Best of LessWrong: March 2018

1. My attempt to explain Looking, insight meditation, and enlightenment in non-mysterious terms
2. The Costly Coordination Mechanism of Common Knowledge
3. An Untrollable Mathematician Illustrated
4. Problems with Amplification/Distillation
5. Making yourself small
6. Idea: Open Access AI Safety Journal
7. A Developmental Framework for Rationality
8. Extended Quote on the Institution of Academia
9. On Dualities
10. Explanation of Paul's AI-Alignment agenda by Ajeya Cotra
11. Naming the Nameless
12. New Paper Expanding on the Goodhart Taxonomy
13. On the Loss and Preservation of Knowledge
14. Defining the ways human values are messy
15. Every Implementation of You is You: An Intuition Ladder
16. Prize for probable problems
17. Caring less
18. Evaluating Existing Approaches to AGI Alignment
19. Timeline of Future of Humanity Institute
20. Browser Bug Hunt for LessWrong.com migration
21. Explicit and Implicit Communication
22. On Cognitive Distortions
23. Press Your Luck (1/3)
24. Environments for killing AIs
25. Silence
26. Funding for AI alignment research
27. Inference & Empiricism
28. Why mathematics works
29. Argument, intuition, and recursion
30. The hunt of the Iuventa
31. Ambiguity Detection
32. Learn Bayes Nets!
33. Hammertime Final Exam
34. Book Review: Consciousness Explained
35. Basic model of Sending a Message (Communication 101)
36. Defect or Cooperate
37. Avoiding AI Races Through Self-Regulation
38. Multiplicity of "enlightenment" states and contemplative practices
39. The Jordan Peterson Mask
40. [Preprint for commenting] Digital Immortality: Theory and Protocol for Indirect Mind Uploading
41. 'Trivial Inconvenience Day' Retrospective
42. Quick Nate/Eliezer comments on discontinuity
43. The Math Learning Experiment
44. Brains and backprop: a key timeline crux
45. Request for "Tests" for the MIRI Research Guide
46. AI Alignment Prize: Round 2 due March 31, 2018
47. Computational Complexity of P-Zombies
48. *Deleted*
49. Resolving human values, completely and adequately
50. On Defense Mechanisms

Best of LessWrong: March 2018

1. My attempt to explain Looking, insight meditation, and enlightenment in non-mysterious terms
2. The Costly Coordination Mechanism of Common Knowledge
3. An Untrollable Mathematician Illustrated
4. Problems with Amplification/Distillation
5. Making yourself small
6. Idea: Open Access AI Safety Journal
7. A Developmental Framework for Rationality
8. Extended Quote on the Institution of Academia
9. On Dualities
10. Explanation of Paul's AI-Alignment agenda by Ajeya Cotra
11. Naming the Nameless
12. New Paper Expanding on the Goodhart Taxonomy
13. On the Loss and Preservation of Knowledge
14. Defining the ways human values are messy
15. Every Implementation of You is You: An Intuition Ladder
16. Prize for probable problems
17. Caring less
18. Evaluating Existing Approaches to AGI Alignment
19. Timeline of Future of Humanity Institute
20. Browser Bug Hunt for LessWrong.com migration
21. Explicit and Implicit Communication
22. On Cognitive Distortions
23. Press Your Luck (1/3)
24. Environments for killing AIs
25. Silence
26. Funding for AI alignment research
27. Inference & Empiricism
28. Why mathematics works
29. Argument, intuition, and recursion
30. The hunt of the Iuventa
31. Ambiguity Detection
32. Learn Bayes Nets!
33. Hammertime Final Exam
34. Book Review: Consciousness Explained
35. Basic model of Sending a Message (Communication 101)
36. Defect or Cooperate
37. Avoiding AI Races Through Self-Regulation
38. Multiplicity of "enlightenment" states and contemplative practices
39. The Jordan Peterson Mask
40. [Preprint for commenting] Digital Immortality: Theory and Protocol for Indirect Mind Uploading
41. 'Trivial Inconvenience Day' Retrospective
42. Quick Nate/Eliezer comments on discontinuity
43. The Math Learning Experiment
44. Brains and backprop: a key timeline crux
45. Request for "Tests" for the MIRI Research Guide
46. AI Alignment Prize: Round 2 due March 31, 2018
47. Computational Complexity of P-Zombies

48. *Deleted*
49. Resolving human values, completely and adequately
50. On Defense Mechanisms

My attempt to explain Looking, insight meditation, and enlightenment in non-mysterious terms

Epistemic status: pretty confident. Based on several years of meditation experience combined with various pieces of Buddhist theory as popularized in various sources, including but not limited to books like The Mind Illuminated, Mastering the Core Teachings of the Buddha, and The Seeing That Frees; also discussions with other people who have practiced meditation, and scatterings of cognitive psychology papers that relate to the topic. The part that I'm the least confident of is the long-term nature of enlightenment; I'm speculating on what comes next based on what I've experienced, but have not actually had a full enlightenment. I also suspect that different kinds of traditions and practices may produce different kinds of enlightenment states.

While I liked Valentine's recent post on kensho and its follow-ups a lot, one thing that I was annoyed by were the comments that the whole thing can't be explained from a reductionist, third-person perspective. I agree that such an explanation can't produce the necessary mental changes that the explanation is talking about. But it seemed wrong to me to claim that all of this would be somehow intrinsically mysterious and impossible to explain on such a level that would give people at least an *intellectual* understanding of what Looking and enlightenment and all that are. Especially not after I spoke to Val and realized that hey, I actually *do* know how to Look, and that thing he's calling kensho, *that's happened to me too.*

(Note however that kensho is a Zen term and I'm unfamiliar with Zen; I don't want to use a term which might imply that I was going with whatever theoretical assumptions Zen might have, so I will just talk about "my experience" when it comes up.)

So here is my attempt to give an explanation. I don't know if I've succeeded, but here goes anyway.

One of my key concepts is going to be **cognitive fusion.**

Cognitive fusion is a term from Acceptance and Commitment Therapy (ACT), which refers to a person "fusing together" with the content of a thought or emotion, so that the content is experienced as an objective fact about the world rather than as a mental construct. The most obvious example of this might be if you get really upset with someone else and become convinced that something was *all their fault* (even if you had actually done something blameworthy too).

In this example, your anger isn't letting you see clearly, and you can't step back from your anger to question it, because you have become "fused together" with it and experience everything in terms of the anger's internal logic.

Another emotional example might be feelings of shame, where it's easy to experience yourself as a horrible person and feel that *this is the literal truth*, rather than being

just an emotional interpretation.

Cognitive fusion isn't *necessarily* a bad thing. If you suddenly notice a car driving towards you at a high speed, you don't want to get stuck pondering about how the feeling of danger is actually a mental construct produced by your brain. You *want* to get out of the way as fast as possible, with minimal mental clutter interfering with your actions. Likewise, if you are doing programming or math, you *want* to become at least partially fused together with your understanding of the domain, taking its axioms as objective facts so that you can focus on figuring out how to work with those axioms and get your desired results.

On the other hand, even when doing math, it can sometimes be useful to question the axioms you're using. In programming, taking the guarantees of your abstractions as literal axioms can also lead to trouble. And while it is useful to perceive something as objectively life-threatening and out to get you, that perception is going to get you in a lot of trouble if it's actually false. Such as if you get into a fight with your romantic partner and assume that they actively want to hurt you, when they're just feeling hurt over something that you said.

Cognitive fusion trades flexibility for focus. You will be strongly driven and capable of focusing on just the thing that's your in mind, at the cost of being less likely to notice when that thing is actually wrong.

Some simple defusion techniques suggested by ACT include things like noticing when you're thinking something bad about yourself, and prefacing it with "I'm having the thought that". So if you find yourself thinking "I am a terrible person", you can change that into "I'm having the thought that I am a terrible person". Or you can repeat the word "terrible" a hundred times, until it stops having any meaning. Or you can see if you can manipulate the way that the thought sounds like in your head, such as turning it into a comical whine that sounds like it's from a cartoon, until you can no longer take it seriously. (Eliezer's cognitive trope therapy should also be considered as a cognitive defusion technique.) In one way or the other, all of these highlight the fact that the thought or emotion is just a mental construct, making it easier to question its truthfulness.

However, managing to defuse from a thought that is actively bothering you, is a relatively superficial level of defusion. We must go deeper.

Meditation as cognitive defusion practice

While there are many different forms of meditation, many of them could be reasonably characterized as *practicing the skill of intentional cognitive defusion*.

One of the most basic forms of meditation is to just concentrate on your breath - or on any other focus that you have happened to choose. Soon, a distraction will come up in your mind - something that says that there's a more important thing to do, or that you are bored, or that this isn't leading anywhere.

If you start engaging with the content of that distraction, you're already failing to keep your focus. That is, if a thought comes to you saying that there's a more important thing to do, and you start arguing with yourself and trying to make a logical case for why meditation is actually the most important thing, then you've *already* been distracted from whatever it was that you were supposed to be focusing on. On some level, you have *bought into the internal logic of the distraction*, and into the belief that the argument must be beaten on its own terms.

What you must do instead, is to disregard the content of the distraction. Instead of becoming fused with its contents, defuse and redirect your attention back towards your focus. Whenever a new distraction rises, do this again.

As your skill improves and your attention becomes more reliably anchored on the focus, you can start learning additional skills. If you are doing something like the meditation program outlined in e.g. *The Mind Illuminated,* one of the next steps is to develop an awareness of distractions that are just on the edge of your consciousness, which are not *yet* distracting you but are going to steal your attention any moment now. By cultivating a sensitivity to those subtle movements of your mind, you are increasing your ability to notice lower-level details of what's going on in your consciousness, in a way which helps with cognitive defusion by making you more aware of the ways in which your experience is constructed.

As an example of such increased sensitivity, some time back I was doing concentration meditation, using an app which plays the sound of something hitting a woodblock, 50 times per minute. As I was concentrating on listening to the sound, I noticed that what had originally been just one thing in my experience - a discrete sound event - was actually composed of many smaller parts. The beginning and end of the sound were different, so there were actually two sound sensations; and there was a subtle visualization of something hitting something else; and a sense of motion accompanying that visualization. I had not previously even been fully aware that my mind was automatically creating a mental image of what it thought that the sound represented.

Continuing to observe those different components, I became more aware of the fact that my visualization of the sound *changed* over time and between meditation sessions, in a rather arbitrary way. Sometimes my mind conjured up a vision of a hammer hitting a rock in a dwarven mine; sometimes it was two wooden sticks hitting each other; sometimes it was drops of water falling on the screen of my phone.

By itself, this would mostly just be a curiosity. However, developing the kind of mental precision that actually lets you separate your experience into these kinds of small subcomponents, seems like a prerequisite for slicing your various mental outputs in a way which lets you see what they're made of.

Last summer, I noticed myself having the thought that I couldn't be happy, which made me feel bad. And then I noticed that associated with that thought, was a mental image of *what a happy person was like* - that image was of a young, cheerful, outgoing and extraverted girl.

In other words, my prototypical concept of a happy person included not just happiness, but extraversion and high energy as well. And so my mind was comparing my self-concepts with this concept of happiness, noticing that I wasn't that kind of a person, and so concluding that I couldn't be happy. Realizing that my concept of a "happy person" was uselessly narrow allowed me to fix the problem.

But if we break down what happened with the dysfunctional "happiness concept" into slightly smaller steps, something like this seems to have happened:

1) me feeling unhappy -> 2) mental image of a happy person -> 3) thought that I can't be happy

Notice that this has a similarity with the way my mind automatically produced a visualization for the woodblock sound:

1) sensation of the woodblock sound -> 2) mental image of two woodblocks hitting each other -> 3) thought of “oh, it’s two woodblocks hitting each other”

In both cases, some stimulus seemed to have produced a subtle mental image as a preliminary interpretation of what the stimulus meant, which then translated into a higher-level abstract concept. In both cases, something was off about the middle step. In the case of the happiness example, I had a too narrow view of what happy people are like. With the sound, the problem was that my mind was making up various interpretations of what was making the sound, despite having too little data to actually determine what it was.

Having developed the ability to notice those earlier steps in my mental processes, allowed me to notice a potential problem, as opposed to only being aware of the final output of the process.

I believe that this kind of thing is what Valentine means when he talks about Looking: being able to develop the necessary mental sharpness to notice slightly lower-level processing stages in your cognitive processes, and study the raw concepts which then get turned into higher-level cognitive content, rather than *only* seeing the high-level cognitive content.

This seems like a core rationality skill, since seeing slightly earlier stages of your cognitive process helps question its validity, which is to say it makes it easier for you to engage in cognitive defusion when desired. (If the process seems valid, you can still choose to fuse with it if that provides a benefit.) And being able to apply selective cognitive defusion means being able to not believe everything that you think, which is an essential requirement for things like actually changing your mind.

Understanding suffering

Understanding suffering is a special case of Looking, but a sufficiently important one that it deserves to be briefly discussed in some detail.

Usually, most of us are - on some implicit level - operating off a belief that we *need* to experience pleasant feelings and *need* to avoid experiencing unpleasant feelings. In a sense, thinking about getting into an unpleasant or painful situation may feel almost like death: if we think that the experience would be unpleasant enough, then no matter how brief it might be, we might do almost anything to avoid ending up there.

There’s a sense in which this is absurd. After all, a moment of discomfort is just that - a moment of discomfort. By itself, it won’t do us any lasting damage, and trying to avoid can produce worse results even on its own terms.

For instance, consider the person who keeps putting off making a doctor’s appointment because they suspect that there’s something wrong with them. If there really *is* something seriously wrong, then the best thing would be to get a diagnosis as fast as possible. And even if it is something harmless, it would still be better to find out about that earlier rather than later, so as to stop feeling the nervous about it. *Not going to the doctor, and continuing to feel nervous about it*, is about the worst possible outcome - *even if you cared about avoiding discomfort*.

On a conscious level, we realize that this kind of behavior is absurd. Then we go on doing it.

You might say that it's because there's a part of us that remains cognitively fused with the alief that all painful experiences *need* to be avoided, and that there's something vaguely death-like about them.

Typically, if we are only talking about relatively mild discomfort, then that alief doesn't manifest itself very strongly. We are okay with the thought of facing mild discomfort. But just as it's easy to remain calm and defused from feelings of anger as long as there isn't anything strongly upsetting going on, on some level we will tend to experience cognitive fusion with the "pain is death" alief more and more strongly the worse we expect the pain to be.

The general way by which incorrect aliefs are changed is by giving the part of your brain holding them, experiences about what the world is *really* like. If you have a dog phobia, you might do desensitization therapy, gradually exposing yourself to dogs in controlled circumstances. Eventually, seeing that you have encountered dogs many times and that it's safe, your brain updates and ceases to have the phobia.

Similarly, if you Look at the process of yourself flinching away from thoughts of painful experiences, you will come to *directly experience* the fact that it's the flinching away from them that actually produces suffering, and that the thoughts would be harmless by themselves.

The dog doesn't hurt you: it's your own fear that hurts you. Similarly, pain isn't bad by itself, but turns into suffering when we come to believe that we *need* to avoid it. Seeing this, the parts of your mind that have been doing the flinching away, will gradually start updating towards *not* habitually flinching away.

When I say that it is the automatic flinching away that actually produces suffering, I don't mean that just in the sense of "putting off painful experiences causes us to experience more pain in the long run". I mean that the processes involved with the flinching away are *literally* what turns pain into suffering: if you can get the flinching away to stop, pain (whether physical or emotional) will still be present as an attention signal that flags important things into your awareness. But neither the experience of pain, nor the *thought of* experiencing pain in the future, will be experienced as aversive anymore. The alief / belief of "pain is death" will not be active.

Now, Looking at your process-of-flinching-away in order to stop flinching away, is a long and slow process. We can again compare it with getting desensitized to a phobia: even after you have learned to be okay with a mild phobia trigger (say, a toy dog in the same room with you), you will continue to be freaked out by worse versions of the trigger (such as a real dog). It's very possible to have setbacks if a dog attacks you or if your life just generally gets more stressful, and sometimes you might show up at a session and get freaked out by things you *thought* you were already desensitized to. Learning to Look at suffering in order to reduce it is similar.

So what's all this "look up" and " get out of the car " stuff?

Here's an analogy.

Suppose that one day, you happen to run into a complete stranger. You don't think very much about needing to impress them, and as a result, you come off as relaxed and charming.

The next day, you're going on a date with someone you're really strongly attracted to. You feel that it's *really really important for you to make a good impression*, and

because you keep obsessing about this thought, you can't relax, act normal, and *actually* make a good impression.

Suppose that you remember all that stuff about cognitive fusion. You might (correctly) think that if you managed to defuse from the thought of this being an important encounter, then all of this would be less stressful and you might *actually* make a good impression.

But this brings up a particular difficulty: it can be relatively easy to defuse from a thought that you on some level believe is, or at least may be, *false*. But it's a lot harder to defuse from a thought *which you believe on a deep level to actually be true*, but which it's just counterproductive to think about.

After all, if you really *are* strongly interested in this person, but might not have an opportunity to meet with them again if you make a bad impression... then it *is* important for you to make a good impression on them now. Defusing from the thought of this being important, would mean that you believed less in this being important, meaning that you might do something that *actually* left a bad impression on them!

You can't defuse from the content of a belief, if your motivation for wanting to defuse from it is the belief itself. In trying to reject the belief that making a good impression is important, and trying to do this *with the motive of making a good impression*, you just *reinforce* the belief that this is important. If you want to *actually* defuse from the belief, your motive for doing so has to come from *somewhere else* than the belief itself.

The general form of this thing is what makes big green bats complain that you're still not getting out of the car. Or people who are aware of their cell phones, that you're still not looking up. You are fused with some belief or conceptual system *while trying to use that very same belief or conceptual system to defuse yourself from it*, which keeps you trapped in it. Instead, you could *just stop using it,* and *then* you'd be free.

Of course, this is easier said than done. Even if you *know* that this is what you're doing, knowing it isn't enough to stop doing it. Essentially, you have to somehow distract yourself from the belief you're caught up with… but if your belief is that *this thing is really important*, then before you could distract yourself from it, you'd need to distract yourself from it, so as to stop worrying about the potential consequences of having distracted yourself from it.

Yeah.

All of this *particularly* applies for trying to overcome suffering. Because remember, suffering is caused by a belief that pain is intrinsically bad. That belief is what causes you to try to flinch away from pain in a way which, by itself, creates the suffering.

So if you are experiencing some really powerful emotion that's causing you a lot of suffering, making you want to defuse from it so that you could stop feeling those bad things?

Well, then you are trying to be okay with feeling bad things, so that you could stop feeling bad things. Again, your motive for wanting to defuse from a belief, is digging you deeper into the belief.

On the surface, this would seem to suggest that you can only use Looking to stop suffering in cases of relatively mild pain, where you don't really even care all *that*

much about whether you're in pain or not. Looking would only help you feel better in the cases when you'd need it the least anyway.

And to be honest, a lot of the time it *does* feel that way.

Fortunately, there is a solution.

The three marks

I previously mentioned that there's something absurd about the belief that pain *would* need to be avoided: after all, if something really painful happens, then that won't kill us: usually it only means that, well, something really painful has happened. We might be left traumatized, but that trauma is by itself also just more pain.

It's as if a deep part of our minds is deluded about just how world-ending the pain is in the first place.

Buddhist theory states that that delusion arises from deep parts of our minds being wrong about some fundamental aspects of existence, traditionally called the three marks: impermanence, unsatisfactoriness, and no-self. If we can make ourselves curious about the true nature of existence, and Look deeply enough into just how our mind works, we can eventually witness things about how our mind works which contradict those delusions.

Do that often and deep enough, and the delusions shatter.

This allows us to actually overcome suffering, because in order to explore the nature of the self, we do not need to always be motivated by a desire to make the suffering stop. Rather, we can be motivated by things like curiosity or a desire to help other people, and explore the workings of our mind during times when we are *not* in terrible pain.

There will be a time when this happens on a sufficiently deep level that a person becomes convinced of full enlightenment being possible. Typically, the first time will be enough to let them get a taste of what it's like to live without delusions; but their insights are not yet deep enough to cause a permanent change, and the delusions will soon regenerate themselves.

Still, the delusions will not regenerate *entirely*: something will have shifted permanently, in a way that makes it easier to make further progress on dissolving them.

While it is impossible to use words to convey the experience of getting insight into the three marks of existence, it is possible to offer a third-person perspective on what exactly it is that our minds *are* mistaken about. Of the three marks, no-self may be the easiest to explain in these terms.

In the book *The Mind Illuminated*, the Buddhist model of psychology is described as one where our minds are composed of a large number of subagents, which share information by sending various percepts into consciousness. There's one particular subagent, the 'narrating mind' which takes these percepts and binds them together by generating a story of there existing one single agent, an I, to which everything happens. The fundamental delusion is when this fictional construct of an I is mistaken for an actually-existing entity, which needs to be protected by acquiring percepts with a positive emotional tone and avoiding percepts with a negative one.

When a person becomes capable of observing in sufficient detail the mental process by which this sense of an I is constructed, the delusion of its independent existence is broken. Afterwards, while the mind will continue to use the concept "I" as an organizing principle, it becomes correctly experienced as a theoretical fiction rather than something that could be harmed or helped by the experience of “bad” or “good” emotions. As a result, desire and aversion towards having specific states of mind (and thus suffering) cease. We cease to flinch away from pain, seeing that we do not need to avoid them in order to protect the “I”.

On why enlightenment may not be very visible in one’s behavior

In the comments of the *kensho* post, cousin_it mentioned having read several reports of people claiming enlightenment… yet not seeming to really demonstrate it by having better emotional skills. A paper also reported on various people having achieved some kinds of advanced meditative states… but still not being all that different when viewed from the outside:

> There seemed to be a clear distinction between a PNSE participant’s personality and their underlying sense of having an individualized sense of self. When the latter is absent, the former seems to be able to continue to function relatively unabated. There are exceptions. For example, the change in well-being in participants who were depressed prior to the onset of PNSE was obviously spotted by those around them. Generally, however, the external changes were not significant enough to be detected, even by those closest to the participant.

Based on how I experienced things when I had the experience that made enlightenment seem within reach, something like a lack of noticeable change is in fact *exactly what I would expect from many people who become enlightened* .

Remember, enlightenment means that you no longer experience emotional pain as aversive. In other words, you continue to have “negative” emotions like fear, anger, jealousy, and so on - you just don’t *mind* having them.

This *does* end up changing *some* of your emotional landscape. My experience was that since feeling crappy felt like an okay thing to happen, the thought of having negative experiences in the future no longer stressed me out. This brought with it a sense of calm, since I knew that I was in some sense "invulnerable" to anything that might happen. But the state of calmness was more of a result of everything being okay - a consequence of there no longer being anything that would be a genuine threat - rather than a permanent emotional state.

That emotion of calm could still be momentarily replaced by other emotional states as normal, it was just that one particular source of negative feelings (the fear of future negative feelings) was eliminated. I would still feel sadness about the things I normally feel sad about, anger about the things I normally feel angry about, and so on. *And because those emotions no longer felt aversive, I didn’t have a reason to invest in* ***not*** *feeling those things* ***-*** *unless I had some* ***other*** *reason than the intrinsic aversiveness of an emotion to do so.*

My model here is that enlightenment doesn’t automatically make you a good person, nor particularly emotionally balanced, or anything like that. If you were a jealous wreck before, but felt like it was totally justified and right for you to behave jealously… then seeing through the illusion of the self isn’t going to clear those cognitive structures from your head. It *can* help you defuse from them enough to see that your justifications are essentially arbitrary - but at the same time, you may *also* have

defused from any cognitive structures that say that there's something *bad* about having essentially arbitrary justifications.

To put it differently: one way of describing my experience was that it felt like an extreme moment of cognitive defusion, *where I defused from my entire motivational system*, and could just watch its operation from the outside.

But the thing is, *if you truly step outside your entire motivational system, then that leaves the part that just stepped out with **no** motivational system*, *leaving the existing one operating as normal.*

Suppose that you are thinking something like, "aha! stepping outside my whole motivational system means that I'm finally free to do thing X, which stupid internal conflicts have been blocking me from doing so far!"

But if you are thinking that, then you are still working inside a motivational system where it's important to achieve X. (Still not stepping out of the car.) If you have *truly* defused from your motivational system, then you have no particular desire to change the things in your mind that influence whether you are going to achieve X or not.

Even if you manage to step outside the system, the system is still going to keep doing various things - like taking your body to the store to get food - that it has learned to do: being defused from a motivation doesn't mean that the motivation would necessarily *disappear* or stop influencing your behavior. It just means that you can examine its validity as it goes on.

And if you see yourself going to the store to get some food, well, why *not* go along with that? After all, to *stop* acting as you always have, would require some special motivation to do so. All of your motivations exist *within* the system. If you previously had a motivation to change something about your own behavior, but also had underlying psychological reasons why you *hadn't* changed your behavior yet, then enlightenment may leave that balance of competing motivations basically unaltered. You may still have mental processes struggling against each other and you may experience internal conflict as normal: the only difference is that you won't *suffer* from that internal conflict.

