



Project Hufflepuff

1. [Project Hufflepuff: Planting the Flag](#)
2. [What exactly is the "Rationality Community?"](#)
3. [Straw Hufflepuffs and Lone Heroes](#)
4. [Background Reading: The Real Hufflepuff Sequence Was The Posts We Made Along The Way](#)
5. [Notes from the Hufflepuff Unconference](#)

Project Hufflepuff: Planting the Flag

"Clever kids in Ravenclaw, evil kids in Slytherin, wannabe heroes in Gryffindor, and everyone who does the actual work in Hufflepuff."

- *Harry Potter and the Methods of Rationality, Chapter 9*

"It is a common misconception that the best rationalists are Sorted into Ravenclaw, leaving none for other Houses. This is not so; being Sorted into Ravenclaw indicates that your strongest virtue is curiosity, wondering and desiring to know the true answer. And this is not the only virtue a rationalist needs. Sometimes you have to work hard on a problem, and stick to it for a while. Sometimes you need a clever plan for finding out. And sometimes what you need more than anything else to see an answer, is the courage to face it..."

- *Harry Potter and the Methods of Rationality, Chapter 45*

I'm a Ravenclaw and Slytherin by nature. I like being clever. I like pursuing ambitious goals. But over the past few years, I've been cultivating the skills and attitudes of Hufflepuff, by choice.

I think those skills are woefully under-appreciated in the Rationality Community. The problem cuts across many dimensions:

- Many people in rationality communities feel lonely (even the geographically tight Berkeley cluster). People want more (and deeper) connections than they currently have.
- There are lots of small pain points in the community (in person and online) that could be addressed fairly easily, but which people don't dedicate the time to fix.
- People are rewarded for starting individual projects more than helping to make existing ones succeed, which results in projects typically depending on a small number of people working unsustainably. (i.e. a single person running a meetup who feels like if they left, the meetup would crumble apart)
- Some newcomers often find the culture impenetrable and unwelcoming.
- Not enough "real-time operational competence" - the ability to notice problems in the physical world and solve them.
- Even at events like EA Global where enormous effort is put into operations and logistics, we scramble to pull things together at the last minute in a way that is very draining.
- Many people communicate in a way that feels disdainful and dismissive (to many people), which makes both social cohesion as well as intellectual understanding harder.
- We have a strong culture of "make sure your own needs are met", that specifically pushes back against broader societal norms that pressure people to conform. This is a good, but I think we've pushed too far in the opposite direction. People often make choices that are valuable to them in the immediate term, but which have negative externalities on the people around them.

In a nutshell, the emotional vibe of the community is preventing people from feeling happy and connected, and a swath of skillsets that are essential for group intelligence and ambition to flourish are undersupplied.

If any one of these things were a problem, we might troubleshoot it in isolated way. But collectively they seem to add up to a cultural problem, that I can't think of any way to express other than "Hufflepuff skills are insufficiently understood and respected."

There are two things I mean by "insufficiently respected":

- Ravenclaw and Slytherin skills come more naturally to many people in the community, and it doesn't even occur to people that emotional and operational skills are something they should cultivate. It feels like a separate magisteria that specialists should do. They're also quick to look at social niceties and traditions that seem silly, make a cursory attempt to understand them, and then do away with them without fully understanding their purpose.
- People who might join the community who value emotional and operational skills more highly, feel that the community is not for them, or that they have to work harder to be appreciated.

And while this is difficult to explain, it feels to me that there is a central way of being, that encompasses emotional/operational intelligence and deeply integrates it with rationality, that we are missing as a community.

This is the first in a series of posts, attempting to plant a flag down and say "Let's work together to try and resolve these problems, and if possible, find that central way-of-being."

I'm decidedly not saying "this is the New Way that rationality Should Be". The flag is not planted at the summit of a mountain we're definitively heading towards. It's planted on a beach where we're building ships, preparing to embark on some social experiments. We may not all be traveling on the same boat, or in the exact same direction. But the flag is gesturing in a direction that can only be reached by multiple people working together.

A First Step: The Hufflepuff Unconference, and Parallel Projects

I'll be visiting Berkeley during April, and while I'm there, I'd like to kickstart things with a Hufflepuff Unconference. We'll be sharing ideas, talking about potential concerns, and brainstorming next actions. (I'd like to avoid settling on a long term trajectory for the project - I think that'd be premature. But I'd also like to start building some momentum towards some kind of action)

My hope is to have both attendees who are positively inclined towards the concept of "A Hufflepuff Way", and people for whom it feels a bit alien. For this to succeed as a long-term cultural project, it needs to have buy-in from many corners of the rationality community. If people have nagging concerns that feel hard to articulate, I'd like to try to tease them out, and address them directly rather than ignoring them.

At the same time, I don't want to get bogged down in endless debates, or focus so much on criticism that we can't actually move forward. I don't expect total-consensus, so my goal for the unconference is to get multiple projects and social experiments running in parallel.

Some of those projects might be high-barrier-to-entry, for people who want to hold themselves to a particular standard. Others might be explicitly open to all, with radical inclusiveness part of their approach. Others might be weird experiments nobody had imagined yet.

In a few months, there'll be a followup event to check in on how those projects are going, evaluate, and see what more things we can try or further refine.

[Edit: The Unconference has been completed. [Notes from the conference are here](#)]

Thanks to Duncan Sabien, Lauren Horne, Ben Hoffman and Davis Kingsley for comments

What exactly is the "Rationality Community?"

[This is the second post in the [Project Hufflepuff sequence](#), which I no longer quite endorse in its original form. But this post is particularly standalone]

I used to use the phrase "Rationality Community" to mean three different things. Now I only use it to mean two different things, which is... well, a mild improvement at least. In practice, I was lumping a lot of people together, many of whom neither wanted to get lumped together nor had much in common.

As Project Hufflepuff took shape, I thought a lot about who I was trying to help and why. And I decided the relevant part of the world looks something like this:

The Rationalsphere



Note: Not to any scale, size and overlap of circles is determined entirely by what made the graph legible as opposed to actual proportion of what people care about.

I. The Rationalsphere

The Rationalsphere is defined in the broadest possible sense - a loose cluster of overlapping interest groups, communities and individuals. It includes people who disagree wildly with each other - some who are radically opposed to one another. It includes people who don't identify as "rationalist" or even as especially interested in "rationality" - but who interact with each other on a semi-regular basis. I think it's useful to be able to look at that ecosystem as a whole, and talk about it without bringing in implications of community.

There is no single feature defining people in the rationalosphere, but there are overlapping habits and patterns of thought. I'd guess that any two people in the cluster share at least one of the following features:

- Being attentive to ways your mind is unreliable
- Desire to understand objective reality.
- Willingness to change one's mind about ideas that are important to you.
- Having goals, which you care about achieving badly enough to decide "if my current habits of thought are an obstacle to achieving those goals, I want to prioritize changing those habits."
- Ambitious goals, that require a higher quality of decision-making than most humans have access to.

People invested in the rationalosphere seem to have three major motivations:

- **Truthseeking** - How do we improve our thinking? How do we use that improved-thinking to better understand the world?
- **Impact** - A lot of things in the world could be a lot better. Some see this in moral terms - there are suffering people and unrealized potential and we have an obligation to help. Others see this purely in opportunity and excitement, and find the concept of "altruism" offputting or harmful. But they share this: an interest in having a big impact, while understanding that 'having a big, intentional impact' is very hard. And confusing. Lots of people have tried, and failed. If we're to succeed, we will need better understanding and resources than we have now.
- **Human/Personal** - Your individual life and the people you know could also be a lot better. How can you and the people you love have as much fulfillment as possible?

For some people in the rationalosphere, "Doing a good job at being human" is a thing they're already doing and don't feel a need to approach from especially "rationality" flavored perspective, but still use principles (such as goal factoring) gleaned from the overall rationality project.

Others specifically do want to be part of a culture lets them succeed at Project Human that is uniquely "rationalist" - either because they want rationality principles percolating through their entire life, or because they like various cultural artifacts.

II. The Broader "Rationality Community"

Within the Rationalosphere, there is a subset of people that specifically want a community. They *also* disagree on a lot, but often want some combination of the following:

1. Social structures that make it easy to make friends, colleagues, and perhaps romantic partners, who also care about one or more of the three focus areas.
2. Social atmosphere that inspires and helps one to improve at one or more of the three focus areas.
3. Institutions that actively pursue one of the three in a serious fashion, and that collaborate when appropriate.
4. Sharing memes/culture/history. Feeling like "these are my people."

