other people have opted into being called on*, but actually require additional inferential steps that are not yet common knowledge nor consensus.

Attention, Mistrust, and Stag Hunts

Meanwhile, another problem here is that, even if Bob and Alice take each other's claims seriously, they might live in a world where lots of people are proposing norms.

Some of those norms are actively bad.

Some people are wielding norm-pushing as a weapon to gain social status or win political fights. (Even the people pushing *good* norms).

Some of the norms are good, but you can only prioritize so many new norms at once. Even people nominally on the same side may have different conceptions of what ingroup boundaries they are trying to draw, what standards they are trying to uphold, and whether a given degree of virtue is positive or negative for their ingroup.

People often model new norms as a stag hunt – if only we all pitched in to create a new societal expectation, we'd reap benefits from our collective action. Unfortunately, most stag hunts are actually schelling coordination games – the question is not "stag or no?", it's "which of the millions of stags are we even trying to kill?"

This all adds up to the unfortunate fact that the schelling choice is rabbit, not stag.

Attention resources are scarce. Not many people are paying attention to any given overton-window-fight. People get exhausted by having too many overton fights in a row. Within a single dispute, people have limited bandwidth before the cost of figuring out the optimal choice in the dispute doesn't seem worth it.

So when someone shows up promoting a new norm, there's a lot of *genuine reason* to be skeptical and react defensively.

Takeaways

This essay may seem kinda pessimistic about establishing new norms. But overall I think new norms are pretty important.

Once upon a time, we didn't have norms against stealing from the outgroup. Over time, we somehow got that norm, and it allowed us to reap massive gains through trade. The story was obviously not nearly so simplistic as Bob. Maybe people started with some incidental trade, and the norm developed in fits and spurts after-the-fact. Maybe merchants (who stood to benefit from the norm) actively promoted it in a self-interested fashion. Or, maybe ancient civilizations handled this largely via redefining ingroups. But somehow or other we got from there to here.

Once upon a time, we didn't even *have* statistics, let alone norms against misusing them to mislead people. Much of society is still statistically illiterate, so it's a hard norm to apply in all contexts. Shared use of statistics is a coordination scheme, which civilization is still in the process of capitally-investing-in.

Part of the point of having intellectual communities is to get on the same page about *novel* ways we can defect on the epistemic commons. So that we can learn not to. So we can push the coordination frontier forward.

(Or, with a more positive spin: part of the point of dedicated communities is to develop new positive skills and habits we can gain, where we can benefit tremendously if lots of people in a network share those skills.)

But this is tricky, because people might have conceptual disagreements about what. (Among people who care about statistics, there are disagreements about how to use them properly. I recently observed an honest-to-goodness fight between a frequentist and bayesian that drove this point home)

Multiplayer Norm Pioneering is legitimately hard

If you're the sort of person who's proactively looking for *better* societal norms, you should *expect* to constantly be running into people not understanding you. The more steps you are beyond the coordination baseline, the less agreement with your policies you should expect.

If you're in a community of people who are *collectively* trying to push the coordination frontier forward via new norms, you should expect to *constantly* be pushing it in different directions, resulting in misunderstandings. This can be a significant source of friction even when everyone involved is well intentioned, trying to cooperate. Part of that friction stems from the fact that we can't reliably tell who is trying to cooperate in improving the culture, and who is trying to get away with stuff.

I have some sense that there are good practices that norm-pioneers can have that make it easier to interact with each other. Ideally, I think when people who are trying to push society forward run into conflict with each other, they have a set of tools where that conflict is resolved as efficiently as possible.

I have some thoughts on how to navigate all this. But each of my thoughts ended up looking suspiciously like "here's a new norm", and I was wary of muddling this meta-level post with object level arguments.

For now, I just want to leave people with the point that developing new norms creates inferential gaps. Efficient coordination generally requires people to be on the same page about what they're coordinating on. It feels tractable to me to get on some meta-level cooperation among norm-pioneers, but exactly how to go about it feels like an unsolved problem.

Coordination Motivation: The Pandemic

I first started thinking about the meta-coordination 4 years ago, in the context of rationalists arguing about community norms. It seemed to me that people were getting into fights that involved a lot of wasted motion, and failing to accomplish what seemed like obvious shared goals.

For a few years, the bulk of my thought process was a vague, dissatisfied "surely we can do better than this, right?". Many of the people arguing eventually went off to focus on their individual orgs and didn't interact as much with each other. Maybe that was the right solution, and all this worrying about meta-coordination and norm arguments was just a distraction.