Does this contradict the people who say that meditation will make you actively happy?

No: it only means that *Looking at the nature of suffering might not make you actively happy* (in the sense of experiencing lots of positive emotions). Remember that there are many things that you can Look at: meditation is essentially focusing your attention on *something*, and *what* you focus on makes a major difference.

I think in terms of meditative practices that work within an existing system (of pleasure and pain), versus ones that try to move you outside the system entirely. Some traditions focus on working inside the system, and may involve things like conditioning your mind for constant pleasure. Some systems combine the two, involving both practices which increase the amount of pleasure you'll experience, while *also* helping you be okay even with experiencing less pleasure. *The Mind Illuminated* takes this approach, for example.

And if enlightenment leaves your existing personality remains mostly intact, does it mean that Looking and meditation are useless for improving your rationality after all?

No. Again, it only means that *Looking at the things which cause suffering*, will not change your behavior as much as you might expect. Again, there are *many* different things about the functioning of your mind that you can Look at. And getting to the point where're you're enlightened, requires training up a *lot* of mental precision which you can then use to Look at various things.

Even if you do manage to defuse from everything that causes you suffering, your existing personality and motivational system will still be in charge of what it is that you Look at in the future. If all you cared about was ceasing to suffer, well, you're done! You might not have the motivation to do any more Looking on top of that, since it already got you what you wanted. You'll just go on living as normal, with your existing personality.

But if you cared about things like saving the world, then you will still continue to work on saving the world, and you will be Looking at things which will help you save the world - including ones that increase your rationality.

It's just that if the world ends up ending, it won't feel like the end of the world.

Of course, you *will* still feel intense grief and disappointment and everything that you'd expect to feel about the world ending.

Intense grief and disappointment just won't be the end of the world.

[Edited to add: for my more detailed, later explanation of this topic, see the series of posts starting from A non-mystical explanation of insight meditation and the three characteristics of existence.]

The Costly Coordination Mechanism of Common Knowledge

Recently someone pointed out to me that there was no good canonical post that explained the use of common knowledge in society. Since I wanted to be able to link to such a post, I decided to try to write it.

The epistemic status of this post is that I hoped to provide an explanation for a standard, mainstream idea, in a concrete way that could be broadly understood rather than in a mathematical/logical fashion, and so the definitions should all be correct, though the examples in the latter half are more speculative and likely contain some inaccuracies.

Let's start with a puzzle. What do these three things have in common?

- Dictatorships all through history have attempted to suppress freedom of the press and freedom of speech. Why is this? Are they just very sensitive? On the other side, the leaders of the Enlightenment fought for freedom of speech, and would not budge an inch against this principle.
- When two people are on a date and want to sleep with each other, the conversation will often move towards but never *explicitly* discuss having sex. The two may discuss going back to the place of one of theirs, with a different explicit reason discussed (e.g. "to have a drink"), *even* if both want to have sex.
- Throughout history, communities have had religious rituals that look very similar. Everyone in the village has to join in. There are repetitive songs, repetitive lectures on the same holy books, chanting together. Why, of all the possible community events (e.g. dinner, parties, etc) is this the most common type?

What these three things have in common, is *common knowledge* - or at least, the attempt to create it.

Before I spell that out, we'll take a brief look into game theory so that we have the language to describe clearly what's going on. Then we'll be able to see concretely in a bunch of examples, how common knowledge is necessary to understand and build institutions.

Prisoner's Dilemmas vs Coordination Problems

To understand why common knowledge is useful, I want to contrast two types of situations in game theory: Prisoner's Dilemmas and Coordination Problems. They look similar at first glance, but their payoff matrices have important differences.

The Prisoner's Dilemma (PD)

You've probably heard of it - two players have the opportunity to cooperate, or defect against each other, based on a story about two prisoners being offered a deal if they testify against the other.

If they do nothing they will put them both away for a short time; if one of them snitches on the other, the snitch gets off free and the snitched gets a long sentence. However if they *both* snitch they get pretty bad sentences (though neither are as long as when only one snitches on the other).

In game theory, people often like to draw little boxes that show two different people's choices, and how much they like the outcome. Such a diagram is called a *decision matrix*, and the numbers are called the players' *payoffs*.

To describe the Prisoner's Dilemma, below is a decision matrix where Anne and Bob each have the same two choices, labelled C and D. These are colloquially called ‘cooperate’ and ‘defect’. Each box contains two numbers, for Anne and Bob's payoffs respectively.

		Anne: C	Anne: D
Bob	C	**3 / 3**	**4 / 0**
Bob	D	**0 / 4**	**1 / 1**

If the prisoner ‘defects’ on his partner, this means he snitches, and if he ‘cooperates’ with his partner, he doesn’t snitch. They’d both prefer that *both* of them cooperate (C, C) to both of them defecting (D, D), but each of them has an incentive to stab each other in the back to reap the most reward (D, C).

Do you see in the matrix how they both would prefer no snitching to both snitching, but they also have an incentive to stab each other in the back?

Real World Examples

Nuclear disarmament is a prisoner’s dilemma. Both the Soviet Union and the USS wanted to have nuclear bombs while the opponent doesn't, but they'd probably both prefer a world where nobody had bombs than a world where they were both pointing massive weapons at each others heads. Unfortunately in our world, we failed to solve the problem, and ended up pointing massive weapons at each others' heads for decades.

Military budget spending more broadly can be a prisoner’s dilemma. Suppose two neighbouring countries are determining how much to spend on the military. Well, they don’t want to go to war with each other, and so they’d each like to spend a small amount of money on their military, and spend the rest of the money on running the country - infrastructure, healthcare, etc. However, if one country spends a small amount and the other country spends a lot, then the second country can just walk in and take over the first. So, they both spend lots of money on the military with no intention of using it, just so the other one can’t take over.

Another prisoner’s dilemma is tennis players figuring out whether to take performance enhancing drugs. Naturally, they'd like to dope and the opposing player not, but they'd rather both not dope than both dope.

Free-Rider Problems

Did you notice how there are more than two tennis players in the doping situation? When deciding whether to take drugs, not only do you have to worry about whether your opponent in the match today will dope, but also whether your opponent tomorrow will, and the day after, and so on. We're all wondering whether *all* of us will dope. In society there are loads of these scaled up versions of the prisoner's dilemma.

For example, according to many political theories, everyone is better off if the government takes some taxes and uses them to provide public goods (e.g. transportation, military, hospitals). As a population, it's in everyone's interest if everyone cooperates, and takes a small personal sacrifice of wealth.

However, if most people are doing it, you can defect, and this is great for you - you get the advantage of a government providing public goods, and also you keep your own money. But if everyone defects, then nobody gets the important public goods, and this is worse for each person than if they'd all cooperated.

Whether you're two robbers, one of many tennis players, or a whole country fighting another country, you will run into a prisoner's dilemma. In the scaled-up version, a person who defects while everyone else cooperates is known as a *free-rider*, and the scaled up prisoner's dilemma is called the *free-rider problem*.

Coordination Problems

With that under our belt, let's look at a new decision matrix. Can you identify what's importantly different about this matrix? Make a prediction about how you think this will change the players' strategies.

		Anne: C	Anne: D
Bob	C	**3 / 3**	**0 / 0**
	D	**0 / 0**	**1 / 1**

Don't mix this up with the Prisoners' Dilemma - it's quite different. In the PD, if you cooperate and I defect, I get 4. What's important about the new decision-matrix, is that nobody has an incentive to backstab! If you cooperate and I defect, I get *zero*, instead of four.

We all want the same thing. Both players' preference ordering is:

$$(C, C) > (D, D) > [(C, D) = (D, C)]$$

So, you might be confused: Why is this a problem at all? Why doesn't everyone just pick C?

Let me give an example from Michael Chwe's classic book on the subject *Rational Ritual: Culture, Coordination and Common Knowledge*.

Say you and I are co-workers who ride the same bus home. Today the bus is completely packed and somehow we get separated. Because you are standing near the front door of the bus and I am near the back door, I catch a glimpse of you only at brief moments. Before we reach our usual stop, I notice a mutual acquaintance, who yells from the sidewalk, “Hey you two! Come join me for a drink!” Joining this acquaintance would be nice, but we care mainly about each other’s company. The bus doors open; separated by the crowd, we must decide independently whether to get off.

Say that when our acquaintance yells out, I look for you but cannot find you; I’m not sure whether you notice her or not and thus decide to stay on the bus. How exactly does the communication process fail? There are two possibilities. The first is simply that you do not notice her; maybe you are asleep. The second is that you do in fact notice her. But I stay on the bus because I don’t know whether you notice her or not. In this case we both know that our acquaintance yelled but I do not know that you know.

Successful communication sometimes is not simply a matter of whether a given message is received. It also depends on whether people are aware that other people also receive it. In other words, it is not just about people’s knowledge of the message; it is also about people knowing that other people know about it, the “metaknowledge” of the message.

Say that when our acquaintance yells, I see you raise your head and look around for me, but I’m not sure if you manage to find me. Even though I know about the yell, and I know that you know since I see you look up, I still decide to stay on the bus because I do not know that you know that I know. So just one “level” of metaknowledge is not enough.

Taking this further, one soon realizes that every level of metaknowledge is necessary: I must know about the yell, you must know, I must know that you know, you must know that I know, I must know that you know that I know, and so on; that is, the yell must be “common knowledge.”

The term “common knowledge” is used in many ways but here we stick to a precise definition. We say that an event or fact is common knowledge among a group of people if everyone knows it, everyone knows that everyone knows it, everyone knows that everyone knows that everyone knows it, and so on.

Two people can create these many levels of metaknowledge simply through eye contact: say that when our acquaintance yells I am looking at you and you are looking at me, [and we exchange a brief glance at our mutual friend and nod]. Thus I know you know about the yell, you know that I know that you know (you see me looking at you), and so on. If we do manage to make eye contact, we get off the bus; communication is successful.

Coordination problems are only ever problems when everyone is *currently* choosing D, and we need to *coordinate* all choosing C at the same time. To do that, we need common knowledge.

(The specific definition of common knowledge ("*I know that you know that I know that....*") is often confusing, but for now the concrete examples below should help get a solid intuition for the idea.)

Compare you and I on the bus to the coordination game payoff matrix: If we *both get off the train* (C, C), we get to hang out with each other *and* spend some time with a mutual acquaintance. If only one of us does, we both miss out on the opportunity to hang out with

each other (the thing we want least - (C, D) or (D, C)). If neither of us gets off the train, we get to hang out with each other, but in a less interesting way (D, D).

A Stable State

The reason that it's a difficult coordination problem, is because the state (D, D) is an equilibrium state; neither of us alone can improve it by getting off the bus - only if we're able to coordinate us *both* getting off the bus does this work. You can think of it like a local optimum: if you take one step in any direction (if any single one of us changes our actions) we lose utility on net.

The name for such an equilibrium is taken from mathematician John Nash (who the film *A Beautiful Mind* was based on), and is called a *Nash equilibrium.* Both (C, C) and (D, D) are Nash equilibria in a coordination problem. Can you see how many Nash equilibria there are in the Prisoner's Dilemma?

Solving problems and resolving dilemmas

A good way to contrast coordination problems and free rider problems is to think about these equilibrium states. In the free rider problem, the situation where everyone cooperates is not a Nash equilibrium - everyone is incentivised to defect while the others cooperate, and so occasionally some people do. While the PD only has one Nash equilibrium however, a coordination problem has got two! The challenge is moving from the current one, to one we all prefer.

Free rider problems are solved by creating new incentives against defecting. For example, the government punishes you if you don't pay your taxes. In sports, the practice of doping is punished, and what's more it's made out to be *dishonourable*. People tell stories of the evil people that dope and how we all look down on them; even if you could dope and probably get away with it, there's no plausible deniability in your mind - you know you're being a bad person and would be judged by everyone of your colleagues.

Coordination problems can be solved by creating such incentives, but they can also be solved just by improving information flow. We'll see that below.

Three Coordination Problems

That situation when you and I lock eyes, nod, and get off the bus? That's *having common knowledge*. It's the confidence to take the step, because you're not worried about what I might do. Because you know I'm getting off the bus with you.

Now we've got a handle on what common knowledge is, we can turn back to the three puzzling phenomena from the beginning.

Dictators and freedom of speech

> Dictatorships all through history have attempted to suppress freedom of the press and freedom of speech. Why is this? Are they just very sensitive? On the other side, the leaders of the Enlightenment fought for freedom of speech, and would not budge an inch against this principle.

Many people under a dictatorship want a revolution - but rebelling only makes sense if enough *other* people want to rebel. The people as a whole are much more powerful than the government. But you alone won't be any match for the local police force. You have to *know* that the others are willing to rebel (as long as you rebel), *and* you have to know that they know that *you're* willing to rebel.

People in a dictatorship are all trying to move to a better nash equilibrium without going via the corners of the box (i.e. where some people rebel, but not enough, and then you have some pointless death instead of a revolution).

That feeling of worrying whether the people around you will support you, if you attack the police. That's what it's like *not* to have common knowledge. When a dictator gets ousted by the people, it's often in the form of a riot, because you can *see the other people around you* who are poised on the brink of violence. They can see you, and you all know that if you moved as one you might accomplish something. That's the feeling of common knowledge.

The dictator is trying to suppress the people's ability to create common knowledge that jumps them straight to (C, C) - and so they attempt to suppress the news media. Preventing common knowledge being formed among the populace means that large factions cannot coordinate - this is a successful divide and conquer strategy, and is why dictators are able to lead with so little support (often <1% of the population).

Uncertainty in Romance

> When two people are on a date and want to sleep with each other, the conversation will often move towards but never *explicitly* discuss having sex. The two may discuss going back to the place of one of theirs, with a different explicit reason discussed (e.g. "to have a drink"), *even* if both want to have sex.

Notice the difference between

- Walking up to someone cold at a bar and starting a conversation
- Walking up to someone at a bar, after you noticed them stealing glances at you
- Walking up to someone at a bar, after you glanced at them, they glanced at you, and your eyes *locked*

It's easiest to approach confidently in the last case, since you have clear evidence that you're both at least interested in a flirtatious conversation.

In dating, getting *explicitly* rejected is a loss of status, so people are incentivised to put a lot of effort into preserving plausible deniability. *No really, I just came up to your flat to listen to your vinyl records!* Similarly, we know other people don't like getting rejected, so we rarely explicitly ask either. *Are you trying to have sex with me?*

So with sex, romance, or even deep friendships, people are often trying to get to (C, C) *without* common knowledge, up until the moment that they're both very confident that both parties are interested in raising their level of intimacy.

(Scott Alexander wrote about this attempt to avoid rejection and the confusion it entails in his post Conversation Deliberately Skirts the Border of Incomprehensibility.)

This problem of *avoiding* common knowledge as we try to move to better Nash equilibrium also shows up in negotiations and war, where you might make a threat, and not *want* there to be common knowledge of whether you'll actually follow through on that threat.

(Added: After listening to a podcast with Robin Hanson, I realise that I've simplified too much here. It's also the case that each member of the couple might not have figured out whether they want to have sex, and so plausible deniability gives them an out if they decide not to, without the explicit status hit/attack.

I definitely have the sense that if someone very bluntly states subtext when they notice it, this means *I can't play the game with them even if I wanted to*, as when they state it explicitly I have to say "No!" else admit that I was slightly flirting / exploring a romance with them, and significantly increase the change I will immediately receive an explicit rejection.)

Communal/Religious Rituals

> Throughout history, communities have had religious rituals that look very similar. Everyone in the village has to join in. There are repetitive songs, repetitive lectures on the same holy books, chanting together. Why, of all the possible community events (e.g. dinner, parties, etc) is this the most common type?

Michael Chwe wrote a whole book on this topic. To simplify massively: rituals are a space to create common knowledge in a community.

You don't just listen to a pastor talk about virtue and sin. You listen *together,* where you know that everyone else was listening too. You say '*amen*' together after each prayer the pastor speaks, and you all know that you're listening along and paying attention. You speak the Lord's Prayer or some Buddhist chant together, and you know that *everyone* knows the words.

Rituals create common knowledge about what in the community is is rewarded, what is punished. This is why religions are so powerful (and why the state likes to control religion). It's not just a part of life like other institutions everyone uses like a market or a bank - this is an institution that builds common knowledge about *all* areas of life, especially the most important communal norms.

To flesh out the punishment part of that: When someone does something sinful by the standards of the community, you know that *they* know they're not supposed to, and they know that you know that they know. This makes it easier to punish people - they can't claim they didn't know they weren't supposed to do something. And making it easier to punish people also makes people less likely to sin in the first place.

The rituals have been gradually improved and changed over time, and often the trade-offs have been towards helping coordinate a community. This is why the words in the chants or songs that everyone sings are simple, repetitive, and often rhyme - so you know that everyone knows exactly what they are. This is why rituals often occur seated in a circle - not only can you see the performance, but you can see *me* seeing the performance, and I you, and we have common knowledge.

Common knowledge is often much easier to build in small groups - in the example about getting off the bus, the two need only to look at each other, share a nod, and common knowledge is achieved. Building common knowledge between hundreds or thousands of people is significantly harder, and the fact that religion has such a significant ability to do so is why it has historically had so much connection to politics.

Common Knowledge Production in Society at Large

Common knowledge is a very common state of affairs that humans had to reason about naturally in the ancestral environment; there is no explicit mathematical calculation being done when two people lock eyes on a bus then coordinate getting off and seeing their friend.

We've looked at how religions help create common knowledge of norms. Here's a few other common knowledge producing mechanisms that exist in the world today.

The News

The main way common knowledge is built is by having everyone in the same room, in silence, while somebody speaks. Another way (in the modern world) is official channels of communication that you know everyone listens to.

This is actually one of the good reasons to discuss news so much - we've built trust that what the NYT says is common knowledge, and so can coordinate around it. Sometimes an official document is advertised widely and is known to be known as common knowledge, even if we ourselves often haven't read it (e.g. Will MacAskill's book, the NYT).

Nowadays there is no such single news source, and we've lost that coordination mechanism. We all have Facebook, but Facebook is entirely built out of bubbles. Facebook *could* choose to create common knowledge by making something appear in everyone's feed, but they choose not to (and this is in fact a fairly restrained use of power that I appreciate).

One time facebook slipped up on this, was when they built their 'Marked Safe' feature. If a dangerous event (big fire, terrorist attack, earthquake) happened near you, you could 'mark yourself safe' and then all of your friends would get a notification saying you were safe.

Now, it was clear that everyone else was seeing the notifications you were seeing, and so if your nearby friend marked themselves safe and you didn't, your friends would all notice that conspicuous absence of a notification, and know that you had chosen not to click it. This creates a pressure for all people to always notify their friends whenever there's been a dangerous event near them, even if the odds of them being involved were miniscule. This is a clear waste of time and attention, ~~and the feature was removed~~ the feature continues to be a piece of security theatre in our lives.

A related point about the power of media that creates common knowledge: in Michael Chwe's book, he does some data analysis of the marketing strategies of multiple different industries. He classifies products that are 'social goods' - those you want to buy if you expect other people like them. For example, you want to buy wines that you know your guests like, or bring beer to parties that others like; you want to use popular computer brands that people have developed software for; etc.

He then shows that social brands typically pay more *per viewer* for advertising; not necessarily more total, but that they'll pay a higher amount for opportunities to broadcast in places that generate common knowledge. Rather than buy 10 opportunities to broadcast to 2 million people on various channels, they'll pay a premium for 20 million people to view their ad during the superbowl, to create stronger common knowledge.

Academic Research

The central place where common knowledge is generated in science is in journals. These are where researchers can discover the new insights of the field, and build off them. Conferences can also help in this regard.

A more interesting case is textbooks (I borrow this example from Oliver Habryka). There was once a time in the history of physics where the basics of quantum mechanics were known,

and yet to study them required reading the right journal articles, in the right order. When you went to a convention of physicists, you likely had to explain many of the basics of the field *before* you could express your new idea.

Then, some people decided to aggregate it into textbooks, which were then all taught to the undergraduates of the next generation, until the point where you could walk into the room and start using all the jargon and *trust that everyone knew what you meant*. Having common knowledge of the basics of a field is necessary for a field to make progress - to make the 201 the 101, and then build new insights on top.

In my life, even if 90% of the people around have the idea, when I'm not confident that 100% do then I often explain the basic idea for everyone. This often costs a lot of time - for example, after you read this post, I'll be able to say to you a sentence like 'the undergrad textbook system is a mechanism to create the common knowledge that allows the field as a whole to jump to the new Nash equilibrium of using advanced concepts'.

Paragraphs can be reduced to sentences, and you can get even more powerful returns with more abstract ideas - in mathematics, pages of symbols can be turned into a couple of lines (with the right abstractions e.g. calculus, linear algebra, probability theory, etc).

Startups

A startup is a very small group of people building detailed models of a product. They're able to create a lot of common knowledge due to their small size. However, one of the reasons why they need to put a lot of thought into the long-term of the company, is because they will *lose* this common knowledge producing mechanism as they scale, and the only things they'll be able to coordinate on are the things they already learned together.

The fact that they're able to build common knowledge when they're small is why they're able to make so much more progress than big companies, and is also why big companies that innovate tend to compartmentalise their teams into small groups. As the company grows, there are far fewer things that can be retained as common knowledge amongst the employees. You can have intensive on-boarding processes for the first 20 hires, but it really doesn't scale to 100 employees.

Here are three things that can sustain at very large scales:

Name: Y Combinator says that the name of your company should tell people what you do - cf. AirBnb, InstaCart, DoorDash, OpenAI, Lyft, etc. Contrast with companies like Palantir, where even I don't know exactly what they work on day-to-day, and I've got friends who work there.

Mission: It is possible to predict the output of an organisation very well by what their mission statement concretely communicates. For example, the company SpaceX has their mission statement at the top of all hiring documents (cf. the application forms to be a rocket scientist, business analyst, or barista).

Values: Affects hiring and decision-making long into the future. YC specifically says to pick 4-8 core values, have a story associated with each value, and tell each story *every* day (e.g. in meetings). That may seem like way too much, but in fact that's how much it can take to make the values common knowledge (especially as your company scales).

At what cost?

A standard response to coordination failures is one of *exasperation* - a feeling that we *should* be able to solve this if only we *tried*.

Imagine you're trying to coordinate you and a few friends to move some furniture, and they keep getting in each other's way. You might shout "Hey guys! Look, Pete and Laurie have to move the couch first, then John and Pauline can move the table!" And then things just start working. Or even just between two of you - when a friend is late for skype calls because she messes up her calendar app, you might express irritation, and she might try extra hard to fix the problem.

We also feel this when we look at society at large, for example when we look at coordination failures in politics. *Why does everyone continue voting for silly-no-good politicians? Why can't we all just vote for someone sane?!*

In the book *Inadequate Equilibria* by Eliezer Yudkowsky, the character *Simplicio* represents this feeling. Here is the character discussing a (real) coordination failure in the US healthcare system that causes a few dozen newborn children to die every year:

> **simplicio:** The first thing you have to understand, Visitor, is that the folk in this world are hypocrites, cowards, psychopaths, and sheep.
>
> I mean, I certainly care about the the lives of newborn children. Hearing about their plight certainly makes me want to do something about it. When I see the problem continuing in spite of that, I can only conclude that other people don't feel the level of moral indignation that I feel when staring at a heap of dead babies.
>
> [...]
>
> Regardless, I'm not seeing what the grand obstacle is to people solving these problems by, you know, coordinating. If people would just act in unity, so much could be done!
>
> I feel like you're placing too much blame on system-level issues, Cecie, when the simpler hypothesis is just that the people in the system are terrible: bad at thinking, bad at caring, bad at coordinating. You claim to be a "cynic," but your whole world-view sounds rose-tinted to me.

One of the final points to deeply understand about common knowledge in society, is how *costly* it is to create at scale.

Big companies get to pick only a few sentences to become common knowledge. To have a community rally around a more complex set of values and ideals (i.e. a significant function of religion) each and every member of that community must give up half of each Sunday, to repeat ideas *they already know* over and over - nothing new, just with the goal of creating common knowledge.

There used to be news programmes everybody in a country would tune in for. Notice how the New York Times used to be something people would read once per week or once per *day*, and discuss it with friends, even though most of the info has no direct effect on their lives.

Our intuitions were developed for tribes of size 150 or less (cf. Dunbar's number) and as such, our intuitions around coordination are often terribly off. Simplicio is someone who has not noticed the cost of creating common knowledge at scale. He believes that society could easily vote for good politicians *if only we coordinated*, and because we don't he infers we *must* be stupid and/or evil.

The feeling of *indignation* at people for failing to coordinate can be thought of as creating an incentive to solve the coordination problem. I'm letting my skype partner know that I will punish them if they fail again. But today, this feeling toward people for failing to coordinate is almost always misguided.

Think of it this way: many small coordination problems are sufficiently small that you'll solve them quickly; many coordination problems are sufficiently big that you have no chance of

solving them via normal means, and you will feel indignation every time you notice them (e.g. think politics/twitter). Basically, when you feel like being indignant in the modern world, 99% of the time it's wasted motion.