The overlapping social structures for each focus benefit each other. Here are some examples. (I want to note that I don't think all of these are unambiguously good. Some might trigger alarm bells, for good reason)

- The [Center for Applied Rationality](#) (CFAR) is able to develop techniques that help people communicate better, think more clearly, be more effective, and choose to work on more high-impact (I think even with their [more explicit shift](#) towards "help with AI", they will continue to have this effect in areas non-adjacent to AI).

- In addition to helping their alumni progress on their own truthseeking, impact and human-ing, CFAR leaves in its wake a community more energetic about trying additional experiments of their own.
- [Giving What We Can](#) encourages people to fund various projects (“EA” and non-EA) more seriously than they otherwise would - in a cluster of people who might otherwise fail to do so at all.
- Startup culture helps encourage people to launch ambitious projects of various stripes, which build people’s individual skills in addition to hopefully having a direct impact on the world of some sort.
- There are spaces where the Human and Truthseeking foci overlap, that create an environment friendly for people who like to think and talk deeply about complex concepts, for whom this is an important part of what they need to thrive as individuals. It’s really hard to find environments like this elsewhere.
- [Givewell](#) and the [Open Philanthropy Project](#) helps people concerned with Impact in obvious ways, but this in turn plays an important role for the Human focus - it gives people who don’t intrinsically care that much about effective altruism a way to contribute to it *without* taking too much of their attention. This is good for the world *and* their own sense of meaning and purpose. (Although I want to note that getting too attached to a particular source of meaning can be harmful, if it makes it harder to change your mind)
- Various other EA orgs that need volunteer work done provide an outlet for people who are not ready to jump-head-first into a major project, also providing sense-of-purpose as well as.
- Parties, meetups, etc (whether themed around Human-ing, Rationality, or EA) provide value to all three projects. They’re fun and mostly satisfy human-ing needs in the moment, but they let people bump into each other and swap ideas, network, etc, valuable for Impact and Understanding.
- A community need not be fully unified. Literal villages include people who disagree on religion, morality or policy - but they still come together to build marketplaces and community-centers and devise laws and guidelines on how to interact.

In addition to the “broader rationality community”, there are local groups that have more specific cultures and needs. (For example, NYC has a Less Wrong and Effective Altruism meetup group, which have different cultures both from each other, and from similar groups in Berkeley and Seattle)

This can include both physical-meet-space communities and some of the tighter knit online groups.

Where does Project Hufflepuff fit into this?

I think each focus area has seen credible progress in the past 10 years - both by developing new insights and by getting better at combining useful existing tools. I think we’ve gotten more and more glimpses of the [Something More That's Possible](#).

At our best, there's a culture forming that is clever, innovative, compassionate, and most all - takes ideas seriously.

But we're often not at our best. We make progress in fits and spurts. And there's a particular cluster of skills, surrounding interpersonal dynamics, that we seem to be systematically bad at. I think this is crippling the potential of all three focus areas.

We've [made progress over the past 10 years](#) - in the form of people writing individual blogposts, facebook conversations, dramatic speeches and just plain in-person-effort. This has helped shift the community - I think it's laid the groundwork for something like Project Hufflepuff being possible. But the thing about interpersonal dynamics is that they require

common knowledge and trust. We need to believe (accurately) that we can rely on each other to have each other's back.

This doesn't mean sacrificing yourself for the good of the community. But it means accurately understanding what costs you are imposing on other people - and therefore the costs they are imposing on you, and what those norms mean when you extrapolate them community-wide. And making a reflective decision on what kind of community we want to live so we can actually achieve our goals.

Since we don't all share the same goals and values, I expect this not to mean that community overall shifts towards some new set of norms. I'm hoping for individual people to think about the tradeoffs they want to make, and how those will affect others. And I suspect that this will result in a few different clusters forming, with different needs, solving different problems.

In the next post, I will start diving into the grittier details of what I think this requires, and why this is an especially difficult challenge.

Straw Hufflepuffs and Lone Heroes

I was hoping the next Project Hufflepuff post would involve more "explain concretely what I think we should do", but as it turns out I'm still hashing out some thoughts about that. In the meanwhile, this is the post I actually have ready to go, which is as good as any to post for now.

Epistemic Status: Mythmaking. This is tailored for the sort of person for whom the "Lone Hero" mindset is attractive. If that isn't something you're concerned with and this post feels irrelevant or missing some important things, note that my vision for Project Hufflepuff has multiple facets and I expect different people to approach it in different ways.

For good or for ill, the founding mythology of our community is a Harry Potter fanfiction.

This has a few ramifications I'll delve into at some point, but the most pertinent bit is: for a community to change itself, the impulse to change needs to come from within the community. I think it's easier to build change off of stories that are already a part of our cultural identity.*

** with an understanding that maybe part of the problem is that our cultural identity needs to change, or be more accessible, but I'm running with this mythos for the time being.*

In J.K Rowling's original Harry Potter story, Hufflepuffs are treated like "generic background characters" at best and as a joke at worst. All the main characters are Gryffindors, courageous and true. All the bad guys are Slytherin. And this is strange - Rowling clearly was setting out to create a complex world with nuanced virtues and vices. But it almost seems to me like Rowling's story takes place in an alternate, explicitly "Pro-Gryffindor propaganda" universe instead of the "real" Harry Potter world.

People have trouble taking Hufflepuff seriously, because they've never actually seen the real thing - only lame, strawman caricatures.

Harry Potter and the Methods of Rationality is... well, Pro-Ravenclaw propaganda. But part of being Ravenclaw is trying to understand things, and to use that knowledge. Eliezer makes an earnest effort to steelman each house. What wisdom does it offer that actually makes sense? What virtues does it cultivate that are rare and valuable?

When Harry goes under the sorting hat, it actually tries to convince him not to go into Ravenclaw, and specifically pushes towards Hufflepuff House:

Where would I go, if not Ravenclaw?

"Ahem. 'Clever kids in Ravenclaw, evil kids in Slytherin, wannabe heroes in Gryffindor, and everyone who does the actual work in Hufflepuff.' This indicates a certain amount of respect. You are well aware that Conscientiousness is just about as important as raw intelligence in determining life outcomes, you think you will be extremely loyal to your friends if you ever have some, you are not frightened by the expectation that your chosen scientific problems may take decades to solve -"

I'm lazy! I hate work! Hate hard work in all its forms! Clever shortcuts, that's all I'm about!

"And you would find loyalty and friendship in Hufflepuff, a camaraderie that you have never had before. You would find that you could rely on others, and that would heal something inside you that is broken."

But my plans -

"So replan! Don't let your life be steered by your reluctance to do a little extra thinking. You know that."

In the end, Harry chooses to go to Ravenclaw - the obvious house, the place that seemed most straightforward and comfortable. And ultimately... a hundred+ chapters later, I think he's still visibly lacking in the strengths that Hufflepuff might have helped him develop.

He does work hard and is incredibly loyal to his friends... but he operates in a fundamentally lone-wolf mindset. He's still manipulating people for their own good. He's still too caught up in his own cleverness. He never really has true friends other than Hermione, and when she is unable to be his friend for an extended period of time, it takes a huge toll on him that he doesn't have the support network to recover from in a healthy way.

The story does showcase Hufflepuff virtue. Hermione's army is strong precisely because people work hard, trust each other and help each other - not just in big, dramatic gestures, but in small moments throughout the day.

But... none of that ends up really mattering. And in the end, Harry faces his enemy alone. Lip service is paid to the concepts of friendship and group coordination, but the dominant narrative is Godric Gryffindor's Nihil Supernum:

No rescuer hath the rescuer.

No lord hath the champion.

No mother or father.

Only nothingness above.

The Sequences and HPMOR both talk about the importance of groups, of emotions, of avoiding the biases that plague overly-clever people in particular. But I feel like the communities descended from Less Wrong, as a whole, are still basically that eleven-year-old Harry Potter: abstractly understanding that these things are important, but not really believing in them seriously enough to actually change their plans and priorities.

Lone Heroes

In Methods of Rationality, there's a pretty good reason for Harry to focus on being a lone hero: he literally is alone. Nobody else really cares about the things he cares about or tries to do things on his level. It's like a group project in high school, which is supposed to teach cooperation but actually just results in one kid doing all the work while the others either halfheartedly try to help (at best) or deliberately goof off.

Harry doesn't bother turning to others for help, because they won't give him the help he needs.

He does the only thing he can do reliably: focus on himself, pushing himself as hard as he can. The world is full of impossible challenges and nobody else is stepping up, so he shuts up and does the impossible as best he can. Learning higher level magic. Learning higher level strategy. Training, physically and mentally.