Then a pandemic hit. Coordination became much more practical and important to me, and the concept of coordination pioneering became more directly relevant.

Here were some issues that felt coordination-shaped to me. In this post, I'm speaking largely from my experiences with the Bay Area rationality community, but I think many of the issues generalize.

- **Negotiating policies and norms within a single household.** Do you lock down? If so, how do you go about it? What do you do if people disagree on how dangerous covid is, what practices are effective, or what's worth trading off for safety?
- **Community contract tracing.** If someone at a party later gets covid, are people entitled to share that information? How do we negotiate with each other about sharing that information? This includes concerns about privacy, public safety, and how to socially navigate trading those off against each other during a crisis.
- **Maintaining social connection.** This might involve negotiation with your housemates over covid policy, or the housemates of your friends. Even if you and a friend each live alone, figuring out what kind of contact to have during a pandemic is at least a two-player game.
- **Housemate swapping/matchmaking.** Housemates hadn't generally been selected for "having similar preferences of how to handle pandemics". There were several reasons people might have wanted to relocate. But people also had reason to not necessarily want to advertise that they were looking for new housemates – they might risk antagonizing their current roommates, or airing drama that was still unfolding. Switching houses is also an effortful, high cost decision that was difficult during an already stressful time.
- **Allocation of labor (intellectual and otherwise).** There was a lot of stuff to figure out, and to do. There was an initial flurry of activity as everyone scrambled to orient. I think there was a fair amount of duplicate labor, and a fair amount of labor allocated to "figure out wtf is up with the pandemic?" that could have been spent on people's day job or other non-pandemic personal projects.
- **Maintaining organizational sync**. Most organizations went remote. I think some organizations can do a decent job working remote, but I think it comes with costs. Some forms of communication translate easily to zoom, and some are much harder when you can't bring things up briefly without scheduling a call being A Whole Deal. This prompts two questions of "What were the best ways to

shift to remote?” as well as “Was it actually necessary to shift to fully remote? Could better coordinated orgs have found ways to stay in person without undue risk?”, or “Were there third options?”

From my perspective, these all feed into two primary goals:

- The physical and mental health of my social network.
- The capacity of the rationality and EA communities to continue doing important work. (In particular, this *could* have been a year where AI safety research made differential progress relative to AI capabilities research. But my sense is that this didn’t happen)

I think all the previous bullet points are meaty topics, that each warrant at least one blogpost worth of retrospective. I’m not sure which topics I’ll end up deep diving into. In this post, I wanted to give a broad overview of *why* coordination innovation feels so important to me.

“Coordination” is a somewhat vague word to cluster all those topics together with. I think, ultimately, it’s helpful if you can taboo “coordination”, and focus on individual problems and processes. But as I write this, I’m still in the process of thinking through exactly what went wrong, or what could have been improved, and how to cluster those problems/solutions/concepts. In some cases I think the issue was more like "actually making use of existing good practices for coordination (at the object level)", and in some cases I think metacoordination, and the coordination frontier, are more relevant.

What all of those items share is that they are multiplayer games. In each case, individuals made choices, but some good outcomes required multiple people to agree, or to make synergistic choices in tandem.

This blogpost is the first of a few posts for helping me organize my own thoughts.

There are a few frames that stand out to me to look at the situation:

- Skills that could have helped.
- Outlooks and orientation that could have helped.
- Systems that could have helped.
- Organizational structures or leadership that could have helped.

And then maybe a fairly different frameset around "*Who's 'we', exactly?*". I think there's multiple scales that it's worth looking at through a coordination lens – a couple individual people, a loose network of friends and colleagues, particular organizations, the vaguely defined "rationality community", and the broader structure of different cities, states, and countries.

Analogies to future crises

I expect to learn many things from a Pandemic Coordination Case Study, that I'd wish I'd known in 2020. But the most important question is "whether/how will this be relevant to future crises?"

It's possible there will literally be another pandemic in our lifetimes, and that many lessons will directly transfer.

My biggest current worry is "accelerating AI technology either disrupt the economy, and create situations of high-stakes negotiations, where *some* of the lessons from the pandemic transfer." There are different ways that this could play out (a few individuals within an organization, negotiations between leaders of organizations, government regulation, industry self-regulation, intergovernmental treaties).

And then, of course, there could be entirely novel crises that aren't currently on my radar.

Coordination Skills I Wish I Had For the Pandemic

In the previous post I noted the pandemic as a wakeup call for coordination-capabilities. There are a few different lenses to look at this through.