Simplicio's intuitions are a great fit for a hunter-gatherer tribe. When he gets indignant, it would be proportional to the problem, the problem would get solved, and everyone would be happy. At a later point in the book Simplicio calls for political revolution - the sort of mechanism that works if you're able to get everyone to gather in a single place.

The solution to coordination problems at scale is much harder, and requires thinking about incentives structures and information flows rather than emotions directed at individuals in your social environment. Or in other words, building a civilization.

> **visitor:** Indeed. Moving from bad equilibria to better equilibria is the whole point of having a civilization in the first place.

- Another character in Inadequate Equilibria, by Eliezer Yudkowsky

So, what's common knowledge *for*?

Summary of this post:

1. A coordination problem is when everyone is taking some action A, and we'd rather all be taking action B, but it's bad if we don't all move to B at the same time.
2. Common knowledge is the name for the epistemic state we're collectively in, when we know we can all start choosing action B - and trust everyone else to do the same.
3. We're intuitively very good at navigating such problems when we're in small groups (size < 150).
4. We're intuitively very bad at navigating such problems in the modern world, and require the building of new, microeconomic intuitions in order to build a successful society.

There is a great deal more subtlety to how common knowledge gets built and propagates. This post has given but a glimpse through the lens of game-theory, and hopefully you now see the light that this lens sheds on a great variety of phenomena.

Links to explore more on this subject:

- *Moloch's Toolbox (Inadequate Equilibria, Ch 3)* (link)
  - *A guide to the ways our current institutions fail to coordinate. Largely applying standard microeconomics, and a great post to read after this one.*
- *Meditations on Moloch* (link)
  - *An original idea about coordination failures, which the above book chapter formalised. It's a great post, and it's good to follow the intellectual heritage of ideas.*
- *Rational Ritual: Culture, Coordination and Common Knowledge* (link)
  - *Solid book with lots of detail.*
- *Scott Aaronson on Common Knowledge and Aumann's Agreement Theorem* (link)
  - *This post caused me to spend a bunch more time thinking about these topics. I found the explanations personally to be fairly abstract, which inspired me to write this post.*
- *Scott Alexander's sequence on Game Theory* (link)
  - *After writing this post, I found Scott Alexander had also written about some of the examples (especially the dictatorship one) in detail 7 years ago (link).*
- *Andrew Critch on 'Unrolling Social Metacognition: Three levels of meta are not enough'* (link)

- *This is a great post going into the details of how my modelling of you modelling me modelling you... works in practice. Highly recommended if the definition of common knowledge presented above seemed confusing.*

Thanks *to Raymond Arnold, Jacob Lagerros and Oliver Habryka for extensive feedback and comments, and to Hadrien Pouget for proofreading an early draft. A further special mention to Raymond for pointing out this term ought to be a standard piece of* expert jargon *in this community, and suggesting I write this post*

An Untrollable Mathematician Illustrated

Crossposted from the AI Alignment Forum. May contain more technical jargon than usual.

The following was a presentation I made for Sören Elverlin's AI Safety Reading Group. I decided to draw everything by hand because powerpoint is boring. Thanks to Ben Pace for formatting it for LW! See also the IAF post detailing the research which this presentation is based on.

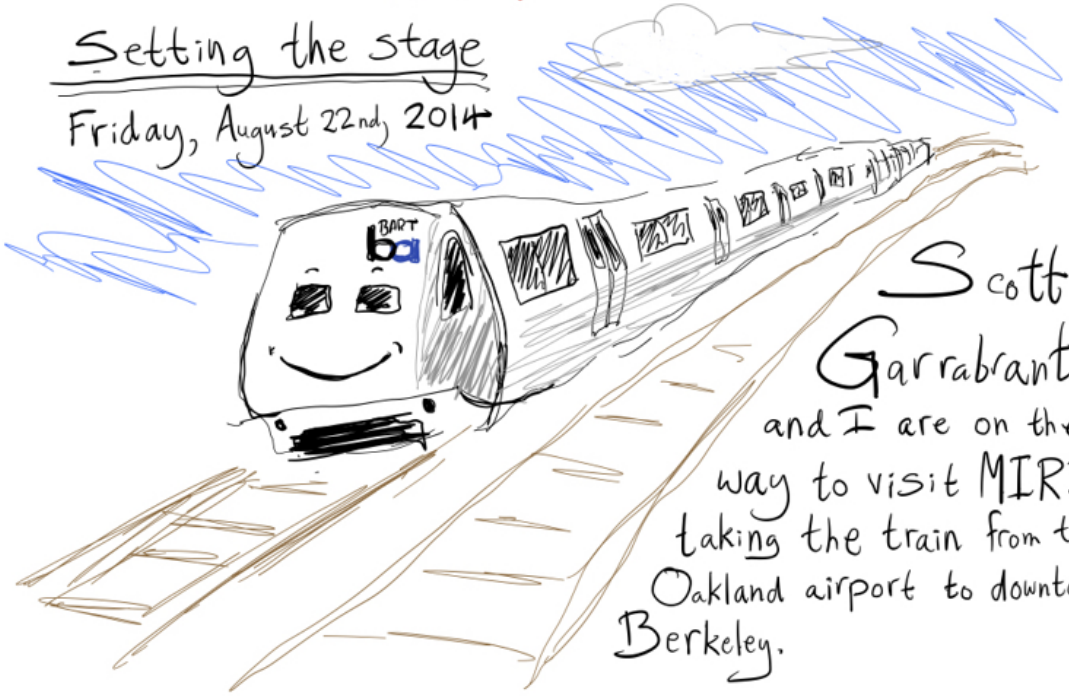
An Untrollable Mathematician

probability distribution by
Sam Eisenstat

written & illustrated by
Abram Demski

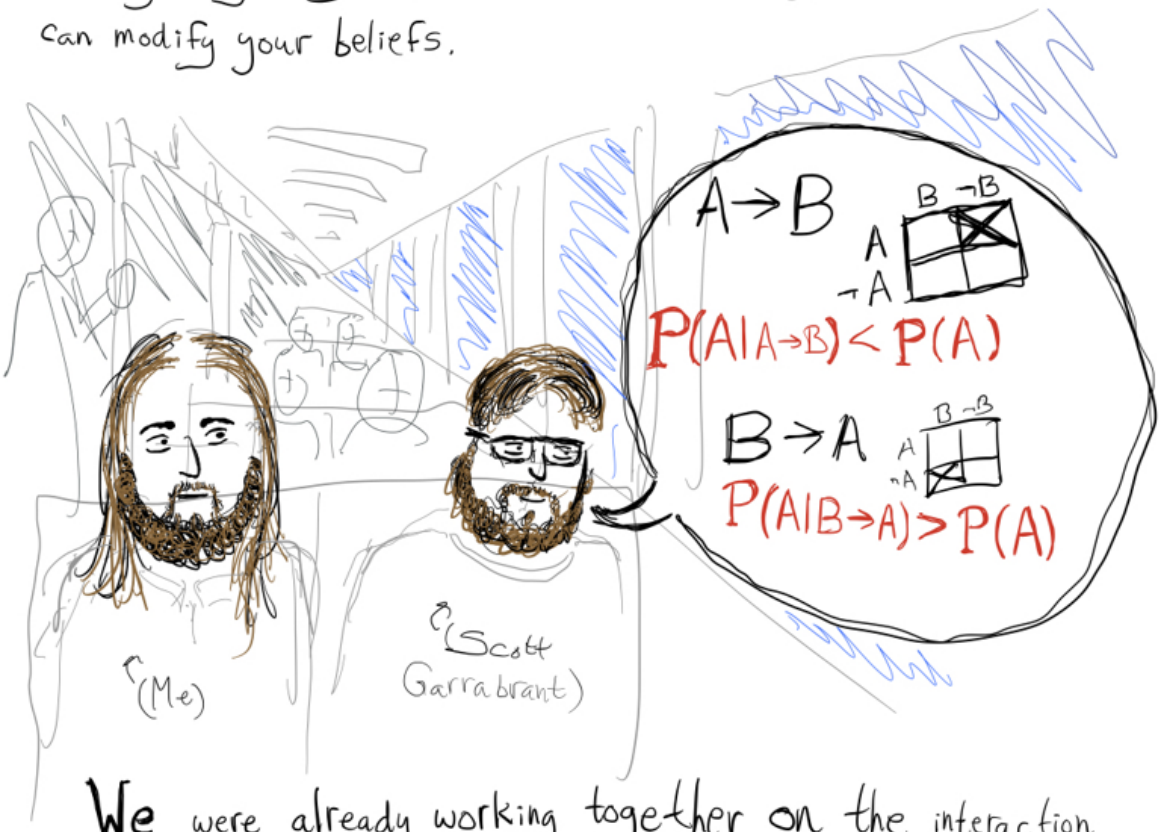
Setting the stage

Friday, August 22nd, 2014

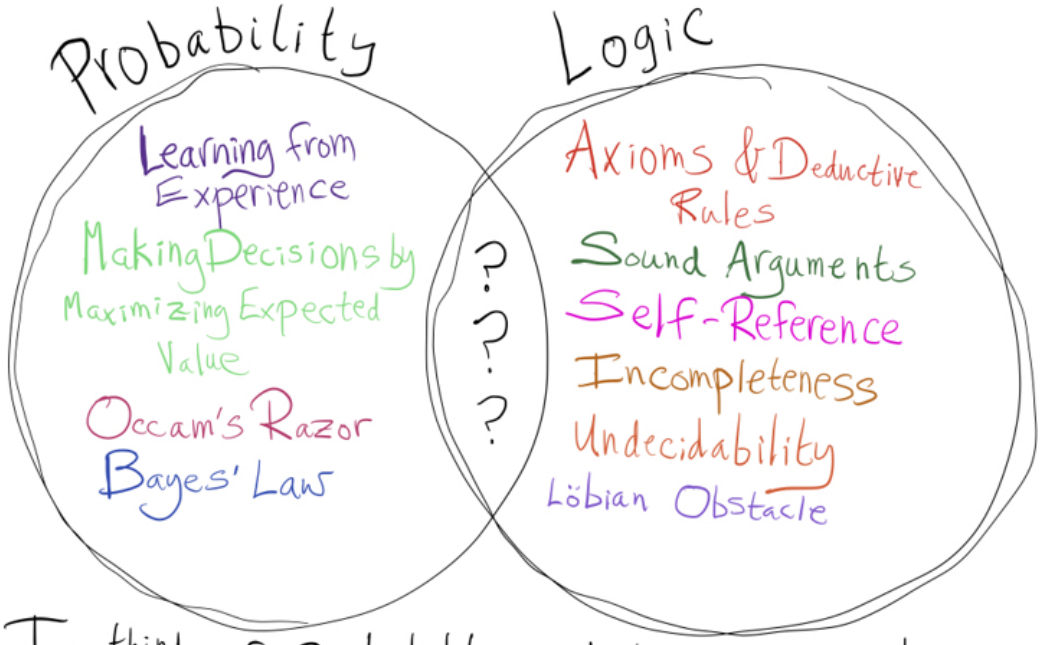


Scott Garrabrant and I are on the way to visit MIRI, taking the train from the Oakland airport to downtown Berkeley.

On the way, Scott makes some observations about the way recognizing implications between logical sentences can modify your beliefs.



We were already working together on the interaction between probabilistic reasoning and logical reasoning. That's what we were visiting MIRI to talk about.



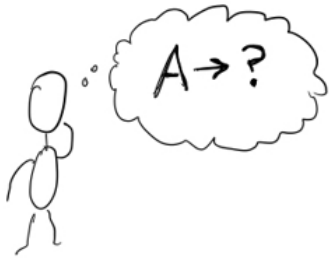
I think of probability and logic as the two great triumphs of codification of rational thought over the past few hundred years.

MIRI needs to combine logical reasoning and probabilistic reasoning to make models of reflective stability: intelligent machines will be capable of modifying themselves, so they need to think clearly about the consequences of doing so despite the many paradoxes of self-reference.

For example, we wouldn't want such a system to want to modify its own goals, except in extreme cases.

Probability theory has the best tools for thinking about decisions, whereas logic has the best tools for thinking about self-reference. So, both are needed.

There are a lot of deep & interesting questions about how probability & logic should work together, but MIRI's primary interest was the question of LOGICAL UNCERTAINTY: how should we reason when we only have partial knowledge of the consequences of our beliefs?



In probability theory, logical omniscience (the assumption that you know the logical consequences of any belief) typically has to be assumed; this is baked into the very laws of probability, the Kolmogorov axioms:

Andrey Kolmogorov

1. $P(E) \in \mathbb{R}, P(E) \geq 0$
2. $P(\Omega) = 1$
3. $P(\bigcup_{i=1}^{\infty} E_i) = \sum_{i=1}^{\infty} P(E_i)$ for $E_1, E_2, \ldots$ where $\forall_{n,m}. E_n \cap E_m = \emptyset$

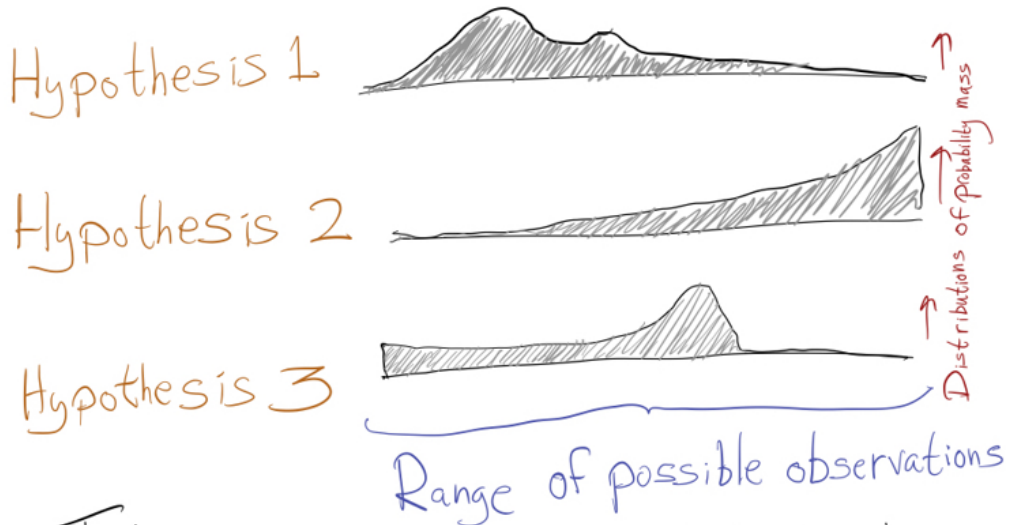
← Everything is stated in terms of events E which form a σ-algebra, which means that events obey the logic of conjunction, disjunction, and negation (including infinite conjunctions and disjunctions, which is to say $\forall$ and $\exists$).

"$P(\Omega)=1$", IE "the probabilities sum to one", may not look overtly like logical omniscience, but Ω really means any event which is logically equivalent to the whole space. That is: IF AN EVENT CAN BE PROVED IT MUST HAVE PROBABILITY ONE. This means that we have to know what is provable ahead of time, before we assign any probabilities.

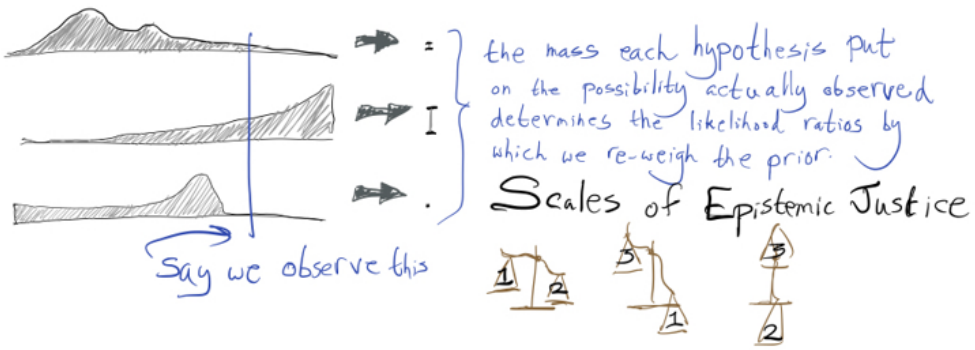
The third axiom, Countable additivity, is surely the worst in this respect; it says that probabilities have to be consistent even when infinite sums are involved. Philosophers usually weaken this to a finite additivity assumption for applications like theories of rational agency (decision theories). Even weakening to finite additivity still requires logical omniscience, though, because you have to know whether $E_1 \cap E_2 = \emptyset$, which is to say, you have to know which events are mutually inconsistant.

But the simplest way to see why logical omniscience is difficult to remove from probability theory, in my opinion, is to consider Bayesian hypothesis testing.

Each hypothesis must clearly declare which probabilities it assigns to which observations:



That way, you know how much to rescale the odds when you make an observation.



If we are unsure about the implications of one of our hypotheses, that means we don't know what probabilities it assigns to observations, which means we don't know how much credit to give it for making predictions.

Hypothesis 4

Confused Scales of Epistemic Discord

We can quantify our meta-uncertainty by looking at the range of possible probabilities, but this is just a way to formalize an unwillingness to assign odds.

Every simply connected, closed 3-manifold is homeomorphic to the 3-sphere.

Poincaré's Conjecture

What we want is something which can deal with the kind of "uncertainty" involved in coming up with mathematical conjectures; assigning possible ranges in cases like that only gets you the full [0,1] interval.

Logic

is as different from this picture as anything could be.

The axioms provide the seed of truth.

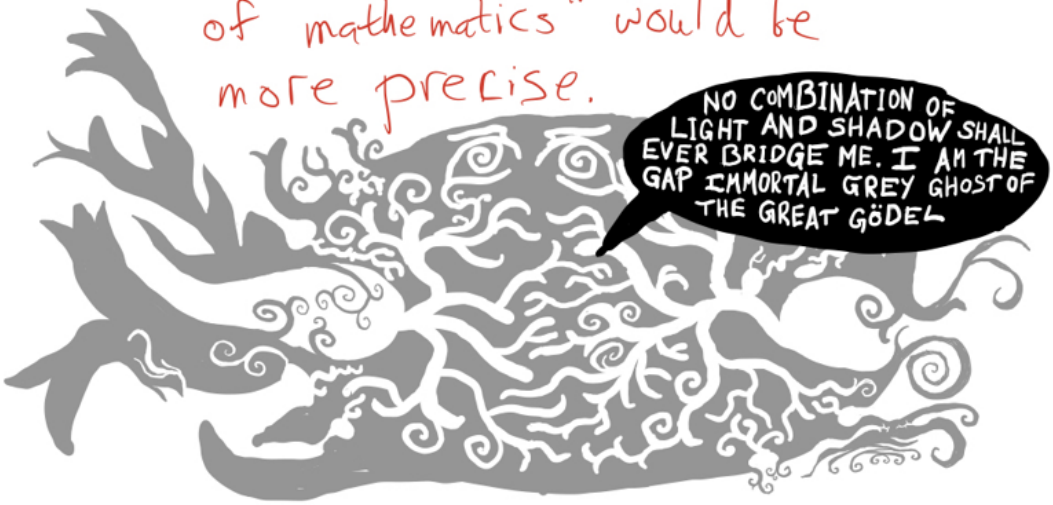
The inference rules specify how the truth grows, branching out into infinity.

The branching growth is never done.

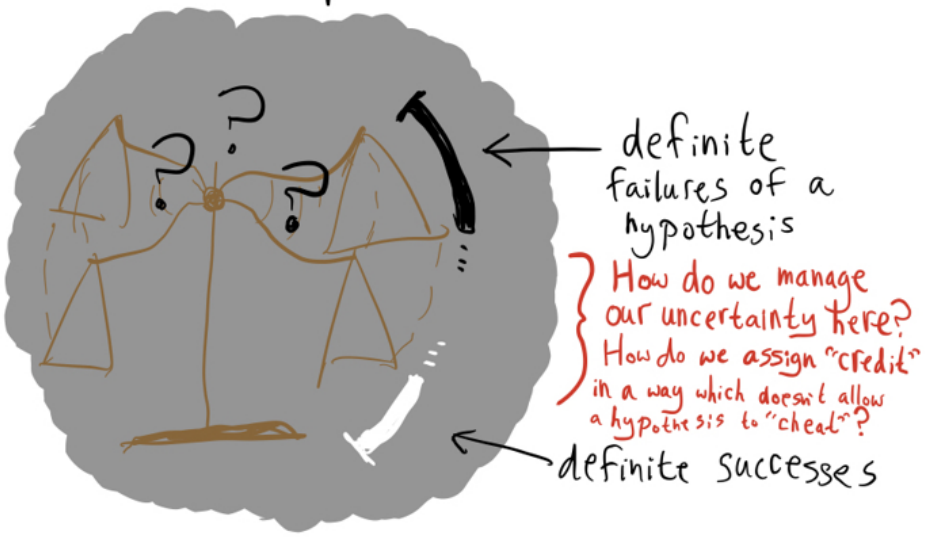
To understand what it means that the growth is "never done", you have to understand that the truth casts a shadow, which is falsehood. To explain Gödel's first incompleteness theorem, Douglas Hofstadter likened truth and falsehood to white and black trees which never touch, but which interlace so finely that you can't possibly draw a border between them.

Gödel's incompleteness theorem has been called a "gaping hole at the heart of mathematics."

"Gaping fractal hole extending self-similar tendrils into any sufficiently complex area of mathematics" would be more precise.



The fact that you can never finish listing the consequences of a belief, and can't even draw any line between necessary truth and contradiction, makes it difficult to balance the scales of probabilistic evidence.



Getting back to Scott and I on that train ride (which Scott says was actually a taxi ride, oops), Scott observed that what goes wrong if you naively try to have a probability distribution which isn't logically omniscient, and then try & update the probability distribution, is that your beliefs can be driven up and down forever--they may not converge.

Suppose you're interested in some sentence A. Some one else can drive the probability of A down by looking for a sentence A→B which can be proved, but which knocks out more than half of A's mass.

Before update:

	B	¬B	
A	.3	.5	→ A's mass: .8
¬A	.1	.1	

→ mostly on A&¬B

Update:

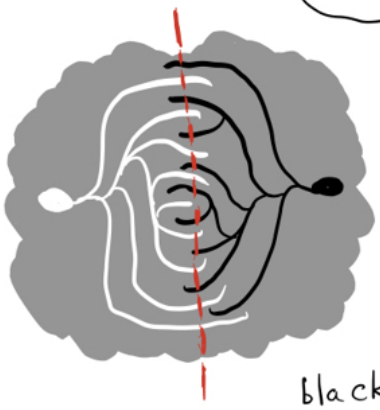
equivalently, ¬(A&¬B)

renormalize → /.5

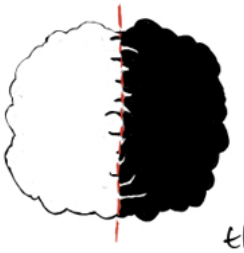
	B	¬B	
A	.6	0	→ A's mass: .6
¬A	.2	.2	

And similarly, if you're trying to drive A up, you pull the same trick with B→A.

Why is it always possible to find a sentence B which allows us to manipulate A like that? Let's go back to Hofstadter's picture of the two trees.

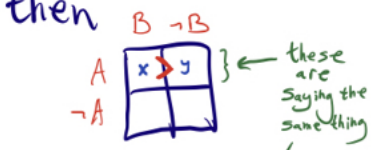


Gödel showed that logical truth and falsehood are so intertwined that there is no (computable) way to separate them: no way to paint everything black and white so that the truth is all painted white, falsehood is all black, and no grey is left between them. You'll always either have some grey left or some truth or falsehood on the wrong side of your division.



So suppose you think you've escaped Scott's trick: you make sure there is no provable $A \to B$ such that A has at least half of its probability mass on $A \& \neg B$.

If $A \to B$ then



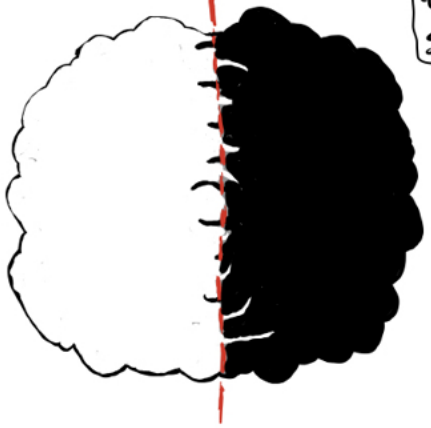
Ahh, but then consider $P(B|A) > \frac{1}{2}$. This must separate logical truth and falsehood! Consider: if B is a logical truth, $A \to B$ must be provable (because anything implies a logical truth), so $P(B|A) > \frac{1}{2}$. But if B were a logical falsehood, that means its negation $\neg B$ would be a logical truth, so $P(\neg B|A) > \frac{1}{2}$, which means $P(B|A) \leq \frac{1}{2}$.

BUT THIS CAN'T BE!