This proves to be barely enough to survive, and not nearly enough to actually play the game. The last chapters are Harry realizing his best still isn't good enough, and no, this isn't fair, but it's how the world is, and there's nothing to do but keep trying.

He helps others level up as best they can. Hermione and Neville and some others show promise. But they're not ready to work together as equals.

And frankly, this does match my experience of the real world. When you have a dream burning in your heart... it is incredibly hard to find someone who shares it, who will not just pitch in and help but will actually move heaven and earth to achieve it.

And if they aren't capable, level themselves up until they are.

In my own projects, I have tried to find people to work alongside me and at best I've found temporary allies. And it is frustrating. And it is incredibly tempting to say "well, the only person I can rely on is myself."

But... here's the thing.

Yes, the world is horribly unfair. It is full of poverty, and people trapped in demoralizing jobs. It is full of stupid bureaucracies and corruption and people dying for no good reason. It is full of beautiful things that could exist but don't. And there are terribly few people who are able and willing to do the work needed to make a dent in reality.

But as long as we're willing to look at monstrously unfair things and roll up our sleeves and get to work anyway, consider this:

It may be that one of the unfair things is that one person can never be enough to solve these problems. That one of the things we need to roll up our sleeves and do even though it seems impossible is figure out how to coordinate and level up together and rely on each other in a way that actually works.

And maybe, while we're at it, find meaningful relationships that actually make us happy. Because it's not a *coincidence* that Hufflepuff is about both hard work *and* warmth and camaraderie. The warmth is what makes the hard work sustainable.

Godric Gryffindor has a point, but Nihil Supernum feels incomplete to me. There are no parents to step in and help us, but if we look to our left, or right...

Yes, you are only one

No, it is not enough—

But if you lift your eyes,

I am your brother

Vienna Teng, [Level Up](#)

-

Reminder that the Berkeley Hufflepuff Unconference is on April 28th. RSVPing on this [Facebook Event](#) is helpful, as is [filling out this form](#).

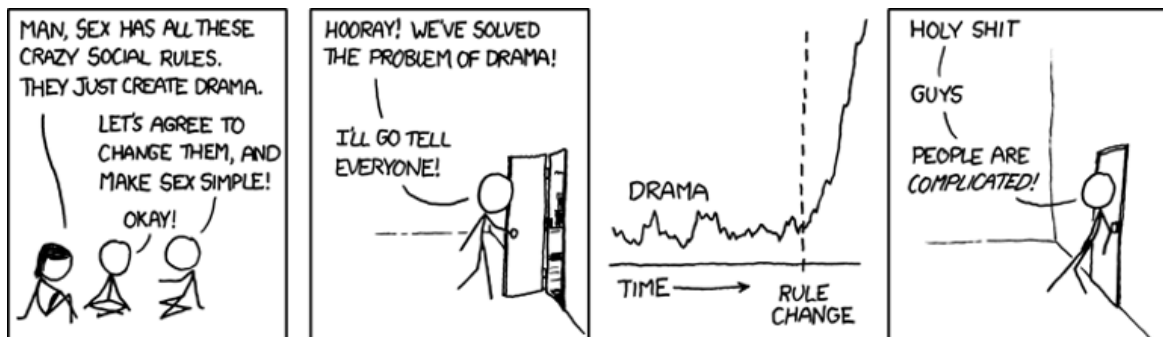
Background Reading: The Real Hufflepuff Sequence Was The Posts We Made Along The Way

This is the fourth post of the Project Hufflepuff sequence. Previous posts:

- [Introduction](#)
- [Who exactly is the Rationality Community?](#) (i.e. who is Project Hufflepuff trying to help)
- [Straw Hufflepuffs and Lone Heroes](#)

Epistemic Status: Tries to get away with making nuanced points about social reality by using cute graphics of geometric objects. All models are wrong. Some models are useful.

Traditionally, when nerds try to understand social systems and fix the obvious problems in them, they end up looking something like this:



Social dynamics is hard to understand with your system 2 (i.e. deliberative/logical) brain. There's a lot of subtle nuances going on, and typically, nerds tend to see the obvious stuff, maybe go one or two levels deeper than the obvious stuff, and miss that it's in fact 4+ levels deep and it's happening in realtime faster than you can deliberate. Human brains are pretty good (most of the time) at responding to the nuances intuitively. But in the rationality community, we've self-selected for a lot of people who:

1. Don't really trust things that they can't understand fully with their system 2 brain.
2. Tend not to be as naturally skilled at intuitive mainstream social styles.
3. Are trying to accomplish things that mainstream social interactions *aren't designed to accomplish* (i.e. thinking deeply and clearly on a regular basis).

This post is an overview of essays that rationalist-types have written over the past several years, that I think add up to a "secret sequence" exploring why social dynamics are hard, and why they are important to get right. This may be useful both to understand some previous attempts by the rationality community to change social dynamics on purpose, as well as to current endeavors to improve things.

(Note: I occasionally have words in [brackets], where I think original jargon was pointing in a misleading direction and I think it's worth changing)

To start with, a word of caution:

[Armchair sociology can be harmful \[link\]](#) - Ozy's post is pertinent - most essays below fall into the category of "armchair sociology", and attempts by nerds to understand and articulate social-dynamics that they aren't actually that good at. Several times when an

outsider has looked in at rationalist attempts to understand human interaction they've said "Oh my god, this is the blind leading the blind", and often that seemed to me like a fair assessment.

I think all the essays that follow are *useful*, and are pointing at something real. But taken individually, they're kinda like the blind men groping at the elephant, each coming away with the distinct impression an elephant is like a snake, tree, a boulder, depending on which aspect they're looking at.

[Fake Edit: Ozy informs me that they were specifically warning against amateur *sociology* and not *psychology*. I think the idea still roughly applies]

i. Cultural Assumptions of Trust

[Guess \[Infer\] Culture, Ask Culture, and Tell \[Reveal\] Culture \(link\)](#)

(by Malcolm Ocean)

Different people have different ways of articulating their needs and asking for help. Different ways of asking require different assumptions of trust. If people are bringing different expectations of trust into an interaction, they may feel that that trust is being violated, which can seem rude, passive aggressive or oppressive.

I'm listing this article, instead of numerous others about Ask/Guess/Tell, because I think: a) Malcolm does a good job of explaining how all the cultures work, and b) I think his presentation of *Reveal* culture is a good, clearer upgrade for Brienne's Tell culture, and I'm a bit sad it didn't seem to make it into the zeitgeist yet.

I also like the suggestion to call Guess Culture "Infer Culture" (implying a bit more about what skills the culture actually emphasizes).

[Guess Culture Screens for Trying to Cooperate \(link\)](#) (Ben Hoffman)

Rationality folk (and more generally, nerds), tend to prefer explicit communication over implicit, and generally see Guess culture as strictly inferior to Ask culture once you've learned to assert yourself.

But there is something Guess culture does which Ask culture doesn't, which is give you evidence of how much people understand you and are trying to cooperate. Guess cultures filters for people who have either invested effort into understanding your culture overall, or people who are good at inferring your own wants.

[Sharp Culture and Soft Culture \(link\)](#) (Sam Rosen)

[WARNING: It turned out lots of people thought this meant something different than what I thought it meant. Some people thought it meant soft culture *didn't involve giving people feedback or criticism at all*. I don't think Soft/Sharp are totally-naturally clusters in the first place, and the distinction I'm interested in (as applies to rationality-culture), is *how* you give feedback.

(i.e. "Dude, your art sucks. It has no perspective." vs "oh, cool. Nice colors. For the next drawing, you might try incorporating perspective", as a simplified example)

Somewhat orthogonal to Infer/Ask/Reveal culture is "Soft" vs "Sharp" culture. Sharp culture tends to have more biting humor, ribbing each other, and criticism. Soft culture tends to value kindness and social harmony more. Sam says that Sharp culture "values honesty more." Robby Bensinger counters in the comments: "My own experience is that sharp culture makes it more OK to be open about certain things (e.g., anger, disgust, power disparities,

disagreements), but less OK to be open about other things (e.g., weakness, pain, fear, loneliness, things that are true but not funny or provocative or badass)."