Today, I want to look through the lens of *skills I wish I had had.*

There are many pandemic-relevant skills that aren't (especially) related to coordination. It'd have been handy to have a strong background in epidemiology, for example. But here I want to focus in skills that I think would be generally useful, and which bear on solving coordination problems in novel circumstances.

Coordination Bottlenecked On Skills

Many day-to-day coordination activities require me to have skills.

For example, I might buy groceries. This is a coordination activity – it's only worth a farmer's time to grow extra food if a middleman is going to buy it and transport it. It's only worth a middleman's time to transport it if a grocery store will stock it. It's only worth the grocery store's time to stock it if people will pay for it. We don't know exactly what things people want. Me buying groceries sends a signal to the economic system to produce more apples and milk.

Participating in this system is a lot easier if I have a collection of skills, such as reading and speaking English, navigating a grocery aisle, doing basic arithmetic so I know how much my apples and milk will cost. It includes soft skills like "the social norms involved with talking to the cashier."

Those skills didn't come for free. Society invested in me having them. It could have invested in me having different skills (such as martial prowess), which would enable *different* coordination patterns.

During the 2 months of the pandemic, I found myself wishing I had several new skills that I hadn't previously developed. Each skill would have taken me a month or so to really wrap my head around if I were trying hard. The fact that I needed all five at once felt very overwhelming. I was struggling to be functional at all and executing on the skills I already had.

Each skill was something I think would have been beneficial in single-player mode. I have a speculative sense that if multiple people around me had them, it'd have enabled compound returns. If people *reliably* had them, it may have been possible to build more complex systems on top of them.

Knowing what I value

When I go to the grocery store and don't have enough money to buy both apples and milk, I have to decide which one I value more than the other. This usually isn't too

hard – my sense of “do I want apples or milk more?” is driven by short term feedback loops I’m pretty familiar with.

When I decide whether to accept a job at one company vs another one, I often have a much harder time knowing which I value more. Jobs are multidimensional, varying in pay, longterm skill growth, coworker rapport, etc. They are also high stakes, high investment decisions. It usually takes me several days or weeks to figure out which is preferable.

The pandemic threw me into a situation where many core pillars of my life were ripped out at the same time, while friends, roommates and coworkers were all having core pillars of *their* life ripped out at the same time. Values I’d normally think of as sacred, and not to be traded off, suddenly had to be traded off against each other. It also included *how I value the people around me*, and how I related to their values.

I didn’t know how much I valued my life, or what tradeoffs were worth making for it. I didn’t know what sacrifices I was willing to make for the sake of how other people valuing their lives, or their social lives.

Negotiating (and Maintaining Relationships) Under Stress

In the beginning, I tried to think carefully and negotiate with roommates about everything. But within 2 months I was exhausted of that, and my roommates were exhausted of that. And then a lot of what *could* have been fairly simple discussions ended up too painful and annoying to contemplate.

I’m least confident about how to improve at this skill. In Takeaways from one year of lockdown, Mingyuan notes:

> It's way harder to be a good rationalist in stressful situations… Negotiating in emotionally fraught situations is a very difficult skill, and despite all the training they receive in talking about feelings and what-not, being a CFAR instructor does not make you good at this skill (source: almost everyone in my house was a CFAR instructor or mentor).

But one of the central things seem to be “Be aware that you have a negotiation exhaustion budget. Try to have a sense of which things are actually worth negotiating over. Try to refactor complex social situations into simpler ones that require less negotiation.”

Grieving

At the beginning of the pandemic, I didn’t have much experience with grief. By the end of the pandemic, I had gotten a *lot* of practice grieving for things. I now think of grieving as a key life skill, with particular ramifications for coordination.

It might work differently for different people. But for me, grieving is the act of wrapping my brain around the fact that something important to me doesn’t exist anymore, or can’t exist right now, or perhaps never existed.

It contains two steps – an “orientation” step, where my brain traces around the lines of the thing-that-isn’t-there, coming to understand what reality is *actually* shaped like now. And then a “catharsis” step, once I fully understand that the thing is gone. The first step can take hours, weeks or months. For me, the second step tends to go quickly once I’ve fully processed.

You can grieve for people who are gone. You can grieve for things you enjoyed. You can grieve for principles that were important to you but aren’t practical to apply right now. You can grieve for concepts like “all of my friends and roommates can coexist happily.”