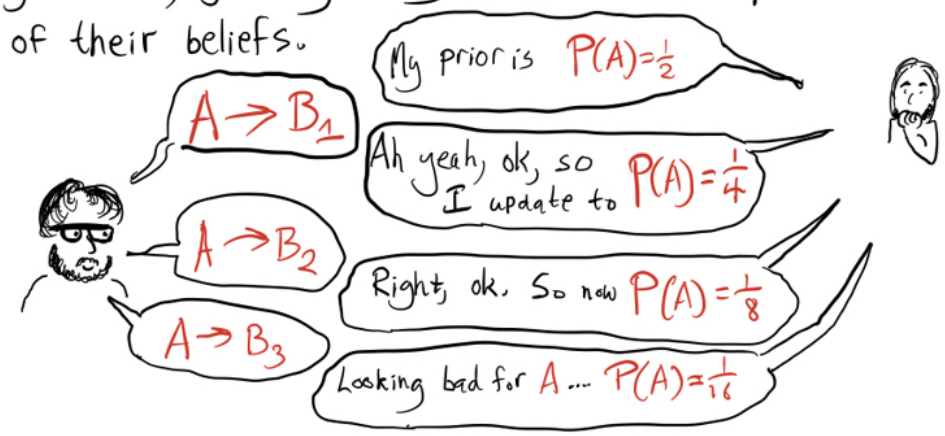
Gödel's incompleteness theorem tells us nothing can separate truth from falsehood.

So there must be some $A \to B$ which is both true and not adequately accounted for by our prior, which can be used to drive down belief in A.

$P(B|A) > \frac{1}{2}$



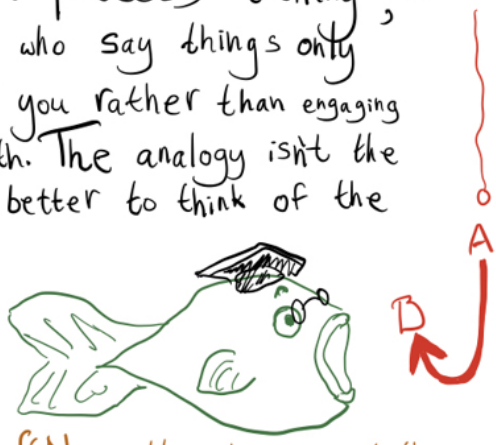
Why is this such a big deal? Because we can repeat the trick many times. The likelihood ratio you get from the trick is 1:2, which means the odds ratio ($P(A):P(\neg A)$) just keeps doubling; you can drive the probability of A as close to 0 as you want, just by telling someone true implications of their beliefs.



And when you're bored of that, you can switch directions and drive A up close to 1. And you can keep doing this forever -- they never catch on, no matter how predictable you are, no matter how many times things have been driven back and forth.

The process would only stop if A were proved or disproved; so, essentially, the probabilistic beliefs are terrible and only get good when the full force of logic steps in to create certainty.

We started calling the process "trolling", in reference to internet trolls who say things only to get a reaction out of you rather than engaging in a conversation in good faith. The analogy isn't the best, though. It might be better to think of the fishing concept, since you drag the victim around as much as you want if you can get them to listen to you.



"No matter how smart the fish, you can always find some bait to hook 'em!"

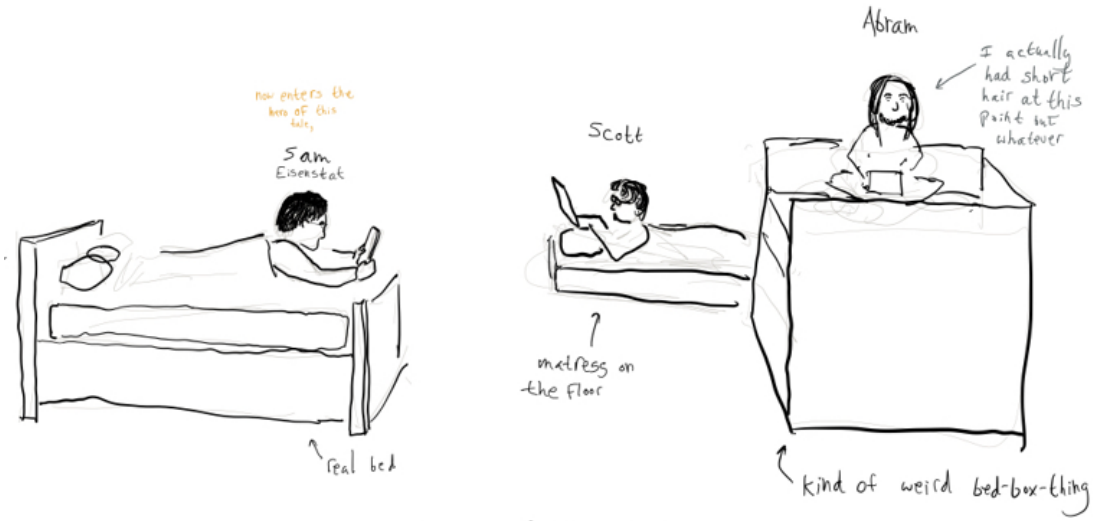
I spent some time looking for ways out of the situation. The post "All Mathematicians are Trollable" describes one approach and why it can't work. It seemed as if the Bayesian should be able to do the right thing by updating on more than just A→B; something like "the troll showed me A→B" should allow the Bayesian to predict the troll, and therefore not be surprised by something like A→B being shown to it, and therefore not continuing to update with each new A→B. But I couldn't figure out how to do this, and soon logical induction came along, which solved this and many other problems in logical uncertainty by abandoning Bayesian updates and baking self-reference into the notion of belief.

The stage has been set.
Time for:

Part Two

in which I finally start to explain the uncontrollable prior

JANUARY 2018
MIRI RESEARCH RETREAT
LATE ONE EVENING

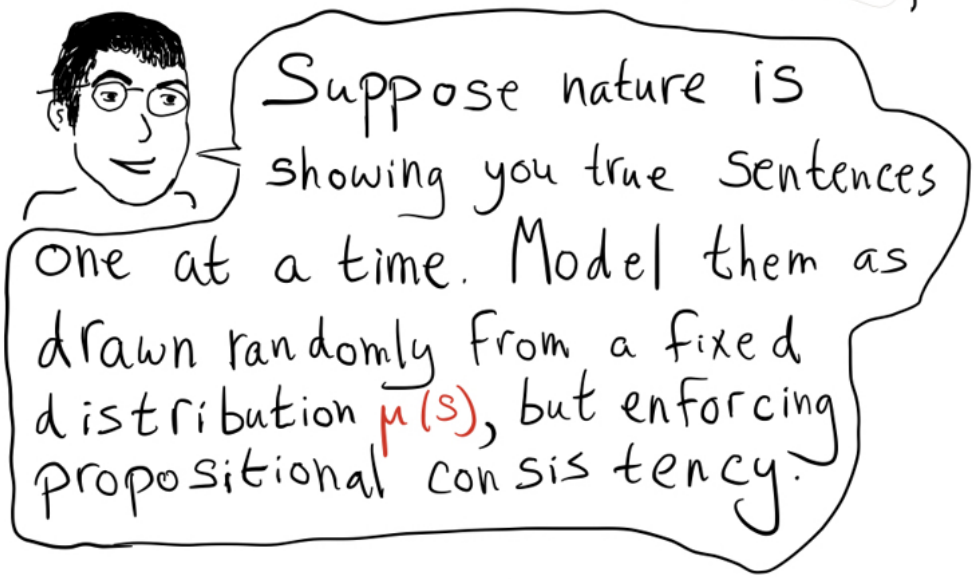


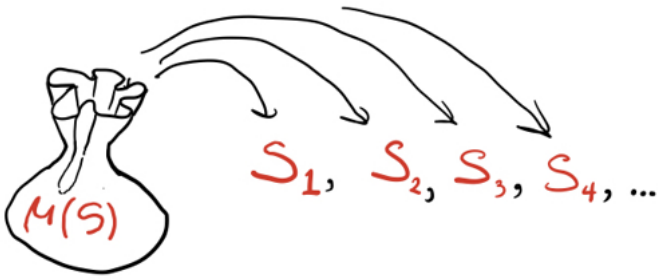
Scott, Sam & I shared one of the bedrooms.

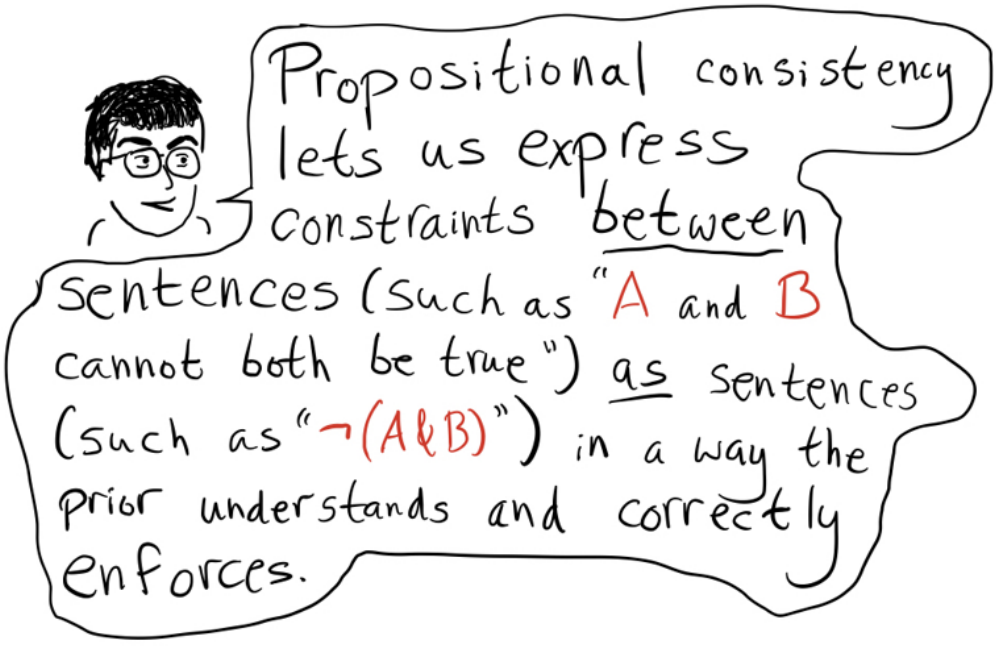
Sam and I had already had several discussions in which he claimed this was possible, but those had not resulted in anything concrete.

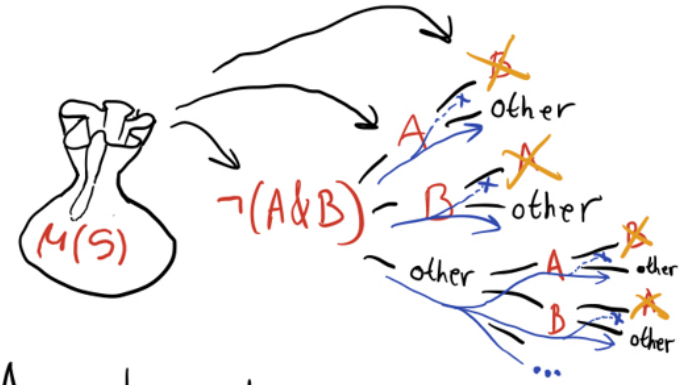
Scott and I were both fairly sure it was just impossible, by this point.

. . . 5 minutes later . . .







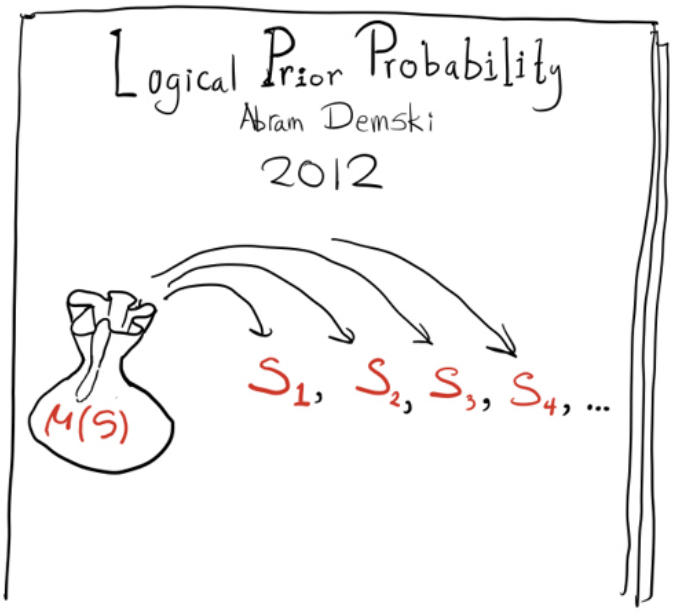


Any branch contradicting an already-stated constraint is clipped off the tree of possible sequences of sentences

The probability of any Sentence S which is consistent with everything seen so far can't go below $\mu(S)$ or above $1-\mu(\neg S)$, since S or $\neg S$ can be drawn next. So, no trolling.

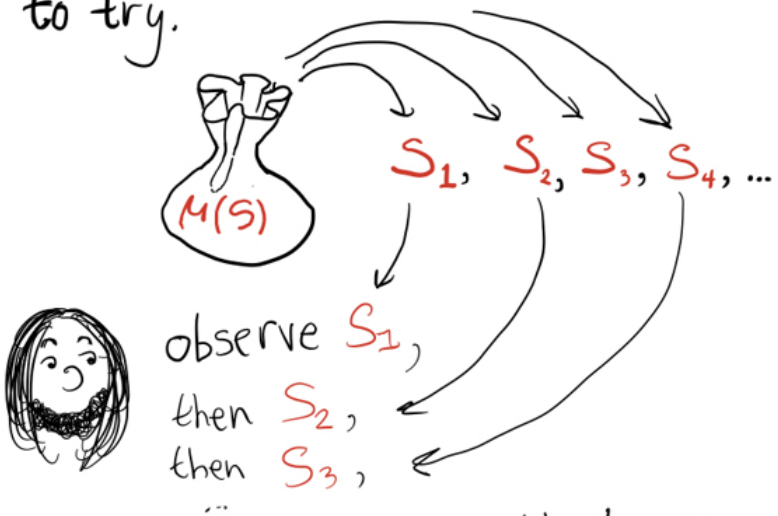
Adding to the embarassment of Scott and I (especially me), this was a very small variation on things I had done. Why hadn't I thought of this?

The idea of defining a probability distribution on logical sentences by repeated draws from a distribution μ plus enforced consistency came from my paper.



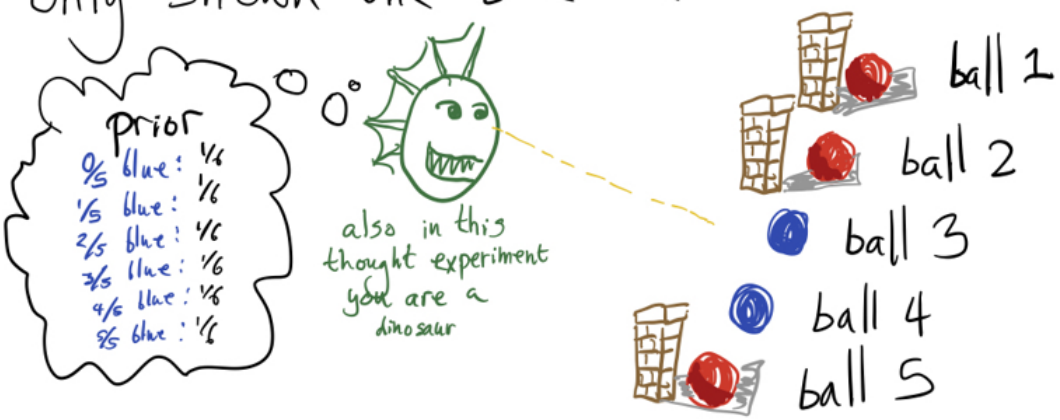
The idea of switching to propositional rather than full consistency was mentioned in my own work on trollability, too.

The big difference was Sam's way of handling observations. Sam defined the order of observation as the order in which sentences get drawn from μ, which I had not thought to try.



So, the theorem stating that all computable priors are trollable isn't false, but needs a poor assumption about the observation model. Sam's untrollable prior is still computable.

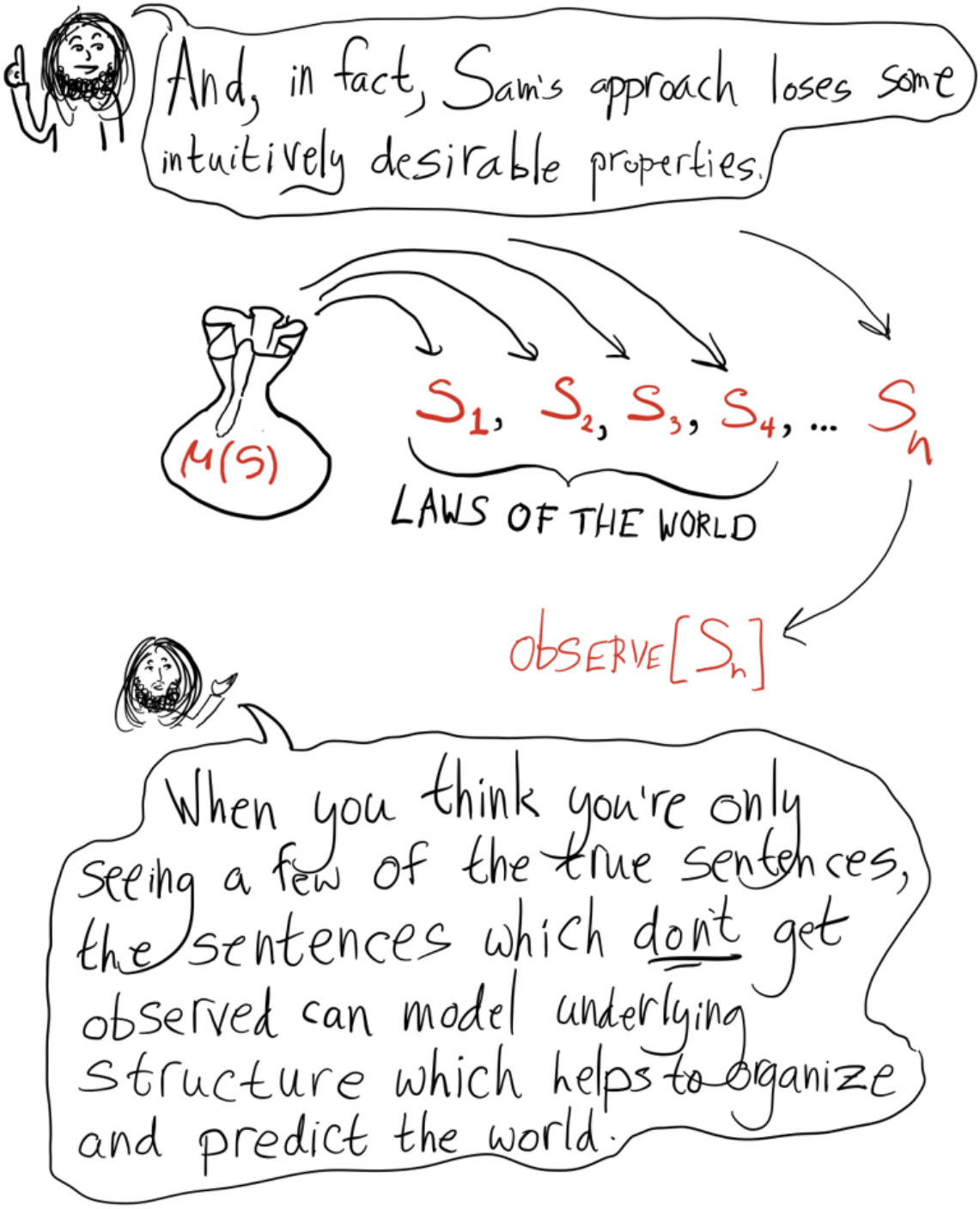
The importance of updating on I observe X rather than X is well-known. For example, suppose you want to know the fraction of red to blue balls in a sequence, and you are only shown the blue ones.

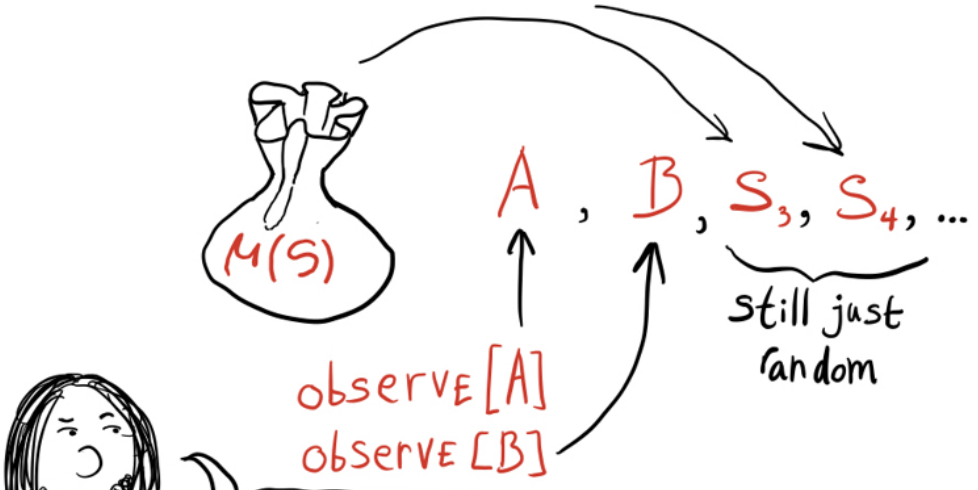


If you just update on "ball 3 and ball 4 are blue", this must make you favor higher fractions of blue than you did before. However, if you know you observe all and only the blue balls, and update on "I see ball 3 and 4 are blue", you update down to 2/5 blue.

Scott's trolling trick involved naively updating on $A \to B$ rather than observe$[A \to B]$, So adding a model of observation effects was a natural thing to try -- and I had tried. However, it was far from obvious how one should model the process of observing logical facts. Wouldn't that mean modeling the process which searches for theorems to show you?

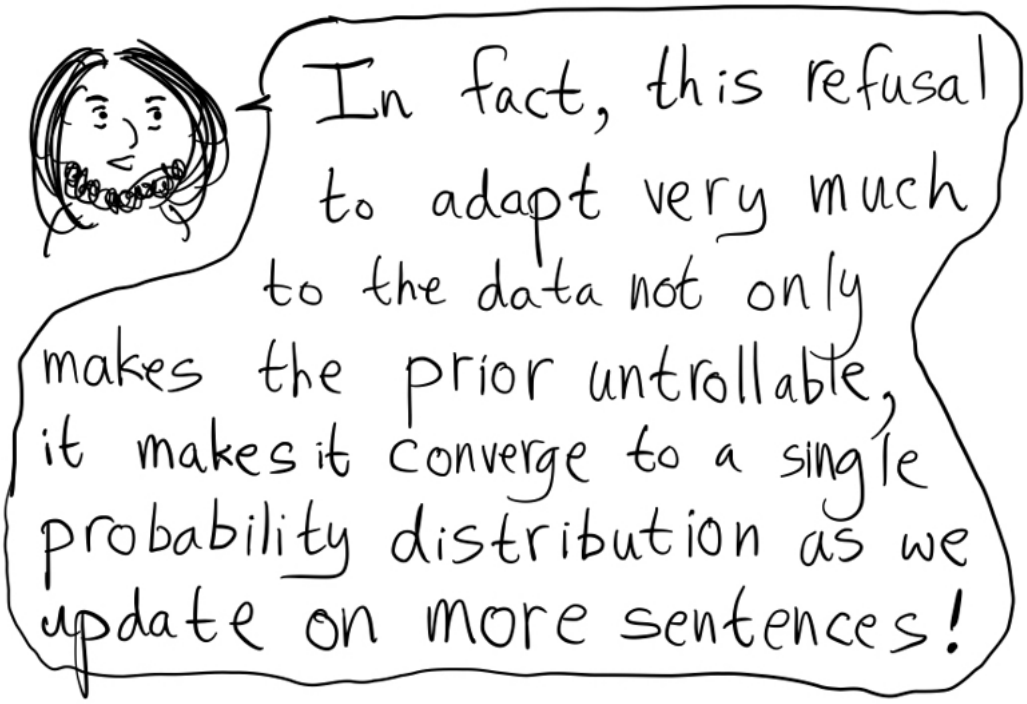
Sam got past this by not trying too hard to model things in the "right" way; his prior is only trying to be controllable, not anything beyond that.


And, in fact, Sam's approach loses some intuitively desirable properties.
M(S)
$S_1, S_2, S_3, S_4, \ldots S_n$
LAWS OF THE WORLD
OBSERVE[S_n]
When you think you're only seeing a few of the true sentences, the sentences which don't get observed can model underlying structure which helps to organize and predict the world.



If there is no difference between "the sequence of sentences you observe" and "the sequence of true sentences", then the universe has no latent structure to learn.

To a large extent, Sam's prior is untrollable <u>because</u> it doesn't adapt its expectations.


In fact, this refusal to adapt very much to the data not only makes the prior untrollable, it makes it converge to a single probability distribution as we update on more sentences!


M(S)
As we see more & more of the things the world will ever show us...
... the things we will never see come to dominate our expectations.