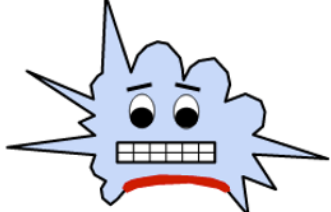
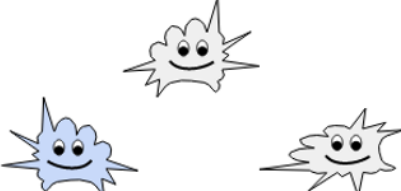
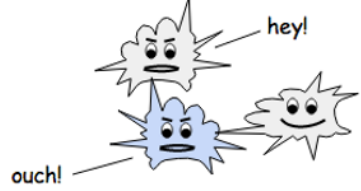




Handshakes, Hi, and What's New: What's Going on With Small Talk? (Ben Hoffman)

Small talk often sounds nonsensical to literally-minded people, but it serves a fairly important function: giving people a structured path to figure out how much time/sympathy/interest they want to give each other. And even when the answer is "not much", it still is, significantly, *nonzero* - you regard each other as persons, not faceless strangers.

Personhood [Social Interfaces?] (Kevin Simler)

This essays gets a lot of mixed reactions, much of which I think has to do with its use of the word "Person." The essay is aimed at *explaining* how people end up treating each other as persons or nonpersons, without making any kind of judgement about it. This includes noting some things human tend to do that you might consider horrible.

Like many grand theories, I think it overstates it's case and ignores some places where the explanation breaks down, but I think it points at a useful concept which is summarized by this adorable graphic:

<p>Here's a human animal. Say hi.</p>  <p>(Don't worry, there's nothing wrong with his face. This is an abstract representation.)</p>	<p>Like all humans, this one has</p>  <p>his goopy parts...</p>
<p>his prickly parts...</p> 	 <p>his soft sensitive underbelly parts... and many other parts.</p>
<p>Now humans, of course, like to spend time with each other...</p> 	<p>But animals in close proximity don't always get along so well.</p> 
<p>That's why we've taken to using our person masks:</p> 	<p>Wearing the masks makes getting along a <u>lot</u> easier.</p> 
<p>Just make sure you wear the right kind of mask....</p> 	<p>or the other animals will say nasty things about you.</p> 

The essay uses the word "personhood". In the original context, this was useful: it gets at why cultures develop, why it matters whether you're able to demonstrate reliably, trust, etc. It helps explain outgroups and xenophobia: outsiders do not share your social norms, so you

can't reliably interact with them, and it's easier to think of them as non-people than try to figure out how to have positive interactions.

But what I'm most interested in is "how can we use this to make it easier for groups with *different* norms to interact with each other"? And for that, I think using the word "personhood" makes it way more likely to veer into judging each other for having different preferences and communication styles.

What makes a person is... arbitrary, but not fully arbitrary.

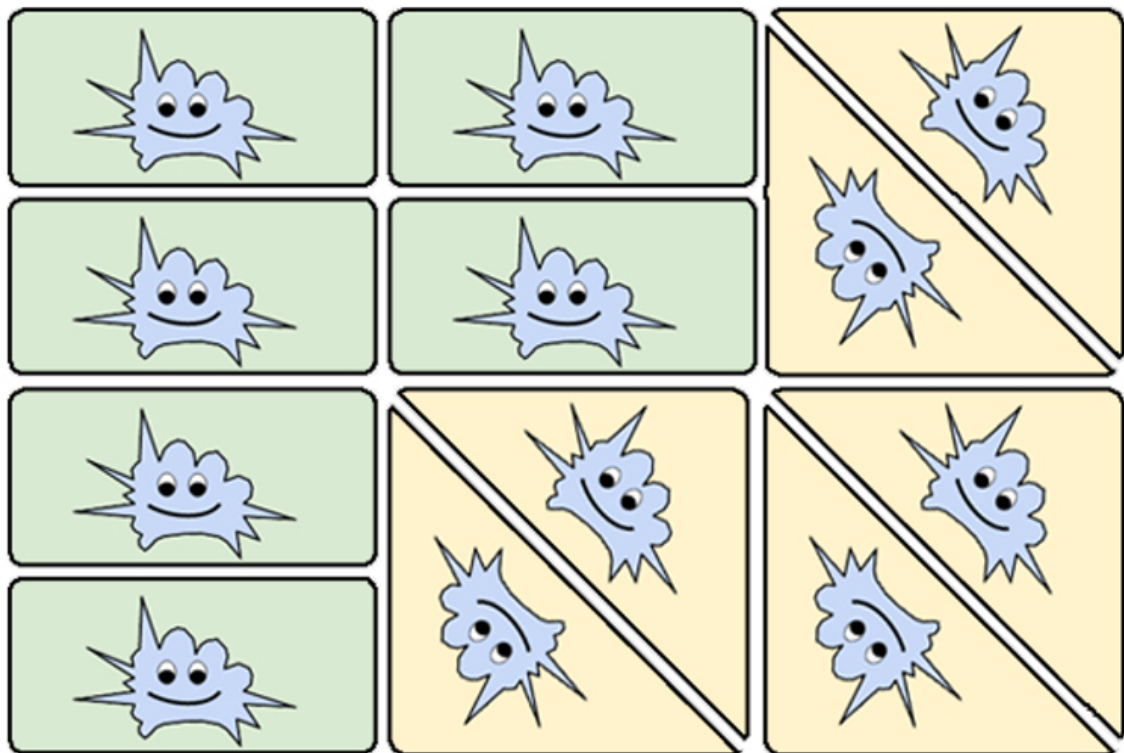
Rationalist culture tends to attract people who prefer a particular style of "social interface", often favoring explicit communication and discussing ideas in extreme detail. There's a lot of value to those things, but they have some problems:

a) this social interface does NOT mesh well with much of the rest of world (this is a problem if you have any goals that involve the rest of the world)

b) this goal does not uniformly mesh well with all people interested in and valuable to the rationality community.

I don't actually think it's possible to develop a set of assumptions that fit everyone's needs. But I do think it's possible to develop *better tools for navigating different social contexts*. I think it may be possible both to tweak sets-of-norms so that they mesh better together, or at least when they bump into each other, there's greater awareness of what's happening and people's default response is "oh, we seem to have different preferences, let's figure out how

Maybe we can end up with something that looks kinda like this:



[Against Being Against or For Tell Culture](#) (Brienne Yudkowsky)

Having said a bunch of things about different cultural interfaces, I think this post by Brienne is really important, and highlights the *end goal* of all of this.

"Cultures" are a crutch. They are there to help you get your bearings. They're better than nothing. But they are not a substitute for actually having the skills needed to navigate arbitrary social situations as they come up so you can achieve whatever it is you want to achieve.

To master communication, you can't just be like, "I prefer Tell Culture, which is better than Guess Culture, so my disabilities in Guess Culture are therefore justified." Justified shmjustified, you're still missing an arm.

My advice to you - my request of you, even - if you find yourself fueling these debates [about which culture is better], is to (for the love of god) move on. If you've already applied cognitive first aid, you've created an affordance for further advancement. Using *even more tourniquettes* doesn't help.

ii. Game Theory, Recursion, Trust

(or, "Social dynamics are complicated, you are not getting away with the things you think you are getting away with, stop trying to be clever, manipulative, act-utilitarian or naive-consequentialist without actually understanding what is going on")

[Grokking Newcomb's Problem and Deserving Trust](#) (Andrew Critch)

Critch argues why it is not just "morally wrong", but *an intellectual mistake*, to violate someone's trust (even when you don't expect any repercussions in the future).

When someone decides whether to trust you (say, giving you a huge opportunity), on the expectation that you'll refrain from exploiting them, they've already run a low-grade simulation of you in their imagination. And the thing is that you *don't know whether you're in a simulation or not when you make the decision whether to repay them*.

Some people argue "but I can tell that I'm a conscious being, and they aren't a literal super-intelligent AI, they're just a human. They can't possibly be simulating me in this high fidelity. I must be real." This is true. But their simulation of you is not based on your thoughts, it's based on your actions. It's *really* hard to fake.

One way to think about it, not expounded on in the article: Yes, if you pause to think about it you can notice that you're conscious and probably not being simulated in their imagination. But by the time you notice that, it's too late. People build up models of each other all the time, based on *very* subtle cues such as how *fast* you respond to something. *Conscious* you knows that you're conscious. But their decision of whether to trust you was based off the half-second it took for *unconscious* you to reply to questions like "Hey, do you think you handle Project X while I'm away?"

The best way to convince people you're trustworthy is to *actually be trustworthy*.

[You May Not Believe In Guess \[Infer\] Culture But It Believes In You](#) (Scott Alexander)

This is short enough to just include the whole thing:

Consider an "ask culture" where employees consider themselves totally allowed to say "no" without repercussions. The boss would prefer people work unpaid overtime so ey gets more work done without having to pay anything, so ey asks everyone. Most people say no, because they hate unpaid overtime. The only people who agree will be those who really love the company or their job - they end up looking really good. More and more workers realize the value of lying and agreeing to work unpaid overtime so the boss thinks they really love the company. Eventually, the few workers who continue refusing look really bad, like they're the only ones who aren't team players, and they grudgingly accept.