Grieving is important in single-player mode – if I’m holding onto something that’s not there anymore, my thoughts and decision-making are distorted. I can’t make good plans if my map of reality is full of leftover wishful markings of things that aren’t there.

I now think of this as relevant for coordination as well – if I’m hanging onto something that’s not real anymore, the distortion in my map also affects people who are trying to negotiate with me and find the least-bad-option available. My clinging becomes their problem.

Grieving is tricky because it’s often unclear when you’re supposed to grieve, and when you’re supposed to fight for something you still care about.

Grieving healthily takes time. But I now think grieving *healthily and quickly* is a skill you can learn. It does, unfortunately, require you to actually experience things that-need-grieving. The biggest things to grieve are (hopefully) rare.

Calibration

There was a whole bunch I didn’t know about the world, which was necessary to make informed choices.

Some uncertainties were about empirical facts in the external world, relating to covid, civil unrest, economic downturn. How likely am I to catch covid, or give it to my friends? How likely are we to die if we do? What exactly *is* civil unrest, and it is a thing I really need to worry about? Will looting increase? Will there be supply chain breakdowns?

Some uncertainties were about *myself*.

Would I be happier if I moved to the countryside for 6 months? Would I reflectively endorse it given my various commitments to friends, coworkers, and significant other?

Is it worth spending more time resolving conflicts between friends or coworkers about how to navigate the pandemic? If we make an agreement, will I turn out to endorse that agreement?

In all these cases, it’d be great to have perfect knowledge. Perfect knowledge is pretty expensive. But I think it’s a more achievable goal to have *calibrated* knowledge – I at least know what I know, and how wide my confidence intervals are.

After a few failed negotiations wherein I couldn't even tell what was worth negotiating for, I decided to boot up PredictionBook and start making predictions, so I could get a sense of my default calibration.

Numerical-Emotional Literacy (Or: "Scope Sensitivity")

I think the first few skills might be prerequisites for a kind of deep Numerical-Emotional literacy. (I'm not sure, because I do not yet have deep Numerical-Emotional-literacy)

I know how to use a spreadsheet. What I don't really know is how to connect a spreadsheet to my emotions and motivations.

In the first month of the pandemic, my house had been defaulting to "just do total lockdown", largely because it was conceptually simple. At some point a housemate said "what if we actually used a spreadsheet to make an informed fermi-model of how dangerous covid could be, and reflect on our values, and use that to consider whether we actually need to be this stringent about lockdown?"

And I agreed with that in principle. But... I just couldn't. I was so stressed out. I didn't have a principled way of valuing my life. I didn't trust myself to be able to do a fermi calc that I'd actually believe in. I didn't trust other roommates to do the fermi calc for me. I didn't have space to learn the skill in a way I *would* trust.

But, man, if I *had* had this skill, and if more of my friends had had it, it would have made a lot of things much easier. In particular because it would have meant we could...

Turning Sacred Values Into Trades

I think it often makes sense to have classes of things that you don't trade away, and that you drop everything to fix if they're threatened.

One of the complaints I heard during the pandemic was "I'd be willing to pay some people, like, $100s or $1000s of dollars for them to come to in-person meetings, but everyone is stuck in this mode where they're not willing to even consider it."

I was one of the people stuck in the mode where I couldn't even consider it. This was in large part due to my obligations to other housemates – everyone was burned out from negotiation and thinking about covid-risk.

At the time, I don't think there was much opportunity to improve on the situation. I think it's pretty harmful to pressure people into accepting deals that they don't feel comfortable making. At least in my corner of the social graph, I know that people were earnestly trying their best and operating with zero cognitive slack for months on end. But it left me wishing for a better world, a world where I, and my friends, *already* had the skills of:

- Having a concrete sense of how we valued our life – how much we'd pay for additional life-hours.

- Having a calibrated sense of how dangerous covid might have been (this could include wide error bars while still having a grounded sense of what the distribution would be, given your current information and your track record predicting things)
- Ability and willingness to recognize when things you previously classified as “key cornerstone of your life you don’t trade away” in fact need to get sacrificed, and the ability to do so with minimal trauma.

If I and several friends had started the pandemic with those skills, I think we’d have been in better positions to figure out where we actually disagreed with each other (as opposed to holing up by default). And then, if we actually disagreed about how much we each valued our lives vs social lives vs working-in-person-together, the additional act of “offer each other trades that are win-win” would have been less overwhelming.

Next post: Systems I Wish Were In Place For the Pandemic. AKA Could we have gotten microcovid.org sooner? Could we have had more numerical-emotional-literacy in the groundwater?