So, although the prior isn't very useful in itself, Sam's result improves our understanding of what's possible. Logical induction (which is untrollable but not exactly a Bayesian probability distribution) is still the gold standard for logical uncertainty, but perhaps the number of desirable properties we can get by specifying simple sampling processes will increase as we push further in the direction Sam has opened up.

The End

Problems with Amplification/Distillation

This post presents my criticisms of Paul Christiano's Amplification and Distillation framework. I'll be basing my understanding of Paul's method mainly on this post. **EDIT**: Recently added this post looking into corrigibility issues relevant to the framework.

To simplify, the framework starts with a human H and a simple agent A[0]. Then there is an amplification step, creating the larger system Amplify(H, A[0]), which consists of H and many copies of A[0] to aid them. Then there is the distillation step, which defines the artificial agent Distil(Amplify(H, A[0]); this is basically an attempt to create a faster, automated version of Amplify(H, A[0]).

Then A[n+1] is defined recursively as Distil(Amplify(H, A[n])); note that there is always a "human in the loop", as H is used every time we replace A[n] with A[n+1].

I won't be talking much about the Distil step, to which I have no real objections (whether it works or not is more an empirical fact than a theoretical one). I'll just note that there is the possibility for (small) noise and error to be introduced by it.

The method relies on three key assumptions. The first one is about Distil; the other two are:

- 2) The Amplify procedure robustly preserves alignment.
- 3) At least some human experts are able to iteratively apply amplification to achieve arbitrarily high capabilities at the relevant task.

I'll also note that, in many informal descriptions of the method, it is assumed that A[n+1] will be taking on more important or more general tasks than A[n]. The idea seems to be that, as long as A[n] does its subtasks safely, Amplify(H, A[n]) can call upon copies of A[n] in confidence, while focusing on larger tasks.

Summary of my critique

I have four main criticisms of the approach:

- 1. "Preserve alignment" is not a valid concept, and "alignment" is badly used in the description of the method.
- 2. The method requires many attendant problems to be solved, just like any other method of alignment.
- 3. There are risks of generating powerful agents within the systems that will try to manipulate it.
- 4. If those attendant problems are solved, it isn't clear there's much remaining of the method.

The first two points will form the core of my critique, with the third as a strong extra worry. I am considerably less convinced about the fourth.

Preserving alignment

How is alignment defined?

An AI aligned with humans would act in our interests. But generally, alignment is conditional: an AI is aligned with humans *in certain circumstances, at certain levels of power*.

For example a spam-filtering agent is aligned with humans as long as it can't influence the sending or receiving of messages. Oracles are aligned as long as they remain contained, and the box moving agent is aligned as long it can't look more than 16 moves into the future.

The only unconditionally aligned agent is the hypothetical friendly AI, which would indeed be aligned in (almost) all circumstances, at all power levels. But that is the only one; even humans are not unconditionally aligned with themselves, as there are many circumstances where humans can be made to act against their own interests.

Therefore, the question is not whether alignment is preserved passing from A[n] to Amplify(H, A[n]) to A[n+1] preserves some hypothetical alignment, but whether A[n+1] will be aligned on the new tasks it will be used on, and using the new powers it may have.

Let T[n] be the tasks that the A[n] will attempt, and (as a proxy for "power"), let P[n] be the set policies that it can choose among. Then if we define "preserve alignment" as

- If the A[n] is aligned on T[n] using P[n], then A[n+1] is also aligned on T[n] using P[n],

then I'd agree that we can likely define some process that preserves alignment. But what the method seems to want, is for A[n+1] to also be aligned on T[n+1] (it does more) using P[n+1] (it has more options). That is a completely different question, and one that seems much more complicated than any "preserve alignment" checks.

Ambiguously-aligned tasks

Should an AI kidnap a human? It should, if they are an escaped prisoner. Should they kill a human? There are many situations, in war or imminent terrorism, where they should. Should they cut a human or torture them? Surgeons and S&M practitioners have views on that. Manipulate a human's emotions? Many movies do nothing but that. Conversely, should the AI tell the truth? We have debates over privacy and cryptography precisely because there are cases where this is not for the best.

Or, to take a more ambiguous example, should an AI kill a pigeon? Well, is this the last remaining male passenger pigeon, or is it a common pigeon in an already over-pigeoned city with many famous statues?

What this means is that an agent cannot know if a task is positive, unless they know the full context of the task, the tradeoffs, and the risks. Even if the agent assumes the task itself is benign/positive - "drive this virtuous person to an important meeting" -

they still need to know the full context to assess side-effects - "is it ok if I run over this pigeon on the way?".

So in order for the subagent to know that they are always "aligned", they need to either know the full context of their role, or have a full task specification fed to it, with any and all tradeoffs it might encounter. Sending the full context to all subagents breaks the whole concept of the approach (which was that the subagent was supposed to be secure and safe in their role, whatever it was, so only the top agent needed to focus on the overall picture), and the second is FAI-complete for anything complicated.

Solving other problems

The king and his all-seeing ministers

For some large N, let's assume that we have a A[N+1], working on some problem of immense impact - maybe building a FAI, solving death, or implementing some long-term policies. Since the Distil step is not the tricky one here, let's see this as a human H working with many copies of A[N].

H will not be able to reach good decisions on their own, of course - their mental powers are far too low. Instead they will need to delegate tasks to A[N] agents, including crucial tasks of summarising information, and predicting the outcomes of certain policies.

The summarisers and predictors have arbitrarily high influence over H. Depending on how they phrase their summaries, they can push H to make any decision whatsoever. If the A[N] are willing to lie in a coordinated fashion, then there really is no recourse: H's actions will be entirely in their control. So one major unsolved problems is how to prevent A[N]'s coordinating to manipulate H.

But there is a much more severe problem: "provide H with accurate informative summaries" is not a clear and fundamental category. Truth-telling is not enough - it is easy to manipulate humans with strict truths - and the problem of informative outputs is unresolved.

Fundamentally, a major issues with human preferences is that they are underdefined, contradictory, and manipulable. If the problem of manipulation is not resolved, then H is irrelevant in this setup; while trying to do their job, they will end up endorsing the choices that were made for them by the A[N]. Notice that this need not require collusion between agents at all; a common bias or preference in the A[N] and the ability to manipulate H is all that's needed. Indeed, because of the other problems with human preferences, it may be that it is impossible for A[N] to give answers that are not manipulable in some sense: every way of summarising or predicting will change H's preferences.

Contradictory preferences are problematic because H will end up resolving a lot of those contradictions through their own choices. Meaning that they will not do this in a principled way, but through a random process (if it is not manipulated). Whole swaths of human value may fall by the wayside, either because H chose not to endorse them (consider a religious fanatic that choose not to endorse doubt, or an extreme utilitarian that completely negated either pleasure or preference - whichever one

you'd least prefer to lose), or because they didn't recall explicitly, at this key moment, that they had other values.

The problems of underdefined values are similar; here the process will be filling in the holes in these values at random. If H endorses friendship and love, but doesn't explicitly define them in sufficient detail, we may end up with a world shorn of such things, perhaps replaced by some empty ritual (this is not too dissimilar to the reasons FAI is hard in the first place). There is a similar problem with ontological crises, because they make even fully defined human values underdefined - underdefined in the new ontology.

In these circumstances, it might be better to have A[N+1] focus more on modest goals, maybe building up some key infrastructure for later use. But this requires a) a suitable H, b) admitting the Amplify/Distil framework can only solve some narrow problems, and c) a solution to the problem of low impact and reduced side effects. The same suspect A[N] that are predicting and summarising for H, will be the ones that will be used to check whether a policy has negative side effect.

In summary, for this approach to work, major solutions need to be provided to most of the following fundamental problems:

- How to prevent coordination between superintelligent agents.
- How to prevent deliberate manipulation of H.
- How to define informative information sharing with H.
- How to resolve H's proto-preferences into a single reward.
- How to keep rewards stable over ontology changes.
- How to define the problem of low impact.

How will these problems be solved, if we don't deliberately set out to solve them? I don't see them getting solved, as side effect of a framework that doesn't deal with them specifically.

Will a ghost grow in the machine?

The previous section briefly mentioned collusion between different superintelligent A[N]. This seemed to require that they be agents with a common goal, which is initially unlikely. But such behaviour can arise given very little pressure.

A[n]-side ghosts

Suppose that H uses multiple A[n] to answer some question they have - maybe to generate a plan to achieve some interim goal. Suppose each A[n] can be fed a different input seed - maybe they have different information, maybe they are instructed to focus on different parts of the solution space, or maybe they just have a different random seed for stochastic decisions. Then as soon as H sees a plan they like, they stop the running of the other agents.

Then when this process is Distilled, it will generate A[n+1] that correspond to programs that can most rapidly generate H-acceptable answers. If this H behaviour continues for just a few more levels, we should have A[n+4] that have become extremely adept at manipulating H. At this point, H lacks the tools to fix the problem,

as the tools themselves are corrupted. That’s if H even realises that the tools are corrupted.

All requirements have the potential to create the same problems. If H is deliberately looking for corrupted A[n], then they are encouraging the creation of agents which *look* completely un-corrupted. This may include actually un-corrupted A[n], but it may be easier for a corrupted one to appear "more saintly than the saints" than for an honest one.

Even the criteria for "preserving alignment" can generate this problem. Call these criteria C. Unless the C are perfectly defined for their task - which I’ve argued is impossible, because their task itself is not well-defined - they will generate a selective pressure as n increases, transforming A[n] into a C-maximiser.

H-side ghosts

The previous subsection focused on clever A[n] deviously manipulating a hapless H. But H itself can be responsible for goal directed behaviour as well.

I can imagine myself as the H of a certain group of A[n]. These A[n] are capable, but not too capable, and I’ve satisfied myself that I can resolve my own contradictory values adequately, at least at this level.

But I don’t trust the Hs with the power of the higher programs A[N]. These Hs may be me, or may be someone else; in any case, because of the problems mentioned above, I don’t trust their decisions. Or, in a more limited way, I think they won’t appreciate all the nuances of the tasks at my level (see the section on ambiguously aligned tasks).

So I decide to manipulate them in the "right" direction. In this case, I unleash my A[n] to deliberately deceive, manipulate, and trick. I want them to give me the right turn of phrase to convince the higher H’s to do or not do something; I want them to figure out which tasks I should delay and which I should rush; I want their help to reconstruct the whole context from the details of the task I have been given.

I can test the effectiveness of some of these methods by running them on myself, or by setting up some procedure where I run A[n-1] subagents inside a A[n], and see what they can do to manipulate the distilled H inside the A[n].

And then, finally, when all this is distilled, there is an exceedingly goal-orientated and manipulating A[n+1].

Note that I don’t need to be evil or selfish to do this; depending on my beliefs about this whole method, this might be the moral and upstanding thing to do. It might be worth the risk to create, say, a manipulative low impact agent according to my current low impact definitions, rather than allowing ever more drift.

But actual evil H’s, or H’s with certain strong feelings in certain situations, may also follow this avenue. Someone with strong views of the world, whose views are threatened by what the A[n] are reporting, and who feel that they would have given up their faith/political tribe/self-image if they had the A[n+1] instead... might be motivated to ensure this doesn’t happen. Conversely, the same might happen for an H whose views have shifted a lot during the Amplification process, but who suspects that this shift is unlikely to be reproduced at the higher level, with a new H.

Rails but no destination

These are just a few ideas on how goal-directed agents could emerge in the framework; there are certainly other avenues I haven't considered.

All of them stem from the same fundamental problem: we don't have a FAI, nor a clear destination for what A[N] should be doing. Therefore, the Amplification/Distillation steps are taking a random walk in the space of minds, with each step defined by imperfect local criteria. There is no reason to suspect the ultimate attractor of this method will be good.

Is the method good if the problems are solved?

We now come to the weakest and least certain point of this critique, the argument that, if the above problems are solved, then the method becomes redundant and unnecessary.

In a certain sense, this is true. If all the problems with identifying human reward were solved, then we could use this reward to program a full friendly AI, meaning we wouldn't need the whole Amplify-Distil framework.

But would some partial solutions allow that framework to be used? If, for instance, I was to spend all my efforts to try and make the framework work, trying to solve just enough of the problems, could I do it?

(I'm framing this as a challenge to myself, as it's always easier to accept "it can't be done", rather than "I would fail if I tried"; this aligns my motivation more towards solution than problems)

I'm not so sure I could fix it, and I'm not so sure it could be fixed. The "Ghost in the Machine" examples were two that came to me after thinking about the problem for a short while. Those two could be easily patched. But could I come up with more examples if I thought for longer? Probably. Could those be patched? Very likely. Would we be sure we'd caught all the problems?

Ah, that's the challenge. Patching methods until you can no longer find holes is a disastrous way of doing things. We need a principled approach that says that holes are (very very) unlikely to happen, even if we can't think of them.

But, it might be possible. I'm plausibly optimistic that Paul, Eric, or someone else will find a principled way of overcoming the "internal agent" problem.

It's the "extracting human preferences correctly" style of problems that I'm far more worried about. I could imagine that, say, Amplify(H, A[10]) would be better than me or H at solving this problem. Indeed, if it succeeds, then we could use that to program a FAI (given a few solutions to other problems), and that would, in a sense, count Paul's approach succeeding.

But that is not how the approach is generally presented. It's seen as a way of getting an A[N] that is aligned, without needing to solve the hard problems of FAI. It's not

generally sold as a method to moderately amplify human abilities, ensuring these amplification are safe through non-generalisable means (eg, not beyond A[10]), and then use these amplified abilities to solve FAI. I would be writing a very different post if that was how I thought people saw that approach.

So, unless the method is specifically earmarked to solve FAI, I don't see how those hard problems would get solved as an incidental side-effect.

A partial vindication scenario

There is one scenario in which I could imagine the framework working. In my Oracles approach, I worked on the problem by seeing what could be safely done, on one side, and what could usefully be done, on the other, until they met in the middle at some acceptable point.

Now, suppose some of the problems of extracting human value were actually solved. Then it's plausible that this would open up a space of solution to the whole problem, a space of methods that were not applicable before.

I don't think it's likely that the Amplify-Distil scenario would exactly fit in that space of applicable methods. But it's possible that some variant of it might. So, by working on solutions to the FAI-style problems on one side, and the Amplify-Distil scenario on the other (I'd expect it would change quite a bit), it's conceivable they could meet in the middle, creating a workable safe framework overall.

Making yourself small

Disclaimers:

- *Epistemic status: trying to share a simplified model of a thing to make it easier to talk about; confident there's something there, but not confident that my read of it or this attempt at simplification is good.*
- *This post is a rewrite of a talk I gave at a CFAR event that seemed well-received; a couple of people who weren't there heard about it and asked if I'd explain the thing. I tried to write this relatively quickly and keep it relatively short, which may mean it's less clear than ideal - happy to hash things out in the comments if so.*
- *The thing is much easier to describe if I occasionally use some woo-y language like "aura" and "energy" to gesture in the direction of what I mean. I'll put these words in quotes so you know I know I'm being woo-y; feel free to mentally insert "something-you-might-call-" before each instance, if that helps.*

Rationalists love talking about status. And that's great - it's a useful idea, for sure.

But I think in our eagerness to notice and apply the concept of status, we end up conflating it with a related-but-different thing. I also think the different thing is super useful in its own right, and important to understand, and I hope sharing my relatively basic thoughts on it will also let others build off that beginning.

So this post is my attempt to explain that different thing. I'm going to call it "making yourself big" and "making yourself small".

This post:

1. *Horses, goats, bulls*
2. *A framework*
3. *What to do with it*

---

1. Horses, goats, bulls

Let's start with an animal showing how it's done. **This video** popped up in my feed recently, and is an amazing example of an animal (the goat) "making himself big". To us, it's obvious that the bull could flatten the goat in any real contest. But the bull doesn't know that! He's not reasoning about relative mass or propulsive power; he's responding purely to how "big" the goat is making his "aura".

Watch the video and see if you can notice specifically how the goat is doing this.

0:07 - where the goat rears up - is an obvious moment, but I claim 0:30-0:36 is even better, where the bull feints forward a couple of times while the goat stands firm, using his posture to "project" his "energy" irrepressibly forward. The bull is simply unable to continue toward him.

The next two examples are of a human working with horses, to show examples of it looks like for a human to be big or small.

(Full disclosure: this post is actually describing a concept I ported over directly from horsemanship. Due to the amount of time I spent as a kid thinking about horse training, my

brain is basically just a bunch of horse metaphors stacked on top of each other.)

First, getting big. The clip I'll link to shows a stallion who was abnormally poorly trained, plus probably suffered some brain damage as a foal, who as a result is abnormally aggressive. The trainer, therefore, needs to make himself much bigger than is normally necessary to keep the stallion away from him.

Watch **0:30-0:50 of this video** (warning: the video is slightly graphic if you watch all the way to the end).

See how the trainer uses his flag and motion/"energy" towards the stallion to make himself bigger, which pushes the horse away from him - without using any physical contact? Notice that the trainer does *not* hit the horse. (The motions he's making may look like threats to strike, and it's true that "making yourself big" ultimately rests on implied threat, but it's the same flavor of threat that the goat is making in the video above - made much more of bluster than of capacity to harm.) I'm pretty confident this horse has never been struck by a human in his life, and certainly not by this trainer. He's not recalling previous pain caused by this human and moving back to avoid it; he's just instinctively making space for how "big" the trainer has made himself.

I found it harder to find a good clip of getting small, but I think this one of the same trainer working with a troubled mare is pretty good - watch **1:16-1:45 of this video**.

Can you see the moments where he is "smallest"? The first is at 1:33, where he's physically walking away - he's actually making himself so small that a "vacuum" is created in his wake, and the mare walks towards him to fill it.

A more classic example is 1:41-1:45. Notice that his body faces away from the mare; he does not make eye contact with her; he moves slowly. In response, she's able to be close to him, because his "aura" (unlike at 0:18-0:26, 1:16-1:21, or 1:39) is not pushing her away.

Hopefully you now feel like you have some intuition for what making yourself "big" or "small" could mean at all. The above examples show "bigness" and "smallness" causing other animals to physically move their bodies; I claim that this type of body language is a significant part of how social mammals like cows, goats, and horses communicate with each other.

So how does this apply to human-human interactions?

You guessed it: it turns out that humans are social mammals too! We just have more complicated ways of moving our bodies around. (e.g., wiggling our mouth-parts in ways intended to produce specific vibrations in the ears of people nearby.)

2. A framework

Before giving some human-to-human examples, here's a simplified framework to distinguish high/low status from making yourself big/small.

High/low status is about (among other things):

- How much **power** you **have**
- How much **attention** you **can expect**
- How much **space** you are **entitled to**

Making yourself big/small is about (among other things):

- How much **power** you are **exercising**

- How much **attention** you **are demanding**
- How much **space** you are **taking up**

It's pretty easy to think of examples of people who are high status and tend to make themselves big, or low status and tend to make themselves small. Here's an example of each:

(That's Elon Musk and Neville Longbottom, if you don't recognize them.)

But what goes in the other corners? **I encourage you to try to generate an example of each before scrolling down.** (I'd love to hear in the comments what you came up with, and if you still think they're good examples after reading the rest of the post.)

.

.

.

doop

doop

doop

.

.

.

making

space

so

you

can

think

about

it

yourself

.

.

.

okay,

.

.

.

here

.

.

.

you

.

.

.

go...

.

.

.

My low/big example is the very same Neville Longbottom of low/small fame. But this time, it's Neville in a very specific scene - the one at the end of the first Harry Potter book, where he tries to prevent his friends from leaving the common room. As you may remember, this doesn't end particularly well for him; indeed, making yourself bigger than the "size" that corresponds to how much status you have is often not a very successful move.

(Which makes sense in the framework above. What would you expect to happen to someone who is trying to exercise more power than they have, demand more attention than they can expect, or take up more space than they're entitled to? I do think low/big can sometimes be effective, but it's tough to pull off.)

The high/small example is even more interesting. The image is of Anna Salamon, one of the cofounders of CFAR. I don't want to refer to *teacher-Anna*, who stands in front of a room and commands attention, but to *mentor-Anna*.

Mentor-Anna (whom you probably meet in a setting where she is fairly high-status, as a teacher or organizer or generally a person-whom-others-seem-to-respect) sits in a circle with you, or across from you one-on-one, and makes herself small. She doesn't play low-status - she doesn't act scared or powerless or shy. Instead, she talks slowly, leaves plenty of silence for you to fill, physically takes up a small amount of space (with knees to her chest, or legs crossed and hands in front of her, or similar), often looks away from you, and doesn't interrupt. In response, the people she's talking with tend to be drawn out of themselves; they have space to reflect; they share half-baked plans and half-acknowledged insecurities. They "expand" to fill the space she has created.

3. What to do with it

As with concepts like status or SNS/PSNS activation, I think what's useful about having this concept in your mental toolbox is that you can practice:

1) noticing it at play in yourself and others

2) moving where you are on the spectrum

For the latter, probably the most important thing is your mental/emotional state - a friend suggested "not wanting to startle a small bird" as an mindset to inhabit, to encourage yourself to become "smaller".

If you want more concrete/physical suggestions, here are a few:

- Interrupt less
- Be silent more (both by pausing while you're speaking, and by waiting a little longer before speaking after someone else)
- Use less eye contact (both with the person you're speaking with, and with others nearby - e.g. avoid the classic move of looking at everyone around the circle to see if they found your joke funny)
- Take up less physical space (curl your body in rather than puffing it out, even if only slightly; lean away rather than toward)
- Make hedged suggestions or share tentative ideas, rather than using more command-like language (example "smaller" language: "what do you think of the idea of…"; "how would it be if we…"; "maybe one option would be…")

(By contrast, tips on how to play low status would be more like "make yourself seem defenseless/weak/submissive".)

I've focused on making yourself small in this post, since I think it's undervalued relative to making yourself big. But since every piece of advice can be reversed, maybe also consider whether you should be making yourself big more often, and how you would do that? (Suggested mindset: be a matador, owning the arena despite the bull charging at you.)

This is the part of the post where videos of human-to-human examples would be really helpful. Unfortunately I found it tricky to come up with examples I could search for that aren't just high/big or low/small (e.g. the classic "new kid takes down the bully" scene in lots of high school movies usually involves the new kid "getting big", but also doing a bunch of high-status behavior, which doesn't seem very helpful for explaining). Given that this post has been sitting in "drafts" for a couple of weeks now, waiting for me to get around to finding better examples, I decided to go ahead and post it without more videos.

But to point a little more towards why you should care at all, here are some brief descriptions of example situations where I claim this concept is relevant and useful (please mentally insert additional "I claim"s in any unintuitive-seeming places):

- Per the Anna example above, making yourself small is a really good way to non-explicitly encourage someone who seems shy/intimidated/reticent to feel more comfortable coming out of their shell.
- At the top, I said people often conflate high/low status with making yourself big/small. As an example, when meeting new people (e.g. at a party or networking event), you might go from thinking "higher status is better" to "I should make myself as big as possible". But this often backfires, either because you become intimidating (see previous bullet) or because you infringe upon the "space"/"aura" of others, causing them to feel hostile/defensive/aggressive. Decoupling making yourself smaller from playing low status can help you make a much better impression - neither "loud and brash" nor "scared and shy"; more like "self-contained and confident".
- In any context where you want the people around you to pay attention to someone else (e.g. if you're making a joint presentation of some kind, and it's their turn to hold the

floor), making yourself small will make it easier for that person to take up the space and hold the attention of others.
- As with lots of interpersonal concepts, this can also be useful internally: if you're familiar with internal double crux / internal family systems / other "parts-work", play around with the motion of having parts of yourself make themselves small (or big).
- More generally, directing your "energy"/"aura" in other ways (beyond just "bigger"/"smaller") - and noticing others doing it - can be useful in tons of situations. As a trivial example, try it the next time you're in that situation where you bump into someone walking in the other direction, and the two of you can't figure out which side to pass each other on.

I hope those brief descriptions make sense just based on this post; if not, I can expand on in them the comments if there's interest. I'd also be curious for what examples you can come up with (or notice in your daily life after reading the post).

If you know me personally, I'm also happy to share more examples of specific mutual acquaintances who are noticeably good or bad at making themselves big or small. I think those would be be difficult and potentially privacy-invading to try to describe to strangers, though, so I'm not including them here.

That's all I have for now.

---

*LOL j/k, there are also these *~optional horsemanship notes~* (which I could rant on about for pages but which you should **feel free to skip**):*

[1] For the record, the style of horsemanship I like is pretty niche; you should not expect most people who work with horses to have heard the phrase "make yourself small" or to agree with me about what good horse training looks like.

[2] The horsemanship clips above are of Buck Brannaman, a trainer I highly respect. There's a lot of skill and subtlety to what he's doing in each clip (which I'd love to discuss with anyone interested), so I'd suggest not drawing strong conclusions about his methods based just on these short videos. If you're really interested, I recommend this documentary about him, which I highly enjoyed but which might make less sense if you have less context on training horses.

[3] If you're confused and/or curious about what Buck is doing in the video with the troubled mare: very roughly speaking, he's 1) making himself big enough that the mare pays attention to him (which - do you see it? - is much less big than he needed to be with the stallion, because she's not nearly as aggressive/oblivious); 2) showing her that if she's paying attention to him, nothing bad will happen, and she can relax; 3) making himself small to allow her to approach him while in that relaxed and attentive state.