Only now the boss notices that the employees hate their jobs and hate the boss. The boss decides to only ask employees if they will work unpaid overtime when it's absolutely necessary. The ask culture has become a guess culture.

How this applies to friendship is left as an exercise for the reader.

The Social Substrate (Lahwran)

A fairly in depth look into how common knowledge, signaling, newcomb-like problems and recursive modeling of each other interact to produce "regular social interaction."

I think there's a lot of interesting stuff here - I'm not sure if it's exactly accurate but it points in directions that seem useful. But I actually think the most important takeaway is the warning at the beginning:

WARNING: An easy instinct, on learning these things, is to try to become more complicated yourself, to deal with the complicated territory. However, my primary conclusion is "simplify, simplify, simplify": try to make fewer decisions that depend on other people's state of mind. You can see more about why and how in the posts in the "Related" section, at the bottom.

When you're trying to make decisions about people, you're reading a lot of subtle cues off them to get a sense of how you feel about that. When you [generic person you, not necessarily *you* in particular] can tell someone is making complex decisions based on game theory and trying to model all of this explicitly, it a) often comes across as a bit *off*, and b) even if it doesn't, you still have to invest a lot of cognitive resources figuring out *how* they are modeling things and whether they are actually doing a good job or missing key insights or subtle cues. The result can be draining, and it can output a general response of "ugh, something about this feels untrustworthy."

Whereas when people are able to cache this knowledge down into a system-1 level, you're able to execute a simpler algorithm that looks more like "just try to be a good trustworthy person", that people can easily read off your facial expression, and which reduces overall cognitive burden.

System 1 and System 2 Morality (Sophie Grouchy)

There's some confusion over what "moral" means, because there's two kinds of morality:

- System 1 morality is noticing-in-realtime when people need help, or when you're being an asshole, and then doing something about it.
- System 2 morality is when you have a complex problem and a lot of time to think about it.

System 1 moralists will pay back Parfit's Hitchhiker because doing otherwise would be being a jerk. System 2 moralists invent Timeless *[Functional?]* decision theory.

You want a lot of people with System 2 morality in the world, trying to fix complex problems. You want people with System 1 morality in your social circle.

The person who wrote this post eventually left the rationality community, in part due to frustration due to people constantly violating small boundaries that seemed pretty obvious (things in the vein of "if you're going to be 2 hours late, text me so I don't have to sit around waiting for you.")

Final Remarks

I want to reiterate - all models are wrong. Some models are useful. The most important takeaway from this is not that any particular one of these perspectives is true, but that *social dynamics has a lot of stuff going on that is more complicated than you're naively imagining*, and that this stuff is *important enough to put the time into getting right*.

Notes from the Hufflepuff Unconference

April 28th, we ran the Hufflepuff Unconference in Berkeley, at the MIRI/CFAR office common space.

There's room for improvement in how the Unconference could have been run, but it succeeded the core things I wanted to accomplish:

- Established common knowledge of what problems people were actually interested in working on
- We had several extensive discussions of some of those problems, with an eye towards building solutions
- Several people agreed to work together towards concrete plans and experiments to make the community more friendly, as well as build skills relevant to community growth. (With deadlines and one person acting as project manager to make sure real progress was made)
- We agreed to have a followup unconference in roughly three months, to discuss how those plans and experiments were going

Rough notes [are available here](#). (Thanks to Miranda, Maia and Holden for taking really thorough notes)

This post will summarize some of the key takeaways, some speeches that were given, and my retrospective thoughts on how to approach things going forward.

But first, I'd like to cover a question that a lot of people have been asking about:

What does this all mean for people outside of the Bay?

The answer depends.

I'd personally *like* it if the overall rationality community got better at social skills, empathy, and working together, sticking with things that need sticking with (and in general, better at recognizing skills other than metacognition). In practice, individual communities can only change in the ways the people involved actually want to change, and there are other skills worth gaining that may be more important depending on your circumstances.

Does Project Hufflepuff make sense for your community?

If you're worried that your community *doesn't* have an interest in any of these things, my actual honest answer is that doing something "Project Hufflepuff-esque" probably does not make sense. I did not choose to do this because I thought it was the single-most-important thing in the abstract. I did it because it seemed important *and* I knew of a critical mass of people who I expected to want to work on it.

If you're living in a sparsely populated area or haven't put a community together, the first steps do not look like this, they look more like putting yourself out there, posting a meetup on Less Wrong and just **trying things**, any things, to get something moving.

If you have enough of a community to step back and take stock of what *kind* of community you want and how to strategically get there, I think this sort of project can be worth learning from. Maybe you'll decide to tackle something Project-Hufflepuff-like, maybe you'll find something else to focus on. I think the most important thing is have *some* kind of vision for something your community can do that is worth working together, leveling up to accomplish.

Community Unconferences as One Possible Tool

Community unconferences are a useful tool to get everyone on the same page and spur them on to start working on projects, and you might consider doing something similar.

They may not be the right tool for you and your group - I think they're most useful in places where there's enough people in your community that they *don't* all know each other, but *do* have enough existing trust to get together and brainstorm ideas.

If you have a sense that Project Hufflepuff is *worthwhile* for your community but the above disclaimers point towards my current approach not making sense for you, I'm interested in talking about it with you, but the conversation will look less like "Ray has ideas for you to try" and more like "Ray is interested in helping you figure out what ideas to try, and the solution will probably look very different."

Online Spaces

Since I'm actually very uncertain about a lot of this and see it as an experiment, I don't think it makes sense to push for any of the ideas here to directly change Less Wrong itself (at least, yet). But I do think a lot of these concepts translate to online spaces in some fashion, and I think it'd make sense to try out some concepts inspired by this in various smaller online subcommunities.

Table of Contents:

I. Introduction Speech

- Why are we here?
- The Mission: Something To Protect
- The Invisible Badger, or "What The Hell Is a Hufflepuff?"
- Meta Meetups Usually Suck. Let's Try Not To.

II. Common Knowledge

- What Do People Actually Want?
- Lightning Talks

III. Discussing the Problem (Four breakout sessions)

- Welcoming Newcomers
- How to handle people who impose costs on others?
- Styles of Leadership and Running Events
- Making Helping Fun (or at least lower barrier-to-entry)

IV. Planning Solutions and Next Actions

V. Final Words

I. Introduction: It Takes A Village to Save a World

(A more polished version of my opening speech from the unconference)

[Epistemic Status: This is largely based on intuition, looking at what our community has done and what other communities seem to be able to do. I'm maybe 85% confident in it, but it is

my best guess]

In 2012, I got super into the rationality community in New York. I was surrounded by people passionate about thinking better and using that thinking to tackle ambitious projects. And in 2012 we all decided to take on really hard projects that were pretty likely to fail, because the expected value seemed high, and it seemed like even if we failed we'd learn a lot in the process and grow stronger.

That happened - we learned and grew. We became adults together, founding companies and nonprofits and creating holidays from scratch.

But two years later, our projects were either actively failing, or burning us out. Many of us became depressed and demoralized.

There was nobody who was *okay* enough to actually provide anyone emotional support. Our core community withered.

I ended up making that the dominant theme of the 2014 NYC Solstice, with a call-to-action to get back to basics and take care each other.

I also went to the Berkeley Solstice that year. And... I dunno. In the back of my mind I was assuming "Berkeley won't have that problem - the Bay area has so many people, I can't even imagine how awesome and thriving a community they must have." (Especially since the Bay kept stealing all the movers and shakers of NYC).

The theme of the Bay Solstice turned out to be "Hey guys, so people keep coming to the Bay, running on a dream and a promise of community, but that community is not actually there, there's a tiny number of well-connected people who everyone is trying to get time with, and everyone seems lonely and sad. And we don't even know what to do about this."

Next year, in 2015, that theme in the Berkeley Solstice was revisited.

So I think that was the initial seed of what would become Project Hufflepuff - noticing that it's not enough to take on cool projects, that it's not enough to just get a bunch of people together and call it a community. Community is something you actively tend to. Insofar as Maslow's hierarchy is real, it's a foundation you need before ambitious projects can be sustainable.

There are other pieces of the puzzle - different lenses that, I believe, point towards a Central Thing. Some examples:

Group houses, individualism and coordination.

I've seen several group houses where, when people decide it no longer makes sense to live in the house, they... just kinda leave. Even if they've literally signed a lease. And everyone involved (the person leaving and those remain), instinctively act as if it's the remaining people's job to fill the leaver's spot, to make rent.

And the first time, this is kind of okay. But then each subsequent person leaving adds to a stressful undertone of "OMG are we even going to be able to afford to live here?". It eventually becomes depressing, and snowballs into a pit that makes newcomers feel like they don't WANT to move into the house.

Nowadays I've seen some people explicitly building into the roommate agreement a clear expectation of how long you stay and who's responsibility it is to find new roommates and pay rent in the meantime. But it's disappointing to me that this is something we *needed*, that we weren't instinctively paying to attention to how we were imposing costs on each other in the first place. That when we *violated a written contract*, let alone a handshake agreement,

that we did not take upon ourselves (or hold each other accountable), to ensure we could fill our end of the bargain.

Friends, and Networking your way to the center

This community puts pressure on people to improve. It's easier to improve when you're surrounded by ambitious people who help or inspire each other level up. There's a sense that there's some cluster of cool-people-who-are-ambitious-and-smart who've been here for a while, and... it seems like everyone is trying to be friends with those people.

It also seems like people just don't quite get that friendship is a skill, that adult friendships in City Culture can be hard, and it can require special effort to make them happen.

I'm not entirely sure what's going on here - it doesn't make sense to say anyone's obligated to hang out with any particular person (or obligated NOT to), but if 300 people aren't getting the connection they want it seems like *somewhere people are making a systematic mistake*.

(Since the Unconference, Maia has tackled this [particular issue in more detail](#))

The Mission - Something To Protect

As I see it, the Rationality Community has three things going on: Truth. Impact. And "Being People".

In some sense, our core focus is the practice of truthseeking. The thing that makes that truthseeking feel *important* is that it's connected to broader goals of impacting the world. And the thing that makes this actually fun and rewarding enough to stick with is a community that meets our needs, where we can both flourish as individuals and find the relationships we want.

I think we have made major strides in each of those areas over the past seven years. But we are nowhere near done.

Different people have different intuitions of which of the three are most important. Some see some of them as instrumental, or terminal. There are people for whom Truthseeking is **the point**, and they'd have been doing that even if there wasn't a community to help them with it, and there are people for whom it's just one tool of many that helps them live their life better or plan important projects.

I've observed a tendency to argue about which of these things is most important, or what tradeoffs are worth making. Inclusiveness versus high standards. Truth vs action. Personal happiness vs high achievement.

I think that kind of argument is a mistake.

We are falling *woefully* short on all of these things.

We need something like 10x our current capacity for seeing, and thinking. 10x our capacity for doing. 10x our capacity for **being healthy people together**.

I say "10x" not because all these things are intrinsically equal. The point is not to make a politically neutral push to make all the things sound nice. I have no idea exactly how far short we're falling on each of these because the targets are so far away I can't even see the end, and we are doing a complicated thing that doesn't have clear instructions and might not even be possible.

The point is that all of these are incredibly important, and if we cannot find a way to improve **all** of these, in a way that is **synergistic** with each other, then we will fail.

There is a thing at the center of our community. Not all of us share the exact same perspective on it. For some of us it's not the most important thing. But it's been at the heart of the community since the beginning and I feel comfortable asserting that it is the thing that shapes our culture the most:

The purpose of our community is to make sure this place is okay:

The world isn't okay right now, on a number of levels. And a lot of us believe there is a strong chance it could become dramatically less okay. I've seen people make credible progress on taking responsibility for pieces of our little blue home. But when all is said and done, none of our current projects really give me the confidence that things are going to turn out all right.

Our community was brought together on a promise, a dream, and we have not yet actually proven ourselves worthy of that dream. And to make that dream a reality we need a lot of things.

We need to be able to criticize, because without criticism, we cannot improve.

If we cannot, I believe we will fail.

We need to be able to talk about ideas that are controversial, or uncomfortable - otherwise our creativity and insight will be crippled.

If we cannot, I believe we will fail.

We need to be able to do those things without alienating people. We need to be able to criticize without making people feel untrusted and discouraged from even taking action. We need to be able to discuss challenging things while earnestly respecting the notion that *talking about ideas gives those ideas power and has concrete effects on social reality*, and sometimes that can hurt people.

If we cannot figure out how to do that, I believe we will fail.

We need more people who are able and willing to try things that have never been done before. To stick with those things long enough to *get good at them*, to see if they can actually work. We need to help each other do impossible things. And we need to remember to check for and do the *possible*, boring, everyday things that are in fact straightforward and simple and not very inspiring.

If we cannot manage to do that, I believe we will fail.

We need to be able to talk concretely about what the *highest leverage actions in the world are*. We need to prioritize those things, because the world is huge and broken and we are small. I believe we need to help each other through a long journey, building bigger and bigger levers, building connections with people outside our community who are undertaking the same journey through different perspectives.

And in the process, we need to not make it feel like if *you cannot personally work on those highest leverage things, that you are not important*.

There's the kind of importance where we recognize that some people have scarce skills and drive, and the kind of importance where we remember that **every* person has intrinsic worth*, and you owe **nobody** any special skills or prestigious sounding projects for your life to be worthwhile.

This isn't just a philosophical matter - I think it's damaging to our mental health and our collective capacity.

We need to recognize that the distribution of skills we tend to reward or punish is NOT just about which ones are actually most valuable - sometimes it is simply founder effects and blind spots.

We cannot be a community for everyone - I believe trying to include anyone with a passing interest in us is a fool's errand. But there are many people who had valuable skills to contribute who have turned away, feeling frustrated and un-valued.

If we cannot find a way to accomplish all of these things at once, I believe we will fail.

The thesis of Project Hufflepuff is that it takes (at least) a village to save a world.

It takes people doing experimental impossible things. It takes caretakers. It takes people helping out with unglorious tasks. It takes technical and emotional and physical skills. And while it *does* take some people who specialize in each of those things, I think it also needs many people who are least a *little* bit good at each of them, to pitch in when needed.

Project Hufflepuff is not the *only* things our community needs, or the most important. But I believe it is one of the *necessary* things that our community needs, if we're to get to 10x our current Truthseeking, Impact and Human-ing.

If we're to make sure that our home is okay.

The Invisible Badger

"A lone hufflepuff surrounded by slytherins will surely wither as if being leeches dry by vampires."

- Duncan

[Epistemic Status: My evidence for this is largely based on discussions with a few people for whom the badger seems real and valuable, and who report things being different in other communities, as well as some of my general intuitions about society. I'm 75% sure the badger exists, 90% that's it worth leaning into the idea of the badger to see if it works for you, and maybe 55% sure that it's worth trying to see the badger if you can't already make out it's edges.]

If I **had** to pick a clear thing that this conference is about without using Harry Potter jargon, I'd say "Interpersonal dynamics surrounding trust, and how those dynamics apply to each of the Impact/Truth/Human focuses of the rationality community."

I'm not super thrilled with that term because I think I'm grasping more for some kind of gestalt. An overall way of seeing and *being* that's hard to describe and that doesn't come naturally to the sort of person attracted to this community.

Much like the blind folk and the elephant, who each touched a different part of the animal and came away with a different impression (the trunk seems like a snake, the legs seem like a tree), I've been watching several people in the community try to describe things over the past few years. And maybe those things are separate but I feel like they're secretly a part of the same invisible badger.

Hufflepuff is about hard work, and loyalty, and camaraderie. It's about emotional intelligence. It's about seeing value in day to day things that don't directly tie into epic narratives.

There's a bunch of skills that go into Hufflepuff. And part of what I want is for people to get better at those skills. But I think it's a *mindset*, an *approach*, that is fairly different from the typical rationalist mindset, that makes those skills easier. It's something that's *harder* when you're being rigorously utilitarian and building models of the world out of game theory and incentives.

Mindspace is deep and wide, and I don't expect that mindset to work for everyone. I don't think everyone should be a Hufflepuff. But I do think it'd be valuable to the community if more people at least had access to this mindset and more of these skills.

So what I'd like, for tonight, is for people to lean into this idea. Maybe in the end you'll find that this doesn't work for you. But I think many people's first instinct is going to be that this is alien and uncomfortable and I think it's worth trying to push past that.

The reason we're doing this conference together is because the Hufflepuff way doesn't really work if people are trying to do it alone - I think it requires trust and camaraderie and persistence to really work. I don't think we can have that required trust all at once, but I think if there are multiple people trying to make it work, who can incrementally trust each other more, I think we can reach a place where things run more smoothly, where we have stronger emotional connections, and where we trust each other enough to take on more ambitious projects than we could if we're all optimizing as individuals.