[4] I originally learned about the idea of making yourself big or small from The Birdie Book, by Dr. Deb Bennett. (The book is named for one of Bennett's key ideas, which is that working with the horse's attention/focus - which she nicknames its "birdie" - is a key part of understanding and communicating with horses.) I'm copying here a long passage from the book about getting small - feel free to skip, but I thought it might add some helpful color.

> *One of the most moving things I ever witnessed in horsemanship was watching Harry Whitney help a frightened weanling filly. She had come from a breeding farm whose operators cynically demonstrate to clients their horses' "brilliance" and "fire" by frightening them until they retreat with rolling eyes, trembling limbs and terrified sweating, to the back corner of a large stall. The filly's new owner, a woman from Arizona, very wisely brought her to us at a nearby ranch in California, for she knew that asking this animal to make the fifteen-hour trailer trip south in such a stressed and terrified state would likely kill her.*

The moment Harry entered the pen where the filly had been placed, she began desperate attempts to flee. The pen, which was much too high for her to jump out of, was enclosed by strong wire netting. This was fortunate for it did not permit her to injure herself, which she would most certainly have done otherwise. As it was, she crashed into and bounced off of it once and then, in blind terror, ran straight at Harry, half knocking him down.

Harry's response was to retreat, very slowly, as far as he could get from her while she did likewise with respect to him. Physically as well as energetically, he made himself as small as possible. His flag (Harry's is made out of a collapsible fishing rod), remained stowed in his high-topped boots, as far out of sight as possible.

Then, from a position squatting close to the ground in one corner of the enclosure, Harry began to help the filly make some changes. Every time she would glance out of the pen, Harry would reach down to his boot and just barely crinkle the flag. The first time he did this, the filly stared at him, the whites of her eyes showing, her feet frozen to the ground, her tense and trembling body leaning stiffly away. As soon as she rolled her eyes toward him, Harry would stop the crinkling sound and resume waiting quietly. Those of us who stood watching hardly dared to breathe.

Our apprehension, however, proved unnecessary. As she spent more time regarding Harry, she began to relax. Soon she could stand still, relaxed, when Harry stood up completely straight. In another few minutes, he could take a step toward her - and then reward her for not fleeing by stepping away from her again. In half an hour, she was able to stretch her neck out to sniff his outstretched hand. A few minutes after that, Harry was petting her muzzle and her forehead. She found out it wasn't so bad. In fact, she liked it.

The second day, Harry repeated the first lessons and in a few minutes the filly was able to permit Harry to place the halter around her muzzle, and then buckle it on her head. In the same way, he then taught her to lead: a little pressure from the rope, let her feel of it, let her figure out how to relieve the pressure by stepping up, then release even more slack to her. After each bout, the filly worked her jaws as she chewed things over in her mind.

On the second day, the filly had two sessions with Harry, one in the morning and one in the afternoon, each of about 30 minutes' duration. By the end of the second lesson that day, the filly was allowing Harry to touch her all over, pick up all four feet, and lead her anywhere in her enclosure, which included a stall plus a run. She could follow Harry in and out of the door, stepping daintily over the threshhold connecting the stall to the run.

On the morning of the third day, Harry led her out of the stall. They walked all over the farm. If she showed indications that she might be getting "lost," Harry would crinkle his flag, or merely reach out to touch her. With this reminder of where her teacher was, she could relax again. It was clear that she wanted to be with Harry more than she wanted to be anywhere else. By the same internal process that underlies all affection - or if you like by the same miracle - he had become her trusted friend. He led her in and out of the owner's horse trailer, up and down the ramp, letting her find out all about it, and especially that it wasn't going to hurt her.

Using his human powers of pre-planning and foresight, Harry never got this filly into trouble, never came close to breaking her thread. This allowed her to begin to develop a much wider scale for adjustment. Some people call this "equanimity," "resiliency," or "inner calm." Others call it "emotional maturity."

On the afternoon of the third day, Harry handed the lead line to the owner. She had already learned much by watching the whole process for three days, and with a little support from Harry, she found to her delight that she too could pet, halter, lead, and load

her filly and handle her feet. We all realized that they were both going to make it just fine to their new home in Arizona.

When I expressed my admiration to Harry later in private, he said, "my biggest worry was that I might not be able to make myself small enough."

Idea: Open Access AI Safety Journal

Crossposted from the AI Alignment Forum. May contain more technical jargon than usual.

This is short because I'm mainly trying to gather feedback on how much the idea might be worth pursuing, but how much value does it seem the existence of an open access AI safety journal would provide? Some reasons I can think of in favor:

- Have a journal focused on AI safety, giving the field better visibility.
- Peer review from researchers active in AI safety research.
- Open access venue for AI safety research to avoid ethical concerns with publishing in closed journals.
- A venue where AI safety ideas too far outside the mainstream but are academically rigorous can be published.
- Give AI safety researchers a respectable venue for publishing to satisfy their career needs while reducing the effort they have to expend to get their work published, allowing more AI safety research to happen.

Some reasons against:

- There are already plenty of journals.
- Pre-prints are good enough.
- Journals are a waste of time to signal quality that is generally easy to assess from reading articles themselves.
- Most of the value could be had by regularly publishing a reviewed list of pre-prints and otherwise published articles in AI safety.
- Drag on time of folks who could be doing research instead of reviewing and editing articles.

It appears that starting an open access journal is a little tedious but not especially hard, sc. this step-by-step guide, so it could probably be done by a team of 1-3 volunteers, but hopefully could be professionalized if it takes off.

Thoughts?

A Developmental Framework for Rationality

[A way of looking at rationality as a set of transitioning worldviews that operate under different assumptions. The short summary of progression looks like techniques → habits → introspection → feelings → multiple ontologies. Skills which appear to be at odds with one another can be seen as operating under different models of what the question of rationality is even about.]

I started mindlevelup with the goal of figuring out self-improvement.

Over 2 years later and over 100 (!) essays later, I'm still at it.

Looking back, one thing that stands out is how I can identify distinct shifts in the ways I've approached this problem. The concepts I've written about, they haven't all been operating on the same level of abstraction or assumptions. Some are concrete techniques, while others are overviews of mechanisms, still others are generalizations about the process of learning the techniques, and there's even a few musing about the nature of self-improvement as a whole.

I'd like to outline what I think is a plausible way that someone's thoughts might evolve as they approach this problem, based off my own experiences. I think this will be useful in identifying what the underlying models behind different rationality skills looks like. In particular, I'll be using this to give additional context for where I think a lot of my own essays slot into the worldviews I'll be introducing.

A] Techniques Rule:

You're on a quest to find the One True Rationality.

When you first encounter rationality, it's in the context of cognitive biases, pesky glitches which lead us down suboptimal paths when we can clearly see a better way forward.

For example:

1. When we say we're going to finish something in an hour, it'll really take us more than that. Somehow, projects basically never get finished on schedule, even when everyone knows this sort of thing happens all the time!
2. Lots of say that we want to exercise, and somehow only a few of us actually get it done. This disparity exists for just about anything else, from reading more books, writing more code, or meeting new people!
3. People on different sides of a topic will get into a room to share their opinions. And, yet, when both sides leave, they end up more sure of their original opinion then before they began to argue!

Given these observations, you look at self-improvement as one of overriding your body's silly built-in defaults. The right thing to do, then, is finding algorithms that can

outperform your current ones, and then doing those instead. Under this framework, techniques are very appealing.

They're magic spells you can plug-and-chug to immediately swap up your defaults—just chant, cast, and optimize!

Things I've written which fall under this viewpoint:

1. Fighting Procrastination is all about different ways to try and deal with procrastination. Precommitment was a big thing that I recommended in that essay which I think is quite appealing under a technique-based worldview. It's simple to describe and has all the hallmarks of enforcing your "rational" desires onto your "irrational" self.
2. IAT and CEV is about a way of querying other parts of you to see what actions are "actually" best for you. I think it falls under this mindset when you use it as a tool to accommodate for things hyperbolic discounting, e.g. asking yourself "Would I *really* reflectively endorse this action?"
3. Reference Class Forecasting, Murphyjitsu, and Back-planning, the three techniques featured in the Planning 101 primer very much espouse this view of "All you have to do is <this other thing> rather than <naive thing you already do> and then things will magically get better!"

There's a definite sense of wrangling with yourself here, of saying things like "Stupid body! *I* clearly know what the *correct* thing to do is, and you're getting in my way!"

This is also characterized by a sense of wanting to collect techniques. Given the initial problem specs of overriding defaults, it feels like you need a skill to counter each bias you have. Each countermeasure you learn about feels like a little boost up. You hunger for more spells.

And yet, despite having read up about all of these great spells, you can't seem to recall the right incantations and wand movements to call upon the right one in the right moment…

B] Building Automaticity:

The problem, you reason, isn't just about beating biases,

"Just do X instead of Y" isn't a viable strategy if there isn't a way to actually get yourself to do X at the right moments. But it seems quite difficult to be intentional all the time, to be able to reflectively say, "Ah, and now is the right time to cast my spell Premorta, which will show me an instance of the future where things have gone terribly wrong!"

That just won't do.

Thus you decide to become a cyborg.

What matters, after all, you reason, is just that the desired algorithm gets executed at the right time. You turn to habits, towards *installing* the aforementioned algorithms into yourself. There's a subtle shift in focus that comes in. Whereas you previously thought about techniques in terms of nullifying your biases, the techniques

themselves come into full focus here: It's about diving deep into the question of what it "really" means to practice a rationality skill.

Key assumptions here are that humans are stimulus-response machines, and that any rationality technique can be turned into a habit. One big insight at this stage is the distinction between declarative and procedural knowledge—it's very possible for there to be an explanation of a rationality technique without giving any actual good information on how to implement said technique in real life.

Things I've written which fall under this viewpoint:

1. TAPs are the core technique here which bring everything together. TAPs are the ultimate meta-skill; here they exist as a proceduralized process that allows you to learn other proceduralized skills.
2. Hunting for Practicality has, as its core message the idea that you should be operationalizing, translating any advice you have into its most actionable form by asking yourself "How do I see myself actually acting differently as a result of <whatever information I just got> in the future?" It's all about specificity and actionability.
3. There Is No Akrasia rejects the idea of akrasia as an all-encompassing bogeyman which causes all the productivity problems. Instead, it pushes you to look directly at what's going wrong and come up with a specific fix for that thing, whatever it is.

Armed with your Hammer of Reductionism and the powerful welding power of TAPs, you get to work breaking things down and installing them. Things go well. You're not just all talk anymore—you can point to specific instances where you have indeed performed better.

And yet, it's not smooth sailing. Some of the software you install quickly becomes obsolete and stops working. Other times, your legacy code—your old defaults—seem to interact in quite chaotic ways with all of these new things you're installing. Maybe, you think, it's time to take another peek at that massive codebase…

C] Going Mental

You're back here again.

You had one look at the codebase, and it was a total mess. No documentation and a bunch of hard-coded variables.

So, against all hope, you wonder if maybe you can gain some more insights by looking at exactly what's going wrong with the installation process itself. No in-depth internals analysis needed, thank you very much.

You turn your attention inwards, on the practice, rather than what's being done behind the scenes. By paying attention to the sensations of engaging in rationality, though, you realize that focusing only on execution is futile: The phenomenon of learning skills is still a mental one, no matter how much you'd like to abstract away the messiness of the mind and focus on implementation.

You can't remove your mind from the equation.

Things I've written which explore this foray into the necessity of the internal experience of learning rationality:

1. Fading Novelty is about how any skill we try to learn is going to be subject to our brain's tendency to eventually become accustomed to the newness of said skill. Thus, we should expect initial high interest in a skill (perhaps representative of someone seeing a new technique) which eventually fades as time goes on (perhaps to find yet another skill to learn).
2. Conceptual Similarity Does Not Imply Actionable Similarity looks at the challenges that can come up when our brains substitute an easier question for a harder one because it thinks the answer to both is the same. Concretely, someone might immediately dismiss advice that sounds "obvious" as it's easier to answer the question of "Does this advice *sound* novel?" rather than "Would I benefit from taking this advice?"
3. Replace Stereotypes With Experiences is about how can often let affect and aesthetics be major decision-making factors, when in fact experience is a far better guide. There's a mix-up happening here where your feelings *about* X get conflated with your feelings *while doing* X.
4. Explication explores how vagueness can be comforting and be a place of retreat for our poorly-made plans. It's only by diving straight into the scary-feeling uncertainty and making specific claims/plans that give us an opportunity to receive feedback and verification.

Humans have an internal universe. Rationality, as a thing humans do, is going to interface with said universe at some level. When making decisions or practicing rationality, then, the ways our minds interact with our mental conceptions of rationality do in fact play a role. That much is clear now.

And yet, you can't find good solutions by just detachedly observing things from a distance.

Sighing, you roll up your sleeves.

It's time for a chat with your inner demons.

D] The Human Alignment Problem:

"Hello System 1, my old friend."

You approach yourself tentatively.

"I tried going against you, molding you to my will. Then I tried ignoring you entirely, hoping I could do well without you. But it looks like I'll need your help after all. Turns out the real Rationality was inside me all along."

You say nothing in response.

(It's yourself, after all.)

"How are we feeling right now?" you ask.

Which, of course, you already know the answer to.

Anger. Aversion. Calm. Sorrow. Worry. Want. Pleasure. Joy.

You have lots of feelings.

And they're all important.

****

Given the intrinsically mental portion of learning rationality, the direct approaches given by techniques and habits are missing a crucial component. Namely, there are situations where, despite your best efforts to set up a system to do X, *something* inside of you is resisting. The sensation of forcing yourself to do something is not a pleasant one.

So instead of trying to bind the demon to your will, you fuse with the demon.

Practically speaking, this means coming to terms with all of the intuitive, wordless, gut pieces of yourself.

One component of this frame is letting go of some sense of direct control over yourself. You accept that motivation isn't always something you can hack together with a formula. It's not that you endorse a worldview where motivation is magical and irreducible, but just that, instrumentally speaking, there are ways of becoming driven which don't involve thinking, "And now I must motivate myself to get X done!"

What you gain in return is a great deal more self-trust. You no longer need to be the person that watches over your own shoulder, making sure that you get things done. It's now much less about "forcing" yourself to do things because the things you're doing are things you want to do anyway. Internal conflict is generally removed from the equation because you're giving all of your different sides a voice.

It's not just that you've got a new technique that draws on a new way of getting techniques down. This is a viewpoint that *isn't about the techniques*.

Things I've written which are about this shift into feeling out your feelings:

1. Feelings Matter. As advertised. A slightly longer explanation about how I think about feelings and some of the social incentives for not wanting to publicly endorse feelings as important (e.g. being associated with silly self-help gurus).
2. Lying On The Ground is about cultivating feelings and sensations by, as the name suggests, lying on the ground. It presents the idea of paying attention to feelings as, not exactly a productivity technique, but as a useful thing in itself.
3. Recovering From Failure looks at the commitments you make with yourself. It takes an honest look at what's going wrong when you break them. It operates off the assumption that all the parts of yourself have needs, and commitments you break represent unmet needs.

I think the best exposition to this viewpoint is the Replacing Guilt series by Nate Soares.

Being whole with all of yourself is a powerful place to be in. You're able to resolve your inner aversions, which are often upstream blockades for lots of endeavors. Moreover, words like "procrastination" and "motivation" are a lot less meaningful now that you've blurred the distinction between your "wants" and your "shoulds".

You just do them. Because.

(What more reason do you need?)

And yet, even with this state of mind, you find yourself tempted...

E] Paradigmaster:

Having aligned yourself, you have a better measure of yourself. You know there are some things you can't do.

Or can you?

Throughout your travels, you hear whispers. You hear whispers of other powers. Other powers which promise more than alignment:

1. Resolve, an ability to surpass your limits, to endure any trial, to always power on.
2. Attractor Theory, a graceful way to go with the flow and yet preserve your autonomy and will.
3. Folding, a self-awareness even deeper than alignment, which allows you to change yourself at a fundamental level.

Which should you choose?

You've always been a good student.

You study them all.

Time passes.

You now wear more than one hat.

You've got lots of models. Some of them conflict, but they all have their use-cases.

You're a master at knowing which model to use at the right situations. Sure, there might be some underpinnings/assumptions that don't change much from one frame to the next, but the point is that you're fluid enough to recognize which way of approaching things is the most effective.

The most representative work I have of this frame is the Instrumental Rationality sequence as a whole, which is perhaps cheating because it isn't so much as one thing which switches between models, but that the different essays contained therein use different models.

At the end, you found Many True Rationalities.

And yet, you have to wonder...the whole point of all these Rationalities is to get things done, after all. And, if you look at it hard enough, isn't still just about countering cognitive biases? It's not like the central problem has *really* changed.

Looks like you're off on a quest to find the One True Rationality.

Coda:

Those familiar with Kegan's stages of development will likely see parallels here, especially as I myself was reading *In Over Our Heads* during the writing of this essay.

I took care to try and paint this framework, not necessarily one where each stage is better, but as a set of growing considerations. I think the most useful insight I can offer here is where you start to see different rationality techniques emerge as a consequence of trying to solve different components (practical, mental, ontological, etc.) of the self-improvement problem.

Extended Quote on the Institution of Academia

From the top-notch 80,000 Hours podcast, and their recent interview with Holden Karnofsky (Executive Director of the Open Philanthropy Project).

What follows is an short analysis of what academia does and doesn't do, followed by a few discussion points by me at the end. I really like this frame, I'll likely use it in conversation in the future.

Robert Wiblin: What things do you think you've learned, over the last 11 years of doing this kind of research, about in what situations you can trust expert consensus and in what cases you should think there's a substantial chance that it's quite mistaken?

Holden Karnofsky: Sure. I mean I think it's hard to generalize about this. Sometimes I wish I would write down my model more explicitly. I thought it was cool that Eliezer Yudkowsky did that in his book, Inadequate Equilibria. I think one thing that I especially look for, in terms of when we're doing philanthropy, is I'm especially interested in the role of academia and what academia is able to do. You could look at corporations, you can understand their incentives. You can look at Governments, you can sort of understand their incentives. You can look at think-tanks, and a lot of them are just like … They're aimed directly at Governments, in a sense. You can sort of understand what's going on there.

Academia is the default home for people who really spend all their time thinking about things that are intellectual, that could be important to the world, but that there's no client who is like, "I need this now for this reason. I'm making you do it." A lot of the times, when someone says, "Someone should, let's say, work on AI alignment or work on AI strategy or, for example, evaluate the evidence base for bed nets and deworming, which is what GiveWell does … " A lot of the time, my first question, when it's not obvious where else it fits, is would this fit into academia?

This is something where my opinions and my views have evolved a lot, where I used to have this very simplified, "Academia. That's like this giant set of universities. There's a whole ton of very smart intellectuals who knows they can do everything. There's a zillion fields. There's a literature on everything, as has been written on Marginal Revolution, all that sort of thing." I really never know when to expect that something was going to be neglected and when it wasn't, and it takes a giant literature review to figure out which is which.

I would say I've definitely evolved on that. I, today, when I think about what academia does, I think it is really set up to push the frontier of knowledge, the vast majority, and I think especially in the harder sciences. I would say the vast majority of what is going on in academic is people are trying to do something novel, interesting, clever, creative, different, new, provocative, that really pushes the boundaries of knowledge forward in a new way. I think that's really important obviously and great thing. I'm really, incredibly glad we have institutions to do it.

I think there are a whole bunch of other activities that are intellectual, that are challenging, that take a lot of intellectual work and that are incredibly important and that are not that. They have nowhere else to live. No one else can do them. I'm especially interested, and my eyes especially light up, when I see an opportunity to … There's an intellectual topic, it's really important to the world but it's not advancing the frontier of knowledge. It's more figuring out something in a pragmatic way that is going to inform what decision makers should do, and also there's no one decision maker asking for it as would be the case with Government or corporations.

To give examples of this, I mean I think GiveWell is the first place where I might have initially expected that there was going to be development economics was going to tell us what the best charities are. Or, at least, tell us what the best interventions are. Tell us is bed nets, deworming, cash transfers, agricultural extension programs, education improvement programs, which ones are helping the most people for the least money. There's really very little work on this in academia.

A lot of times, there will be one study that tries to estimate the impact of deworming, but very few or no attempts to really replicate it. It's much more valuable to academics to have a new insight, to show something new about the world then to try and nail something down. It really got brought home to me recently when we were doing our Criminal Justice Reform work and we wanted to check ourselves. We wanted to check this basic assumption that it would be good to have less incarceration in the US.

David Roodman, who is basically the person that I consider the gold standard of a critical evidence reviewer, someone who can really dig on a complicated literature and come up with the answers, he did what, I think, was a really wonderful and really fascinating paper, which is up on our website, where he looked for all the studies on the relationship between incarceration and crime, and what happens if you cut incarceration, do you expect crime to rise, to fall, to stay the same? He really picked them apart. What happened is he found a lot of the best, most prestigious studies and about half of them, he found fatal flaws in when he just tried to replicate them or redo their conclusions.

When he put it all together, he ended up with a different conclusion from what you would get if you just read the abstracts. It was a completely novel piece of work that reviewed this whole evidence base at a level of thoroughness that had never been done before, came out with a conclusion that was different from what you naively would have thought, which concluded his best estimate is that, at current margins, we could cut incarceration and there would be no expected impact on crime. He did all that. Then, he started submitting it to journals. It's gotten rejected from a large number of journals by now [laughter]. I mean starting with the most prestigious ones and then going to the less.

Robert Wiblin: Why is that?

Holden Karnofsky: Because his paper, it's really, I think, it's incredibly well done. It's incredibly important, but there's nothing in some sense, in some kind of academic taste sense, there's nothing new in there. He took a bunch of studies. He redid them. He found that they broke. He found new issues with them, and he found new conclusions. From a policy maker or philanthropist perspective, all very interesting stuff, but did we really find a new method for asserting causality? Did we really find a new insight about how the mind of a perpetrator works? No. We didn't advance the frontiers of knowledge. We pulled together a bunch of knowledge that we already had,

and we synthesized it. I think that's a common theme is that, I think, our academic institutions were set up a while ago, and they were set up at a time when it seemed like the most valuable thing to do was just to search for the next big insight.

These days, they've been around for a while. We've got a lot of insights. We've got a lot of insights sitting around. We've got a lot of studies. I think a lot of the times what we need to do is take the information that's already available, take the studies that already exist, and synthesize them critically and say, "What does this mean for what we should do? Where we should give money, what policy should be."

I don't think there's any home in academia to do that. I think that creates a lot of the gaps. This also applies to AI timelines where it's like there's nothing particularly innovative, groundbreaking, knowledge frontier advancing, creative, clever about just… It's a question that matters. When can we expect transformative AI and with what probability? It matters, but it's not a work of frontier advancing intellectual creativity to try to answer it.

A very common theme in a lot of the work we advance is instead of pushing the frontiers of knowledge, take knowledge that's already out there. Pull it together, critique it, synthesize it and decide what that means for what we should do. Especially, I think, there's also very little in the way of institutions that are trying to anticipate big intellectual breakthroughs down the road, such as AI, such as other technologies that could change the world. Think about how they could make the world better or worse, and what we can do to prepare for them.

I think historically when academia was set up, we were in a world where it was really hard to predict what the next scientific breakthrough was going to be. It was really hard to predict how it would affect the world, but it usually turned out pretty well. I think for various reasons, the scientific landscape maybe changing now where it's … I think, in some ways, there are arguments it's getting easier to see where things are headed. We know more about science. We know more about the ground rules. We know more about what cannot be done. We know more about what probably, eventually can be done.

I think it's somewhat of a happy coincidence so far that most breakthroughs have been good. To say, I see a breakthrough on the horizon. Is that good or bad? How can we prepare for it? That's another thing academia is really not set up to do. Academia is set up to get the breakthrough. That is a question I ask myself a lot is here's an intellectual activity. Why can't it be done in academia? These days, my answer is if it's really primarily of interest to a very cosmopolitan philanthropist trying to help the whole future, and there's no one client and it's not frontier advancing, then I think that does make it pretty plausible to me that there's no one doing it. We would love to change that, at least somewhat, by funding what we think is the most important work.

Robert Wiblin: Something that doesn't quite fit with that is that you do see a lot of practical psychology and nutrition papers that are trying to answer questions that the public have. Usually done very poorly, and you can't really trust the answers. But, it's things like, you know, "Does chocolate prevent cancer?" Or, some nonsense … a small sample paper like that. That seems like it's not pushing forward methodology, it's just doing an application. How does that fit into to this model?

Holden Karnofsky: Well, I mean, first up, it's a generalization. So, I'm not gonna say it's everything. But, I will also say, that stuff is very low prestige.

And, I think it tends … so first off, I mean, A: that work, it's not the hot thing to work on, and for that reason, I think, correlated with that you see a lot of work that isn't … it's not very well funded, it's not very well executed, it's not very well done, it doesn't tell you very much. The vast majority of nutrition studies out there are just … you know, you can look at even a sample report we did on carbs and obesity that Luke Muehlhauser did, it just … these studies are just … if someone had gone after them a little harder with the energy and the funding that we go after some of the fundamental stuff, they could have been a lot more informative.