Meta-Meetups Suck. Let's Not.

This unconference is pretty meta - we're talking about norms and vague community stuff we want to change.

Let me tell you, meta meetups are the worst. Typically you end up going around in circles complaining and wishing there were more things happening and that people were stepping up and maybe if you're lucky you get a wave of enthusiasm that lasts a month or so and a couple things happen but nothing really **changes**.

So. Let's not do that. Here's what I want to accomplish and which seems achievable:

1) Establish common knowledge of important ideas and behavior patterns.

Sometimes you DON'T need to develop a whole new skill, you just need to notice that your actions are impacting people in a different way, and maybe that's enough for you to decide to change somethings. Or maybe someone has a concept that makes it a lot easier for you to start gaining a new skill on your own.

2) Establish common knowledge of who's interested in trying which new norms, or which new skills.

We don't actually **know** what the majority of people want here. I can sit here and tell you what **I** think you should want, but ultimately what matters is what things a critical mass of people want to talk about tonight.

Not everyone has to agree that an idea is good to try it out. But there's a lot of skills or norms that only really make sense when a critical mass of other people are trying them. So, maybe of the 40 people here, 25 people are interested in improving their empathy, and maybe another 20 are interested in actively working on friendship skills, or sticking to commitments. Maybe those people can help reinforce each other.

3) Explore ideas for social and skillbuilding experiments we can try, that might help.

The failure mode of Ravenclaws is to think about things a lot and then not actually get around to doing them. A failure mode of *ambitious* Ravenclaws, is to think about things a lot and then do them and then assume that because they're smart, that they've thought of everything, and then not listen to feedback when they get things subtly or majorly wrong.

I'd like us to end by thinking of experiments with new norms, or habits we'd like to cultivate. I want us to frame these as *experiments*, that we try on a smaller scale and maybe promote more if they seem to be working, while keeping in mind that they may not work for everyone.

4) Commit to actions to take.

Since the default action is for them to peter out and fail, I'd like us to spend time bulletproofing them, brainstorming and coming up with trigger-action plans so that they actually have a chance to succeed.

Tabooing "Hufflepuff"

Having said all that talk about The Hufflepuff Way...

...the fact is, much of the reason I've used those words is to paint a rough picture to attract the sort of person I wanted to attract to this unconference.

It's *important* that there's a fuzzy, hard-to-define-but-probably-real concept that we're grasping towards, but it's *also* important not to be talking past each other. Early on in this project I realized that a few people who I thought were on the same page actually meant

fairly different things. Some cared more about empathy and friendship. Some cared more about *doing things together*, and expected deep friendships to arise naturally from that.

So I'd like us to establish a [trigger-action-plan](#) right now - for the rest of this unconference, if someone says "Hufflepuff", y'all should say "*What do you mean by that?*" and then figure out whatever concrete thing you're actually trying to talk about.

II. Common Knowledge

The first part of the unconference was about sharing our current goals, concerns and background knowledge that seemed useful. Most of the specifics are covered in the [notes](#). But I'll talk here about why I included the things I did and what my takeaways were afterwards on how it worked.

Time to Think

The first thing I did was have people sit and *think* about what they actually wanted to get out of the conference, and what obstacles they could imagine getting in the way of that. I did this because often, I think our culture (ostensibly about helping us think better) *doesn't* give us time to think, and instead has people who are quick-witted and conversationally dominant end up doing most of the talking. (I wrote a post a year ago about this, [the 12 Second Rule](#)). In this case I gave everyone 5 minutes, which is something I've found helpful at small meetups in NYC.

This had mixed results - some people reported that while they can think well by themselves, in a group setting they find it intimidating and their mind starts wandering instead of getting anything done. They found it much more helpful when I eventually let people-who-preferred-to-talk-to-each-other go into another room to talk through their ideas outloud.

I think there's some benefit to both halves of this and I'm not sure how common which set of preferences are. It's certainly true that it's not *common* for conferences to give people a full 5 minutes to think so I'd expect it to be someone uncomfortable-feeling regardless of whether it was useful.

But an overall outcome of the unconference was that it was somewhat lower energy than I'd wanted, and opening with 5 minutes of silent thinking seemed to contribute to that, so for the next unconference I run, I'm leaning towards a shorter period of time for private thinking (Somewhere between 12 and 60 seconds), followed by "turn to your neighbors and talk through the ideas you have", followed by "each group shares their concepts with the room."

"What is do you want to improve on? What is something you could use help with?"

I wanted people to feel like active participants rather than passive observers, and I didn't want people to just think "it'd be great if *other* people did X", but to keep an internal locus of control - *what can *I* do to steer this community better?* I also didn't want people to be thinking entirely individualistically.

I didn't collect feedback on this specific part and am not sure how valuable others found it (if you were at the conference, I'd be interested if you left any thoughts in the comments). Some anonymized things people described:

- When I make social mistakes, consider it failure; this is unhelpful

- Help point out what they need help with
- Have severe akrasia, would like more “get things done” magic tools
- Getting to know the bay area rationalist community
- General bitterness/burned out
- Reduce insecurity/fear around sharing
- Avoiding spending most words signaling to have read a particular thing; want to communicate more clearly
- Creating systems that reinforce unnoticed good behaviour
- Would like to learn how to try at things
- Find place in rationalist community
- Staying connected with the group
- Paying attention to what they want in the moment, in particular when it’s right to not be persistent
- Would like to know the “landing points” to the community to meet & greet new people
- Become more approachable, & be more willing to approach others for help; community cohesiveness
- Have been lonely most of life; want to find a place in a really good healthy community
- Re: prosocialness, being too low on Maslow’s hierarchy to help others
- Abundance mindset & not stressing about how to pay rent
- Cultivate stance of being able to do helpful things (action stance) but also be able to notice difference between laziness and mental health
- Don’t know how to respect legit safety needs w/o getting overwhelmed by arbitrary preferences; would like to model people better to give them basic respect w/o having to do arbitrary amount of work
- Starting conversations with new people
- More rationalist group homes / baugruppe
- Being able to provide emotional support rather than just logistics help
- Reaching out to people at all without putting too much pressure on them
- Cultivate lifelong friendships that aren’t limited to particular time and place
- Have a block around asking for help bc doesn’t expect to reciprocate; would like to actually just pay people for help w stuff
- Want to become more involved in the community
- Learn how to teach other people “ops skills”
- Connections to people they can teach and who can teach them

Lightning Talks

Lightning talks are a great way to give people an opportunity to not just share ideas, but get some practice at public presentation (which I've found can be a great gateway tool for overall confidence and ability to get things done in the community). Traditionally they are 5 minutes long. CFAR has found that 3.5 minute lightning talks are better than 5 minute talks, because it cuts out some rambling and tangents.

It turned out we had more people than I'd originally planned time for, so we ended up switching to two minute talks. I actually think this was even better, and my plan for next time is do 1-minute timeslots but allow people to sign up for multiple if they think their talk requires it, so people default to giving something short and sweet.

Rough summaries of the lightning talks can be [found in the notes](#).

III. Discussing the Problem

The next section involved two "breakout session" - two 20 minute periods for people to split into smaller groups and talk through problems in detail. This was done in an somewhat impromptu fashion, with people writing down the talks they wanted to do on the whiteboard and then arranging them so most people could go to a discussion that interested them.

The talks were:

- Welcoming Newcomers
- How to handle people who impose costs on others?
- Styles of Leadership and Running Events
- Making Helping Fun (or at least lower barrier-to-entry)
- Circling session

There was a suggested discussion about outreach, which I asked to table for a future unconference. My reason was that outreach discussions tend to get extremely meta and seem to be an attractor (people end up focusing on how to bring more people into the community without actually making sure the community is *good*, and I wanted the unconference to focus on the latter.)

I spent some time drifting between sessions, and was generally impressed both with the practical focus each discussion had, as well as the way they were organically moderated.

Again, more [details in the notes](#).

IV. Planning Solutions and Next Actions

After about an hour of discussion and mingling, we came back to the central common space to describe key highlights from each session, and begin making concrete plans. (Names are crediting people who suggested an idea and who volunteered to make it happen)

Creating Norms for Your Space (Jane Joyce, Tilia Bell)

The "How to handle people who impose costs on other" conversation ended up focusing on *minor but repeated* costs. One of the hardest things to moderate as an event host is not

people who are actively disruptive, but people who just a *little* bit awkward or annoying - they'd often be happy to change their behavior if they got feedback, but giving feedback feels uncomfortable and it's hard to do in a tactful way. This presents two problems at once: parties/events/social-spaces end up a more awkward/annoying than they need to be, *and* often what happens is that rather than giving feedback, the hosts stop inviting people doing those minor things, which means a lot of people still working on their social skills end up living in fear of being excluded.