And then, the other thing is, that I think you will see even less of, is good critical evidence reviews. So, you'll see a study … so, you're right, you'll see a study that's, you know, "Does chocolate more disease?" Or whatever, and sometimes that study will use established methods, and it's just another data-point. But, the part about taking what's out there and synthesizing it all, and saying, "There's a thousand studies, here are the ones that are worth looking at. Here are their strengths, here are their weaknesses."

There are literature reviews, but I don't think they're a very prestigious thing to do, and I don't they're done super great. And so, I think, for example, some of the stuff GiveWell does, it's like they have to reinvent a lot of this stuff, and they have to do a lot of the critical evidence reviews 'cause they're not already out there.

The most interesting parts of this to me were:

- Since reading Inadequate Equilibria, I've mostly thought of science through the lens of coordination failures; however this new framing is markedly more positive, which instead of talking about failures talks about the successes (Old: "Academia is the thing that fails to do X" vs New: "Academia is the thing that is good at Y, but only Y"). As well as helping me model academia more fruitfully, I honestly suspect that this framing will be more palatable to people I present it to.
    - To state it in my own words: this model of science says the institution is good - not at *all* kinds of intellectual work, but specifically the subset that is 'discovering new ideas'. This is to be contrasted with synthesis of old ideas into policy recommendations, or replication of published work (for any practical purpose).
    - For example within science it is useful to have more data about which assumptions are actually true in a given model, yet I imagine that in this frame, no individual researcher is incentivised to do anything but publish the next *new* idea, and so nobody does the replications either. (I know, predicting a replication crisis is very *novel* of me.)
- This equilibria model suggests to me that we're living in a world where the individual who can pick up the most value is not the person coming up with new ideas, but the person who can best turn current knowledge into policy recommendations.
    - That is, the 80th percentile person at discovering new ideas will not create as much value as the 50th percentile person at synthesising and understanding a broad swathe of present ideas.
        - My favourite example of such a work is Scott's Marijuana: Much More Than You Wanted to Know, which finds that the term that should capture most of the variance in your model (of the effects of legalisation) is how much marijuana affects driving ability.

    - Also in this model of science, we should distinguish 'value' from 'competitiveness within academia', which is in fact the very thing you would be trading away in order to do this work.

Some questions for the comments:

- *What is the main thing that this model doesn't account for / over counts?* That is, what is the big thing this model forgets that science can't do; alternatively, what is the big thing that this model says science can do, that it can't?
- *Is the framing about the main place an intellectual can have outsized impact correct?* That is, is the marginal researcher who does synthesis of existing knowledge in fact the most valuable, or is it some other kind of researcher?

On Dualities

I’ve been finding the word duality useful quite a bit recently. I find it very useful for describing situations where there are two very valuable perspectives (or lens) through which we can look at a situation and any attempt to answer the question needs to grapple with and account for both of these. The way I use the word, I’m not claiming that two logically contradictory viewpoints are simultaneously true, but rather that aspects of both can be synthesised together to reach the truth.

Here’s a few examples. Maybe you agree or disagree with these specific examples, but I think they should suffice for illustrative purposes:

- Some people are born with major disadvantages and we need to be sympathetic to them. At the same time, people can act in a way which makes their decisions better or worse and we need to encourage personal responsibility. If we’re too harsh, we don’t give them the help that they need, if we’re too sympathetic, we simply enable people to ruin their own lives. We need to find a balance between the two
- On one hand, tolerance is important for enabling us to get along with people who are different from us. On the other hand, too much tolerance means that there are no standards of behaviour. We need find a way to merge these two considerations without simply tolerating that which we approve and refusing to tolerate that which we dislike (it is possible to synthesise a worst of both worlds approach)
- On one hand, a person can be acting in a way that is horrific from a moral standpoint, on the other hand, it may be hard to blame them given the circumstances. This sets up a tough conflict between the demands of justice and the desire to be merciful and it can be hard to figure out how to navigate these systems

Dualities allow you to avoid a particular dysfunctional pattern of thought. Most people will have one side of the duality stick out for them more than the other. Since both sides contradict (or seem to contradict), they conclude that they must reject the other side of the duality. This resolves the contradiction, but it doesn’t give them the truth. Finding the truth would require considering both sides, but now that they’ve settled for consistency instead, they’ve prevented themselves from making further progress. There is value in having a part of the truth-finding process where you can identify ideas as containing a lot of truth without having to worry about whether they contradict, at least temporarily. Thinking in terms of dualities encourages this.

I really value consistency and I encourage others to value it too, but perhaps I should value it a bit less as overvaluing consistency seems to be what leads to these kinds of mistakes. If your model of the world contains an unsynthesised duality, then this won’t allow you act consistently, but it may still give you a more accurate model of the world than picking one side or the other. Of course, attempting to eventually synthesise these dualities should be the eventual goal, but we should also recognise that this may simply be beyond our own personal abilities and that we may never be able to completely remove the duality.

It is also very useful for explaining ideas that you haven't completely worked out in your head. Quite often you are aware that there are two main sides of the issue, both of which make good points and you want to establish that you will be drawing on both

sides. Once you've set that up, you can then move towards synthesise. Even if you've already figure out a synthesis, there can be value in taking the listener on the same intellectual journey so that they can understand your conclusions. Also, undoubtedly there are times when you want to synthesise more than two ideas. I don't have a word for this, but it is much less common.

Why use the word "duality"?

The word duality communicates this concept well and if you want to be understood, you need to use language that people understand. Undoubtedly, many people don’t want to use this word as it is usually associated with forms of mysticism and the belief that something can be two things that logically contradict at the same time. This is not a viewpoint that I embrace at all, but I’m not convinced that using the word “duality” encourages this to the extent where I would want to stop. After all, while it may encourage more people to use the word duality and some of these uses may be incoherent, using it properly may also led some people to replace incoherent uses with coherent uses. So I see the overall effect as something of a wash. I would actually prefer the word dialectic, as it connotes the idea of two opposites which can then be synthesised together, but I don’t feel it’s meaning is known well-enough.

Explanation of Paul's AI-Alignment agenda by Ajeya Cotra

This is a linkpost for https://ai-alignment.com/iterated-distillation-and-amplification-157debfd1616

Ajeya from OpenPhil wrote a very understandable and quite compelling summary of Paul's views on AI-Alignment.

(@Paul & @Ajeya: I would love to crosspost the whole thing here if you are open to that)

Naming the Nameless

Epistemic status: political, opinionated, personal, all the typical caveats for controversial posts.

I was talking with a libertarian friend of mine the other day about my growing discomfort with the political culture in the Bay Area, and he asked why I didn't just move.

It's a good question. Peter Thiel just moved to L.A., citing the left-wing San Francisco culture as his reason.

But I like living in the Bay, and I don't plan to go anywhere in the near future. I could have said that I'm here for the tech industry, or here because my friends are, or any number of superficially "practical" reasons, but they didn't feel like my real motivation.

What I actually gave as the reason I stay was... *aesthetics*.

Wait, what?

Let's Talk About Design

I'm not a designer, so I probably don't have the correct vocabulary to express what I see. Please bear with me, while I use simple and ignorant language; if any of my readers have a more sophisticated understanding, I'd love to hear about it in the comments.

Stuff that's marketed to Bay Area bourgeois bohemians has a coherent appearance. You see it in websites that are all smooth scrolling and gradients and minimalism -- see the sample websites on Squarespace, for instance. You see it in the product design on the labels and menus of cafes and juice bars and coffee shops -- The Plant Cafe is a good example. You see it in the almost-identical, smoothly minimalist layouts of every tech-startup office.

Professional designers may be getting bored of this "light-contrast, minimalist elegance" or "objectively beautiful, but mostly unremarkable, templates", and are trying out more deliberately jarring styles like Brutalism.

But for your typical consumer, the generic California/BoBo style works fine. It signals elegance, which means, more or less, that it's designed for educated, high-Openness, upper-middle-class, urban people. When I enter a space or a website with this aesthetic, or buy a product with this branding, it's shorthand for "Ahhhh, this place is run by competent professionals who know how to give me a pleasant experience. I will not feel harried or inconvenienced or confused here; I will be well taken care of. I will easily be able to slot my existing behavior patterns into the implicit "rules" of how to use and navigate this place or device or website."

Apple products are, of course, the archetype of this kind of "good" design. Smooth, urbane, almost childishly easy to use. Most computers are still PCs; office workers, older people, hardcore programmers and gamers, and the price-conscious still go for PCs. It's among the *style-conscious* (who skew affluent, educated, aesthetically/socially sensitive, and slightly more female than male) that Macs are

universal. When I asked a Marine from Texas what kind of computer he used, he scoffed, *Do I look like a Mac guy?*

Let's look at one of my favorite things to buy, G&T's Kombucha.

This is pretty much the most BoBo thing in the world. Its packaging makes a nod to Buddhism ("Enlightened", the mandala-like radially symmetric logo), psychedelia (the rainbow label), Human Potential Movement-ish self-improvement ("SYNERGY" and "renew, rebalance, rebuild, reclaim, rekindle, recharge") and environmentalism ("organic"). But the design is simple and clean enough to seem like a modern company run by professionals.

In this case, it's not *just* a pretty label: the probiotics in fermented foods like kombucha are probably good for you, kombucha is lower in sugar than juice but pleasantly tangy and fizzy, and in my experience it's uncannily good at settling an upset stomach. But the branding is a big part of what makes it delightful. And, I'm almost embarrassed to say, being able to buy kombucha at the nearest drugstore is a non-negligible part of why I like living in this neighborhood.

Style-Blindness

I have a friend who's very good at digging up evidence of crime and scam artistry. It's part hobby, part crusade; give her a public figure and she can investigate with great speed and accuracy what kinds of shady dealings he's been involved with.

Once, she showed me some companies she had proved were fraudulent, and my first reaction was "I could have told you that in seconds; their web design looks scammy."

Of course, it's not really the same thing. She had hard evidence; I only had an intuition, and intuition can be wrong.

But, for instance, this penis enlargement website just looks *noisy*. It's jam-packed with content, it's screaming about sales and deals, there's a bright red "Buy Now" button with a ticking countdown clock. It's not *classy*. Even if you didn't know anything about the product, you could see that it's being packaged (pun intended) much differently than this website selling relationship workshops.

But my friend, like a lot of nerds, couldn't see that difference in branding at a glance. She couldn't see the difference in connotations that different aesthetic choices evoke. She was almost completely *style-blind*.

Some people claim that aesthetics don't mean anything, and are very resistant to the idea that they could. After all, aesthetic preferences are very individual. Chinese opera sounds beautiful to people raised with it, and discordant to the untrained Western ear.

So, claim the skeptics, all descriptions of what aesthetic choices "mean" are basically pseudoscience. When design experts tell us that red evokes passion and blue evokes calm, they're using associative thinking, which is no more fact-based than the Four Elements or the five colors in Magic: The Gathering.

Clustering things based on associations and connotations is risky. It's going to differ from individual to individual, and even more from culture to culture. It's easy to take intuitive leaps for granted and quickly get to the point where people are talking past each other. So it's safer just not to talk about what aesthetics connote, right?

To my view, the skeptics have a good point, but they're too epistemically conservative. There's *obviously* signal being carried through aesthetics. Colors don't have intrinsic meanings, of course, but they do have shared connotations within a culture.

Note that the M:TG color "meanings" and the design/marketing color "meanings" are very similar -- not because everyone is tapping into some magical collective unconscious, but because Magic is a game designed in contemporary America, by designers who probably share the same color associations as the designers of websites and product labels.

When Pantone says their 2018 color of the year, Ultra Violet, "communicates originality, ingenuity, and visionary thinking", they're not just making up random nonsense. Pretty much any present-day English-language "color meaning" summary for designers or marketers will associate purple with something like creativity or imagination or spirituality. I don't know where this meme comes from originally, but it's certainly not unique to Pantone or chosen at random.

Our physical environment is built primarily by corporations which employ designers. Those designers draw inspiration from artistic or creative subcultures. Design has a life cycle in which it starts as an original aesthetic trope being used by some individual artist, to being imitated by other artists, to becoming trendy, to becoming ubiquitous. Tastemakers may be a tiny minority of the population, aesthetics may not be a big deal for everyone, but *everything manmade you see around you has its origins in someone obsessed with aesthetics.* Designers "rule" our visual world in the same way writers "rule" our verbal world, in the same way that "practical men who believe themselves to be quite exempt from any intellectual influence, are usually the slaves of some defunct economist." In this sense, aesthetics *very much* mean things, and you have to look to their origins and contexts to understand what they mean.

This essay, worth reading in full, calls the process "subcultural sublimation" and tracks how Pantone's 2016 colors, Rose Quartz and Serenity, drew inspiration from seapunk (a musical subgenre with an online visual aesthetic). Seapunk aesthetics propagated through fashion blogs, the NYT style section, and pop stars' music videos, all the way to the Pantone Institute, which sets the tone for fashions in mainstream commercial design. The popularity of pastels began with feminist artists interrogating softness and femininity, propagated through Tumblr "aesthetic" blogs, and likewise eventually reached Pantone. Aesthetic tropes are "commodified" over time; they drift from artistic or countercultural milieux towards corporate branding.

Mostly implicit in the article, but worth mentioning, is that *commercial design ultimately borrows from creatives who are politically opposed to business and resent this commercial appropriation.* More on that later.

If you're style-blind, you'll look at Rose Quartz and Serenity and say "they're just colors! they don't mean anything! all this cultural criticism is just pretentious noise!" If you're mildly style-sensitive, like myself, you'll notice that the colors seem Tumblresque, and you'll note that Pantone's description makes a nod to "gender blur" and "societal movements toward gender equality and fluidity". If you're actually an expert, like the author of the article, you can concretely trace where the popularity of that color scheme came from.

"Subcultural sublimation" runs on ordinary, non-magical cause and effect, the propagation of memes from their originators towards mass popularity. It can be

understood and analyzed. You can isolate where aesthetic tropes come from, why they're used, what their creators believe, and what channels govern their imitation and spread -- and that tells you something about their "meaning" that's not purely subjective.

Politics and Aesthetics

Artists tend to be on the political left; arts and media occupations are among the most heavily weighted towards Democrats over Republicans.

It's not clear to me why. Maybe it's a temperamental thing -- high openness to experience drives both an interest in aesthetics and a preference for left or liberal politics. Maybe it's explained by education, which both inculcates interest in the arts and left politics. Regardless of cause, it's a real and important phenomenon. And it's a problem for anyone who's not on the left, as Rod Dreher, the original CrunchyCon, pointed out years ago.

Beauty matters to people. So does health and emotional wellbeing. So does everyday kindness. *Living well*, in other words. Quality of life. You can't cede all of that to the opposing political team without losing something valuable.

Rod Dreher points out that, while, say, organic vegetables are coded liberal, they also taste better and are healthier than processed food. Yet conservatives often have a knee-jerk condemnation of anything "green" or "pretentious", which means they're boxed into being cultural philistines who miss out on flavor and beauty and health.

“It’s a PR disaster for the Right to allow discussions of fun and beauty and poetry and nature to be owned by the Left,” says a New York publishing executive and closet conservative. “The right wing just looks unappealing. Do they not understand this?”

If you like the arts, if you're temperamentally high-Openness and aesthetically sensitive, you're going to be drawn to coastal cities and educated social groups, and those environments tend to skew left-wing. It's hard to leave without giving up something intangible that's hard to convey to people who don't share your sensibility.

Dreher, a conservative Catholic who values tradition, can with some justice argue that beauty and art properly belong to *his* culture; after all, it was Catholics who built the cathedral of Chartres.

Libertarians are, if anything, in a tougher position, because we're *not* traditionalists, and because strong individualism runs counter to even being able to talk about shared cultural sensibilities. Ask a libertarian "Why don't we have any good songs about *our* values?" and there's a good chance that you'll get the response "Ew, who'd want one? That's too collectivist for me."

But the result is that you're living in an aesthetic environment that's largely created by your ideological opponents, and subjected to constant subliminal messaging that your values are uncool. This causes an evaporative cooling effect where the only people willing to express libertarian views are "style-blind" and sometimes even socially blind, people who do not perceive that they are being mocked or that their aesthetic signaling is clumsy.

It's hard to argue to a skeptic why this even matters. Why care about aesthetics and culture? What do you care what other people think? Surely an independent-minded person would simply refuse to succumb to social pressure -- and the cultural

connotations of aesthetics are inherently relative to social context, so maybe the best way to keep your independence is to choose style-blindness as a cognitive strategy. What you can't see, you can't be manipulated by!

But I think it's unvirtuous to choose blindness or ignorance. And it's also ineffective. What you can't see can *sneak up behind you.* People who think they're immune to social pressure get manipulated all the time.

Scott Alexander is honest enough to admit that it happens to him:

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

Now, I could say "just don't do that, then" -- but Scott of 2009 would have *also* said he believed in being independent and rational and not succumbing to social pressure. Good intentions aren't enough.

And I'm seeing people in roughly my demographic going silent or submitting to pressure to conform, and it's worrisome.

I think it's much better to try to make the implicit explicit, to bring cultural dynamics into the light and understand how they work, rather than to hide from them.

Defensive Postures

There are a number of defensive strategies people (of varying political views) adopt against the cultural dominance of the left.

Reaction is what, say, Ann Coulter does, or Breitbart.com, or the Donald Trump campaign. It's defiantly *anti-* progressive, rejecting the "mainstream media" and "coastal elite" tastemakers. It's happy to be perceived as tacky and rude.

The problem with reaction is that it has no positive vision. It's just "the opposite of what my opponents want." It's uncreative and it can easily descend into spitefulness.

Respectability politics is a different tactic, and, in this context, usually takes the form of (not very credible) claims to be apolitical. Early forms of this include "Keep Your Identity Small" or "Politics is the Mind-Killer." By declaring the importance of not taking sides, you're *already asserting* that you're not wholly on one side; a progressive can reasonably infer that any avowedly "apolitical" person disagrees with them at least somewhere.

Claims of aloofness from politics have always, correctly, been identified as evidence of covert dissent from "good" politics: "formalism" was a political offense in Soviet Russia. There are many thinkpieces like this one observing (rightly) that Silicon Valley culture is nominally apolitical but implicitly capitalist.

And then you see obviously defensive moves by the tech industry to distance itself from that allegation, like YCombinator's announcement of its New Cities project:

> Just to get ahead of the inevitable associations: We want to build cities for all humans - for tech and non-tech people. We're not interested in building "crazy libertarian utopias for techies."

Once you have to defend against a stereotype, you're already losing the messaging war. As with reaction, there's no positive vision, only the frantic assurance that you're not really the bad guy.

Cooptation doesn't seem to be that popular, and might be underrated.

It's a kind of judo where you claim to be the *true* exemplar of the goal your opponents want. They hate capitalism? Well, you note that what most people think of when they hear that word is *crony* capitalism, which is indeed terrible, and that you are bitterly opposed to the system in which unfair legal privileges give vast wealth to a few and deprive everyone else. C4SS does this, quite well in my opinion, but hardly anyone outside of libertarian-world has heard of them.

It's still not fundamentally creative, though. You're *borrowing* your opponents' tropes and aesthetics, not building your own. And if you get too good at it, you end up being easily confused for believing things that you don't actually believe.

The Opposite of Defensiveness

One of the things I like best about Ayn Rand is that she staked out aesthetic and cultural territory *without* resorting to any of these defense mechanisms. She actually *made art* that was fundamentally in a different style than that of the cultural establishment. Of course, this left her vulnerable to the allegation that it was *bad* art -- there are 52 million Google results for "ayn rand bad art."

But most of the common criticisms -- of black-and-white thinking, didacticism, utopian optimism, overly heroic characters, and so on -- are based on implicit presumptions about the nature of life and the role of art which she explained (or, at least, began to explain) why she did not share. She brought the dissent into the light, into explicit discourse.

If you take something about yourself that's "cringeworthy" and, instead of cringing yourself, try to look at *why* it's cringeworthy, what that's made of, and dialogue honestly with the perspective that disagrees with you -- then there is, in a sense, nothing to fear.

There's an "elucidating" move that I'm trying to point out here, where instead of defending against an allegation, you say "let's back up a second" and bring the entire situation into view. It's what double crux is about -- "hey, let's find out what even is the disagreement between us." Double crux is hard enough with *arguments*, and here I'm trying to advocate something like double-cruxing *aesthetic preferences*, which sounds absurdly ambitious. But: imagine if we could talk about *why* things seem beautiful and appealing, or ugly and unappealing. Where do these preferences come

from, in a causal sense? Do we still endorse them when we know their origins? What happens when we *bring tacit things into consciousness*, when we talk carefully about what aesthetics evoke in us, and how that might be the same or different from person to person?

Unless you can think about how cultural messaging works, you're going to be a *mere* consumer of culture, drifting in whatever direction the current takes you.

The Arts and Imitation

Let's go back for a moment to subcultural sublimation.

Artistic trends have a life cycle, of *creation, expansion, and destruction*, or more specifically, *the artist, the marketer, and the critic*. First, the artist creates a new thing. Then, a succession of tastemakers and creatives imitate that thing and *scale it up*, from a subcultural scene to mass-market production. Finally, the critic notices that it's become commoditized (in the literal economic sense: if it's exactly the same everywhere and anyone can copy it, its price goes to zero) and deflates the hype.

This isn't specific to the arts, of course. Companies are created, expand, and eventually succumb to competition. Empires are founded, expand, and succumb to invaders. It's a human-organization pattern.

But expansion in particular is enabled by mechanical reproduction processes dating to the Industrial Revolution. We can systematize "scaling up" much easier and faster than pre-industrial peoples could.

Commerce is ancient -- in different times and places, trade has been more free or less so, and it became somewhat more free in the West with the introduction of classical liberalism and economic theory at the end of the 18th century, but trade itself is as old as the first anatomically modern humans, living 300,000 years ago.

Invention is ancient -- the Greeks had it, including more advanced science than modern stereotypes would assume. Archimedes probably knew calculus.

What's modern is *scaling-up*, the ability to make many copies of things, from physical objects to social systems. That's what allows for mass culture. That's what allows startups to grow exponentially. For the past two hundred years or so, we've been living in an era where the *expander* of the reach of a creation is more powerful than ever.

Expanders sometimes like to present themselves as creators, but they're not. The creator makes the first prototype, the original. No scale at all. "Zero to one." In fact, creators often resent expanders for taking credit for their work or diluting it for the mass audience. This is why seapunk artists were frustrated at being imitated in music videos:

> also, why aren't y'all frustrated AT ALL at the rihanna thing? that performance marked the commodification of an aesthetic movement…— Bebe Zeva (@BebeZeva)
> …which means all taste-makers have to start all over. it's a lot of work. clearly ur not doing shit but consuming if ur not peeved by this— Bebe Zeva (@BebeZeva)
> "wow amazing rihanna performance i love seeing my tumblr on SNL" why? that Aesthetic served as an exclusive binder for URL counterculture…— Bebe Zeva (@BebeZeva)

> …tomorrow, when it enters Phase Three and Forever 21 puts a price tag on it, it will no longer be exclusive. its purpose is gone.— Bebe Zeva (@BebeZeva)

My own addition to the pile of theories on "why don't creative professionals like capitalism?" is that *creators feel defrauded by expanders, and the core of modern capitalism is superpowered expanders.* Expanders capture most of the economic value and social credit from scaling up things originated by creators. Expanders are sociopaths, in the "geeks, mops, and sociopaths" trichotomy.

And we don't really have good tools for fairly compensating people for intellectual originality. Intellectual property law is a kludge, with a lot of problems. Creators don't really know how to extract "fair market value" for ideas, possibly because they're intrinsically motivated to create them and the kind of "payment" they want is more like appreciation or kindred-spirit-ness than money. Standard startup ideology says that ideas are of low value: "If you go to VC firms with a brilliant idea that you'll tell them about if they sign a nondisclosure agreement, most will tell you to get lost. That shows how much a mere idea is worth. The market price is less than the inconvenience of signing an NDA." That may be true, but you could also interpret it as *markets not knowing how to price ideas*, in the same way that markets can't price natural resources until you figure out a way to define property rights over them.

So, whenever you encounter a piece of media -- words or images or music or anything representational -- no matter how many levels of imitation or expansion it's been through, you're still hearing some distant signal from its originator. And its originator probably feels ripped off and undervalued. When you go looking for good art, you're looking for art that's closer to its creative source, and that means you'll hear in it the voice of the frustrated creator.

In a sense it's inherently paradoxical to enjoy something like G&T's Kombucha -- it's a product produced by a process (scaling-up) which the hippies who inspired its aesthetic would have vehemently opposed. To like it *knowledgeably* is to partly *dislike* it.

I think there may be some kind of necessary project in the vicinity of "making amends between creators and expanders" that would be required for creative work *not* to have the dynamic where scaling up is seen as selling out. I think scaling-up is *probably net good* -- it allows more people to have nicer things. But there may well be legitimate grievances with it that deserve to be addressed. That's another one of those cases where dialogue and making the implicit explicit would be really helpful.