Solving this fully requires a few different things at once, and I'm not sure I have a clear picture of what it looks like, but one stepping stone people came up with was creating explicit norms for a given space, and a practice of reminding people of those norms in a low-key, nonjudgmental way.

I think this will require a lot of deliberate effort and practice on the part of hosts to avoid alternate bad outcomes like "the norms get disproportionately enforced on people the hosts like and applied unfairly to people they aren't close with". But I do think it's a step in the right direction to showcase what kind of space you're creating and what the expectations are.

Different spaces can be tailored for different types of people with different needs or goals. (I'll have more to say about this in an upcoming post - doing this right is *really* hard, I don't actually know of any groups that have done an especially good job of it.)

I *was* impressed with the degree to which everyone in the conversation seemed to be taking into account a lot of different perspectives at once, and looking for solutions that benefited as many people as possible.

Welcoming Committee (Mandy Souza, Tessa Alexanian)

Oftentimes at events you'll see people who are new, or who don't seem comfortable getting involved with the conversation. Many successful communities do a good job of *explicitly* welcoming those people. Some people at the unconference decided to put together a formal group for making sure this happens more.

The exact details are still under development, but I think the basic idea is to have a network of people who are interested

he idea is to have a group of people who go to different events, playing the role of the welcomer. I think the idea is sort of a "Uber for welcomers" network (i.e. it both provides a place for people running events to go to ask for help with welcoming, and people who are interested in welcoming to find events that need welcomers)

It also included some ideas for better infrastructure, such as reviving "bayrationality.org" to make it easier for newcomers to figure out what events are going on (possibly including links to the codes of conduct for different spaces as well). In the meanwhile, some simple changes were the introduction of a facebook group for Bay Area Rationalist Social Events.

Softskill-sharing Groups (Mike Plotz and Jonathan Wallis)

The leadership styles discussion led to the concept that in order to have a flourishing community, and to be a successful leader, it's valuable to make yourself legible to others, and others more legible to yourself. Even small improvements in an activity as frequent as communication can have huge effects over time, as we make it easier to see each other as we actually are and to clearly exchange our ideas.

A number of people wanted to improve in this area together, and so we're working towards establishing a series of workshops with a focus on practice and individual feedback. A longer post on why this is important is coming up, and there will be information on the structure of

the event after our first teacher's meeting. If you would like to help out or participate, please fill out this poll:

<https://goo.gl/forms/MzkcsMvD2bKzXCQN2>

Circling Explorations (Qiaochu and others)

Much of the discussion at the Unconference, while focused on community, ultimately was explored through an intellectual lens. By contrast, "Circling" is a practice developed by the Authentic Relating community which is focused explicitly on feelings. The basic premise is (sort of) simple: you sit in a circle in a secluded space, and you talk about how you're feeling in the moment. Exactly how this plays out is a bit hard to explain, but the intended result is to become better both at noticing your own feelings and the people around you.

Opinions were divided as to whether this was something that made sense for "rationalists to do on their own", or whether it made more sense to visit more explicitly Circling-focused communities, but several people expressed interest in trying it again.

Making Helping Fun and More Accessible (Suggested by Oliver Habryka)

Ultimately we want a lot of people who are able and excited to help out with challenging projects - to improve our collective group ambition. But to get there, it'd be really helpful to have "gateway helping" - things people can easily pitch in to do that are fun, rewarding, clearly useful but on the "warm fuzzies" side of helping. Oliver suggested this as a way to get people to start identifying as people-who-help.

There were two main sets of habits that worth cultivating:

- 1) Making it clear to newcomers that they're encouraged to help out with events, and that this is actually a good way to make friends and get more involved.
- 2) For hosts and event planners, look for opportunities to offer people things that they can help with, and make sure to publicly praise those who do help out.

Some of this might dovetail nicely with the Welcoming Committee, both as something people can easily get involved with, and if there ends up being a public facing website to introduce people to the community, using that to connect people with events that could use help).

Volunteering-as-Learning, and Big Event Specific Workshops

Sometimes volunteering just requires showing up. But sometimes it requires special skills, and some events might need people who are willing to practice beforehand or learn-by-doing with a commitment to help at multiple events.

A vague cluster of skills that's in high demand is "predict logistical snafus in advance to head them off, and notice logistical snafus happening in realtime so you can do something about them." Earlier this year there was an Ops Workshop that aimed to teach this sort of skill, which went reasonably but didn't really lead into a concrete *use* for the skills to help them solidify.

One idea was to do Ops workshops (or other specialized training) in the month *before* a major event like Solstice or EA Global, giving them an opportunity to practice skills and making that particular event run smoother.

(This specific idea is not currently planned for implementation as it was among the more ambitious ones, although Brent Dill's series of "practice setting up a giant dome" beach parties in preparation for Burning Man are pointing in a similar direction)

Making Sure All This Actually Happens (Sarah Spikes, and hopefully everyone!)

To avoid the trap of dreaming big and not actually getting anything done, Sarah Spikes volunteered as project manager, creating an Asana page. People who were interested in committing to a deadline could opt into getting pestered by her to make sure things got done.

V. Parting Words

To wrap up the event, I focused on some final concepts that underlie this whole endeavor.

The thing we're aiming for looks something like this:

In a couple months (hopefully in July), there'll be a followup unconference. The theme will be "Innovation and Excellence", addressing the twofold question "how do we encourage more people to start cool projects", and "how do we get to a place where longterm projects ultimately reach a high quality state?"

Both elements feel important to me, and they require somewhat different mindsets (both on the part of the people running the projects, and the part of the community members who respond to them). Starting new things is scary and having too high standards can be really intimidating, yet for longterm projects we may want to hold ourselves to increasingly high standards over time.

My current plan (subject to lots of revision) is for this to become a series of community unconferences that happen roughly every 3 months. The Bay area is large enough with different overlapping social groups that it seems worthwhile to get together every few months and have an open-structured event to see people you don't normally see, share ideas, and get on the same page about important things.

Current thoughts for upcoming unconference topics are:

- Innovation and Excellence
- Personal Epistemic Hygiene
- Group Epistemology

An important piece of each unconference will be revisiting things at the previous one, to see if projects, ideas or experiments we talked about were actually carried out and what we learned from them (most likely with anonymous feedback collected beforehand so people who are less comfortable speaking publicly have a chance to express any concerns). I'd also like to build on topics from previous unconferences so they have more chance to sink in and percolate (for example, have at least one talk or discussion about "empathy and trust as related to epistemic hygiene").

Starting and Finishing Unconferences Together

My hope is to get other people involved sooner rather than later so this becomes a "thing we are doing together" rather than a "thing I am doing." One of my goals with this is also to provide a platform where people who are interested in getting more involved with community leadership can take a step further towards that, no matter where they currently stand (ranging anywhere from "give a 30 second lightning talk" to "run a discussion, or give a keynote talk" to "be the primary organizer for the unconference.")

I also hope this is able to percolate into online culture, and to other in-person communities where a critical mass of people think this'd be useful. That said, I want to caution that I consider this all an experiment, motivated by an intuitive sense that we're missing certain things as a culture. That intuitive sense has yet to be validated in any concrete fashion. I think "willingness to try things" is more important than epistemic caution, but epistemic

caution is still *really important* - I recommend collecting lots of feedback and being willing to shift direction if you're trying anything like the stuff suggested here.

(I'll have an upcoming post on "Ways Project Hufflepuff could go horribly wrong")

Most importantly, I hope this provides a mechanism for us to collectively take ideas more seriously that we're ostensibly supposed to be taking seriously. I hope that this translates into the sort of culture that [The Craft and The Community](#) was trying to point us towards, and, ideally, eventually, a concrete sense that our community can play a more consistently useful role at making sure the world turns out okay.

If you have concerns, criticism, or feedback, I encourage you to comment here if you feel comfortable, or on the [Unconference Feedback Form](#). So far I've been erring on the side of move forward and set things in motion, but I'll be shifting for the time being towards "getting feedback and making sure this thing is steering in the right direction."

-

In addition to the people listed throughout the post, I'd like to give particular thanks to Duncan Sabien for general inspiration and a lot of concrete help, Lahwran for giving the most consistent and useful feedback, and Robert Lecnik for hosting the space.