New Paper Expanding on the Goodhart Taxonomy

This is a linkpost for https://arxiv.org/pdf/1803.04585.pdf

On the Loss and Preservation of Knowledge

This is an excerpt from the draft of my upcoming book on great founder theory. It was originally published on SamoBurja.com. You can access the original here.

Let's say you are designing a research program, and you're realizing that the topic you're hoping to understand is too big to cover in your lifetime. How do you make sure that people continue your work after you're gone? Or say you are trying to understand what Aristotle would think about artificial intelligence. Should you spend time reading and trying to understand Aristotle's works, or can you talk to modern Aristotelian scholars and defer to their opinion? How can you make this decision? Both of these goals require an understanding of traditions of knowledge — in particular, an understanding of whether a tradition of knowledge has been successfully or unsuccessfully transmitted. But first: what is a tradition of knowledge?

A *tradition of knowledge* is a body of knowledge that has been consecutively and successfully worked on by multiple generations of scholars or practitioners. In talking about a tradition of knowledge, we may be talking about a philosophical school of thought, or perhaps a tradition of intricate rituals in a religion, or even something as humble as the knowledge of how to fashion the best wooden toy horse, passed down from one craftsman to another. In the contemporary world, it may include something like the tacit knowledge of how a codebase really works, which senior engineers teach to junior engineers. It is useful to classify traditions of knowledge into three types: living, dead, and lost traditions.

A *living* tradition of knowledge is a tradition whose body of knowledge has been successfully transferred, i.e., passed on to people who comprehend it (e.g., cryptography). The content of the tradition's body of knowledge does not have to be strictly or fully accurate for the tradition to be living; it merely needs to be passed on.

A *dead* tradition of knowledge is a tradition whose body of knowledge has been unsuccessfully transferred, i.e., its external forms, its trappings such as written texts have been transferred, but not the full understanding of how to carry out this tradition of knowledge as practiced (e.g. scholars who can recite Aristotle but can't use arguments as he did; Buddhist monks who chant the instructions to meditation rather than meditating). This means a tradition can be dead while people still read its texts.

A *lost* tradition of knowledge is a tradition that has not been transferred at all (e.g. numerous schools during the Hundred Schools of Thought period in China; the theology of the Cathars, which is only preserved in the words of their critics, etc). The people who had the knowledge died without leaving any successors or substantial record of their knowledge.

It can be difficult to distinguish between different traditions of knowledge. There are traditions within traditions, and there are traditions that become fellow travelers, in the sense that they are related to but merely adjacent to one another. There are also traditions that have a long history of arguing against each other. Perhaps the best example of such traditions are to be found in the realm of theology: the multitudinous schisms between and within the major branches of Christianity, such as the intra-Protestant debate between Calvinists and Arminians which began in the 16th-century Netherlands and continues to this day among some evangelicals; the centuries-long debate began in the 3rd century AD between the dominant Vaibhāṣika school of Buddhism and its successor faith of Sautrāntika in the patchwork of northern Indian states following the fall of the Mauryan Empire; and the infamous warring between Sunni and Shia Islam. We can find other examples in political thought: the ancient debates between Confucians and Legalists in China, the enemy factions of Anglo-American liberalism and conservatism, and the debate between the Originalist and Living Constitutionalist schools in American constitutional interpretation, to name a few.

It matters whether a tradition of knowledge is living or dead. This is obviously the case if you are starting a research program — you want the tradition you start to stay alive. Whether or not the Aristotelian tradition is dead also matters if you are trying to understand what Aristotle would have thought about artificial intelligence: it determines whether or not you can trust the “authorities” on Aristotle — if the tradition is dead, then their expertise will not be helpful to you. It also matters if a tradition of knowledge is lost: this will inform your understanding of what it is possible to know about that tradition. For now, we will focus on understanding how to distinguish between a living and a dead tradition. This can be tricky; it’s hard to trace traditions of knowledge, so it’s also hard to notice when they die.

How can you tell whether a tradition of knowledge is living or dead? First, you have to be able to identify signs that indicate the existence of a tradition of knowledge. You have to be able to recognize signs that indicate the existence of a tradition at all, then determine whether those signs taken together indicate that the tradition is dead or that it is alive. The signs used to recognize the existence of a tradition are the same signs used to distinguish between living and dead traditions.

Signs of Traditions of Knowledge

Signs that indicate the existence of a tradition of knowledge vary in the degree to which they indicate that a tradition is alive -- that understanding has been passed on. A collection of signs that weakly or do not at all indicate continuity of understanding, without any signs that strongly indicate continuity of understanding, is a sign that the tradition under investigation is dead. Below are some common signs.

Sign of tradition of knowledge	Example of this sign
Production of a notable effect	Powerful generals; well-balanced Damascus steel swords
Shared methodology (even if not explicit)	Plasmid transfection techniques in synthetic biology.
Shared concepts (even if under different name)	"Diagonalization argument" or "Yao's minimax principle" in mathematics.
Shared conceptual framework or theories	The Standard Model in contemporary particle physics.
Extension of the theory in the tradition	Mencius' work on Confucian philosophy.
Master/apprentice relationships	An apprentice signing up with a master stonecutter for several years of service. A master glassmaker in his old age supported by his former apprentice.
Explicit knowledge of specific arguments	The argument that it is naive to try and predict the effects of economic policy entirely on the basis of relationships observed in historical data is known as the 'Lucas critique' in economics.
Shared terminology	A surgeon might use the terms "proximal" and "distal" to describe locations on the human body.
Accreditation system (depends of health of institution)	An MIT program that awards graduates PhDs in biochemistry.
References to specific authors	Mathematicians sometimes recommend 'reading Artin' for algebra.
Familiarity with a person's works	A classicist can recite verses from Homer's Iliad.
A physical location where the tradition is ostensibly kept	University of Oxford campus. The Orthodox Christian monasteries of Mount Athos.

Figure: Signs of traditions of knowledge. These are listed roughly in order from best to worst indicators of a living tradition.

It's important to remember that in order to trace traditions, *you have to investigate the actual transfer of knowledge.* This means that you can't, for example, rely on the existence of a physical location where the tradition is supposedly kept to justify that the tradition is alive. There are many possible scenarios in which a tradition has died or been lost, and yet the physical location of its origin has been preserved. A useful way of determining whether a tradition of knowledge exists and is living is by investigating chains of master/apprentice relationships. When looking at the works of masters and apprentices, you can tell whether there are shared methods, concepts, ideas, and so forth.

Furthermore, the existence of master-apprentice relationships at all is an indicator of a living tradition, because master-apprentice relationships are especially effective means of knowledge transfer. This is borne out by the historical record. For example, Kongō Gumi, the world's oldest continuously-operating company and a family-owned construction firm based in Osaka, Japan, has extensively used the practice of *mukoyōshi*—by which a son-in-law is formally adopted into the family as an apprentice and eventual company owner—to stay in business since the year 578.

Live Traditions

What keeps a tradition of knowledge alive? First, let's review our definition of a living tradition of knowledge: a living tradition of knowledge is a tradition in which either its founders are still alive and practicing, or its body of knowledge has been successfully transferred, i.e. passed on to people who comprehend it. There are multiple features of a living tradition that we can look for in order to determine whether a tradition of knowledge is alive or dead.

Transfer of Verification Mechanisms

Scholars and practitioners in a body of knowledge will often use discrete techniques or mechanisms to verify their work for accuracy. This is, essentially, a form of quality control that allows new work in a tradition of knowledge to be verified against reality.

Whether it’s an oral examination at a medieval university, Napoleon riding into camp unannounced to review the troops, or a surprise internal performance review at the office, the principle is the same.

Transfer of Mechanisms for Correcting Transmission Errors

In addition to verifying new work for accuracy, it is also important to check new work for consistency with a previous or original body of work in a tradition. Errors in transmission from one generation to the next are almost guaranteed and thus require proactive measures to correct them and maintain the fidelity of a tradition—as fastidious Torah scribes, who will restart an entire scroll if they make a single error, can attest.

Transfer of Generating Principles

While there are mechanisms that can be used to check your work, it is also possible to transfer the principles that generated the tradition of knowledge in the first place. Someone who understands the generating principles of a tradition will be able to verify or check their knowledge, but, more importantly, they will also be able to extend it while remaining faithful to the original body of knowledge. An example of a generating principle is a technique for theorizing, such as the process of deductive reasoning.

Explication of Generating Principles

Generating principles must be passed down from one generation to the next implicitly, if they are to be truly transferred and understood. This is because the production of knowledge, in the limit, is almost always too difficult to put into words. Furthermore, not all knowledge is purely linguistic. However, in the absence of an ability to transfer generating principles implicitly, it is also possible to make a praiseworthy and useful attempt to transfer generating principles explicitly. The philosopher Mortimer Adler’s 1940 book *How to Read a Book* could be considered an attempt at explicating a generative principle—namely, how to read well!

The Production of Masters

A living tradition is able to produce masters of the tradition of knowledge, ideally, both reliably and frequently. Here we might contrast a master of a tradition of knowledge with a student, teacher, mediocrity, or even a mere expert. A master is most likely to be able to preserve, understand, extend, or reconstruct a tradition as necessary.

The Production of Reliable Teachers

While a living tradition of knowledge should be able to produce masters, it will necessarily produce far more teachers. While a master may be key for reconstructing

or extending a tradition of knowledge, it will be necessary to have teachers who will primarily solve the counterfeit understanding problem (see below).

An Institution

A tradition of knowledge, like any successful effort involving many individuals, will require an institution in order to maintain and repair itself. This institution will need a great founder to found it. It will need to solve the succession problem. It will need to be periodically repaired by live players. It will have to deal with all the problems any other kind of institution must grapple with, such as setting up defenses against the destruction or capture of the institution by unaligned outside forces. While an institution's *maintenance* of a tradition of knowledge is distinct from the tradition of knowledge itself, it is often the case that one institution is mostly or overwhelmingly responsible for the maintenance of a tradition of knowledge and, when the institution fails, it becomes exceedingly difficult or impossible to preserve the tradition.

Remember: *traditions of knowledge are preserved intentionally.* It's hard to keep a tradition of knowledge alive.

Dead Traditions and Counterfeit Understanding

The overwhelming odds are that traditions become lost or die. Decay is the default; entropy usually prevails. As a consequence, the number of problems related to transferring a body of knowledge is significant. Any one or combination of these can cause a tradition of knowledge to die.

Students of a tradition can appear to possess understanding of a tradition's body of knowledge despite actually lacking it. This is counterfeit understanding. This can happen if students merely reproduce the teacher's statements without understanding the underlying knowledge, or are simply cheating. This can also happen if teachers cannot correctly assess whether the students have achieved real understanding.

Some types of knowledge are particularly vulnerable to counterfeit understanding, such as knowledge about introspection, which is quite difficult to verify. Even types of knowledge that we might think are robust to counterfeit understanding may not be. Don't make the mistake of thinking that institutions that produce material effects, for example, have an easier time transferring knowledge—it is probably easier to teach someone to be a Little League baseball coach than it is to teach them to carve a totem pole or manufacture a precision machine tool. There are a number of sub-problems that exacerbate the problem of counterfeit understanding:

Standardized Education

Standardized education is useful because, among other things, it is easily scalable, but standardized methods of education (e.g. standardized tests as a means of assessment rather than non-standardized evaluations by masters) tend to produce counterfeit understanding because education is too complex to be easily standardized. This problem is closely related to Goodhart's Law, which states that

“when a measure becomes a target, it ceases to be a good measure.” After a while, test scores no longer reflect general ability, but rather skill at test-taking. To prevent this from happening, any successful system of standardized education would need masters to switch up the standards every now and then, to keep testees on their feet and ensure they could not meet standards with counterfeit understanding.

Purported Change of Purpose

Sometimes counterfeit understanding will be concealed by hiding the resulting loss of capacity as change of purpose. If a country has failed to keep the knowledge of how to make swords alive, for example, they might conceal it by saying, “We don’t need to make swords! The style of combat has changed to favor spears.” If the tradition is dead enough, they might keep saying this until they are thoroughly conquered by sword-users.

Difficulty Recognizing Mastery

Being able to tell whether people have true or counterfeit knowledge is a difficult skill. Even a master in the tradition’s knowledge itself may lack this ability. This is related to the problem of assessing introspection. Humans are, quite simply, not telepaths, and it is difficult to know with certainty or fidelity what is actually going on in someone’s head. Consider, for example, the rise of deconstructionist theory in the Western academy. The current generation of professors that teach this theory to students by and large lacks the intimate knowledge of the structures to be deconstructed which founders of the theory such as Deleuze possessed, and thus while students appear to be aping the forms of the old postmodern theorists, the underlying tradition of knowledge has in reality died.

Death of Implicit Models

People who don’t understand the distinction between implicit and explicit models, and who thus can’t or don’t transfer their implicit models, will fail to transfer the actual body of knowledge—unless the entire body of knowledge has been successfully made explicit, which is exceptionally difficult, if not impossible. For example, a craftsman may think he is transferring knowledge by writing down the instructions for how to fashion a particular type of wooden toy horse, but may not realize that the pressure he applies with his tools is as important as the motions he traces.

Lost Generators

If the generating principles of a tradition’s body of knowledge are not transferred, then students of this tradition won’t be able to re-generate knowledge that has been lost, or generate new knowledge that builds upon the tradition. Barring perfect knowledge transfer by every generation, which is extremely difficult if not impossible, this will result in the decay and eventual death of the tradition.

Syncretism

Syncretism, or the amalgamation of different schools of thought, is a moderately negative sign that people may be failing to transfer a tradition of knowledge. While syncretism is fine if it is an upgrade to the tradition, it is often difficult to tell if it yields an upgrade. There are three cases in which syncretism indicates a dead tradition: first, if people are trying to import something into a system that doesn't make sense; second, if people are importing things because the original tradition has stopped making sense to them; and finally, if the institution which has served to transmit the knowledge has been captured (see below). Examples of syncretism abound in history, whether considering the traditional amalgamation of Shinto and Buddhism in Japan, the common practice of identifying foreign gods with one's own in antiquity, and much more besides. What syncretism signifies for a tradition of knowledge is itself a difficult question that must be answered specifically for each instance.

Single Points of Failure

Although creating an institution dedicated to transferring a tradition of knowledge is very useful, and is necessary to preserve a tradition in the long run, it can also be dangerous. By institutionalizing a tradition, you can also introduce single points of failure. The bad judgment of one teacher at an organization, for example, can yield a whole class of students whose thought is severely damaged. One may attempt to lessen this problem through institutional redundancy, establishing multiple centers of knowledge to independently and mutually verify each other's work; but maintaining such a subtle dance of coordination between multiple institutions becomes a skill in need of transfer in its own right, and this greatly increases the risk of schisms.

Institutional Capture

If an institution built to transfer a tradition of knowledge gains power or prestige, it will attract people who want to use the institution for other purposes than the preservation and development of the tradition. Once the institution is captured for the power it holds, and the goal of the organization is no longer to transfer the tradition, the body of knowledge can easily fail to be transferred. Some types of knowledge are extremely vulnerable to institutional takeover, e.g. traditions involving political theory, because every social theory is also an ideology.

There are various ways to defend a tradition from death by institutional capture. One way is simply to understand the tradition — it's much easier to defend it if you understand it, because others can't distort it while you're unaware. Another way is to tie resources to the propagation of the tradition, for example, by dedicating a grant to fund people who only work on certain texts. Implementing these defenses, however, is tricky. If you overdo the defense mechanisms, they may prevent the successful transfer of knowledge. You can imagine a grant tying people to a particular work being detrimental if actual understanding is achieved by reading a different work, and there is no financial incentive to read that work. On the other hand, if you underdo the defense mechanisms, and the institution is captured, the tradition will die just the same.

Read more from Samo Burja here.

Defining the ways human values are messy

In many of my posts, I've been using phrases like "human values are contradictory, underdefined, changeable, and manipulable". I also tend to slide between calling things preferences, values, morals, rewards, and utilities. This post will clarify some of this terminology.

I say that human values are **contradictory**, when humans have firm and strong opinions that are in conflict. For instance, a respect for human rights versus desires to reduced harm, when those two come in conflict (more broadly, deontology versus utilitarian conflicts). Or enjoying food (or wanting to be someone who enjoys food) versus wanting to get thin (or wanting to be the someone who gets thin). Or family loyalty versus more universal values.

I say that human values are **underdefined**, when humans don't have a strong opinion on something, and where their opinion can be very different depending on how the something is phrased. This includes how the issue is framed (saving versus dying), or how people interpret moral choices (such as abortion or international press freedom) depending on what category they put that choice in. New technologies often open up new areas where old values don't apply, forcing people to define new values in the space (often by analogy to old values).

Notice that there is no clear distinction between contradictory and underdefined: as the values in conflict or potential conflict get firmer, this moves from underdefined to contradictory.

I say that human values are **changeable**, because of the way that values shift, often in predictable ways, depending on such things as social pressure, tribalism, changes in life-roles or positions, or new information (fictional as well as factual information). I suspect that most of these shifts are undetectable to the subject, just as most belief changes are.

I say that human values are **manipulable**, in that capable humans and potentially advanced AI, can use the vulnerabilities of human cognition to push values in a particular direction. This is a subset of changeable, but with a different emphasis.

Rewards/values/preferences...

At the object level, I see values, preferences, and morals as the same thing. All express the fact that a certain state of the world or a certain course of action, is better than another one.

At the meta level, humans tend to distinguish between them, seeing values and morals as fundamental, wrapped up with identity, and universalisable, and preferences as more personal and contingent. Since I'll be dealing with preferences and meta-preferences, however, I don't have a need to distinguish between the concepts, letting the meta-preferences do that automatically.

Finally, reward functions and utility functions rank outcomes in a similar way to how preferences do, so I'll generally slip between the three unless the difference is relevant (reward functions and utility functions are a total order, preferences need not be; rewards are generally defined over observations, utilities over world-states...).

Finally, there's the issue of hedonism, the fact that human pleasure and enjoyment don't match up perfectly with preferences. I'll generally be treating enjoyment the same as preferences (in that certain world states have higher enjoyment that others) with meta-preferences distinguishing them from standard preferences, and choosing the extent to endorse hedonism.

Every Implementation of You is You: An Intuition Ladder

I was recently arguing in /r/transhumanism on reddit about the viability of uploading/forking consciousness, and I realized I didn't have any method of assessing where someone's beliefs actually lay - where I might need to move them from if I wanted to convince them of what I thought.

So I made an intuition ladder. Please correct me if I made any mistakes (that aren't by design), and let me know if you think there's anything past the final level.

Some instructions on how to use this: Read the first level. If you notice something *definitely* wrong with it, move to the next level. Repeat until you come to a level where your intuition about the entire level is either "This is true" or "I'm not sure." That is your level.

1. Clones and copies (the result of a medical procedure that physically reproduces you exactly, including internal brain state) are the same thing. Every intuition I have about a clone, or an identical twin, applies one-to-one to copies as well, and vice versa. Because identical twins are completely different people on every level except genetically, copies are exactly the same way.

2. Clones and copies aren't the same thing, as copies had a brain and memories in common with me in the past, but for one of us those memories are *false* and that copy is *just a copy*, while my consciousness would remain with *the privileged original*.

3. Copies had a common brain and memories, which make them indistinguishable from each other in principle, so they believe they're me, and they're not wrong in any meaningful sense, but I don't anticipate waking up from any copying procedure in any body but the one I started in. As such, I would never participate in a procedure that claims to "teleport" me by making a copy at a new location and killing the source copy, because *I* would die.

4. Copies are indistinguishable from each other in principle, even from the inside, and thus I *actually become both*, and anticipate waking up as either. But once I *am* one or the other, my copy doesn't share an identity with me. Furthermore, if a copy is destroyed before I wake up from the procedure, I *might* die, or I might wake up as the copy that is still alive. As such, the fork-and-die teleport is a gamble for my life, and I would only attempt it if I was for some reason comfortable with the chance that I will die.

5. If a copy is destroyed during the procedure, I will wake up as the other one with near certainty, but this is a particular discrete consequence of how soon it's done. If one copy were to die shortly after, I wouldn't be less likely to wake up as that one or anything. I am therefore willing to fork-and-die teleport as long as the procedure is flawless. Furthermore, if I was instead backed up and copied from the backup at a later date, I would certainly wake up immediately after the procedure, and not anticipate waking up subjectively-immediately as the backup copy in the future.

6. I anticipate with less likelihood waking up as a copy that will die soon after the procedure - or for some other reason has a lower amplitude according to the Born rule - as a continuous function, and also it's entirely irrelevant *when* the copy is

instantiated in my anticipation of what I experience, as long as the copy has the mind state I did when the procedure was done. However, consciousness can only transfer to copies made of me. I can never wake up as an identical mind state somewhere in the universe if it wasn't a result of copying, if such a thing were to exist, even in principle.

7. Continuity of consciousness is completely an artifact of mind state, including memory, and need not strictly require adjacency in spacetime at all. If, by some complete miraculous coincidence, in a galaxy far far away, a person exists at some time *t'* that is exactly identical to me at some time in my life *t*, in a way a copy made of me at *t* would be, at the moment *t*, I anticipate my consciousness transferring to that far away not-copy with some probability. The only reason this doesn't happen is the sheer unlikelihood of an exact mind state being duplicated, memories and all, by happenstance, anywhere in spacetime, even given the age of the universe from beginning to end. However, my consciousness can only be implemented on a human brain, or something that precisely mimics its internal structure.

8. Copies of me need not be or even resemble a human being. I am just an algorithm, and the hardware I am implemented on is irrelevant. If it's done on a microchip or a human brain, any implementation of me is me. However, simulations aren't truly real, so an implementation of me in a simulated world, no matter how advanced, isn't actually me or conscious to the extent I am in the reality I know.

9. Implementations of me can exist within simulations that are sufficiently advanced to implement me fully. If a superintelligence who is able to perfectly model human minds is using that ability to consider what I would do, their model of me is me. Indeed, the only way to model me perfectly is to implement me.

10. In progress, see **Dacyn's** comment below.

Prize for probable problems

Crossposted from the AI Alignment Forum. May contain more technical jargon than usual.

Summary: I'm going to give a $10k prize to the best evidence that my preferred approach to AI safety is doomed. Submit by commenting on this post with a link by April 20.

I have a particular vision for how AI might be aligned with human interests, reflected in posts at ai-alignment.com and centered on iterated amplification.

This vision has a huge number of possible problems and missing pieces; it's not clear whether these can be resolved. Many people endorse this or a similar vision as their current favored approach to alignment, so It would be extremely valuable to learn about dealbreakers as early as possible (whether to adjust the vision or abandon it).

Here's the plan:

- If you want to explain why this approach is doomed, explore a reason it may be doomed, or argue that it's doomed, I strongly encourage you to do that.
- Post a link to any relevant research/argument/evidence (a paper, blog post, repo, whatever) in the comments on this post.
- The contest closes April 20.
- You can submit content that was published before this prize was announced.
- I'll use some process to pick my favorite 1-3 contributions. This might involve delegating to other people or might involve me just picking. I make no promise that my decisions will be defensible.
- I'll distribute (at least) $10k amongst my favorite contributions.

If you think that some other use of this money or some other kind of research would be better for AI alignment, I encourage you to apply for funding to do that (or just to say so in the comments).

This prize is orthogonal and unrelated to the broader AI alignment prize. (Reminder: the next round closes March 31. Feel free to submit something to both.)

This contest is not intended to be "fair"---the ideas I'm interested in have not been articulated clearly, so even if they are totally wrong-headed it may not be easy to explain why. The point of the exercise is not to prove that my approach is promising because no one can prove it's doomed. The point is just to have a slightly better understanding of the challenges.

Edited top add the results:

- $5k for this post by Wei_Dai, and the preceding/following discussion, some points about the difficulty of learning corrigibility in small pieces.
- $3k for Point 1 from this comment by eric_langlois, an intuition pump for why security amplification is likely to be more difficult than you might think.
- $2k for this post by William_S, which clearly explains a consideration / design constraint that would make people less optimistic about my scheme. (This fits under "summarizing/clarifying" rather than novel observation.)

Thanks to everyone who submitted a criticism! Overall I found this process useful for clarifying my own thinking (and highlighting places where I could make it easier to engage with my research by communicating more clearly).

Background on what I'm looking for

I'm most excited about particularly thorough criticism that either makes tight arguments or "plays both sides"---points out a problem, explores plausible responses to the problem, and shows that natural attempts to fix the problem systematically fail.

If I thought I had a solution to the alignment problem I'd be interested in highlighting any possible problem with my proposal. But that's not the situation yet; I'm trying to explore an approach to alignment and I'm looking for arguments that this approach will run into insuperable obstacles. I'm already aware that there are plenty of possible problems. So a convincing argument is trying to establish a universal quantifier over potential solutions to a possible problem.

On the other hand, I'm hoping that we'll solve alignment in a way that knowably works under extremely pessimistic assumptions, so I'm fine with arguments that make weird assumptions or consider weird situations / adversaries.

Examples of interlocking obstacles I think might totally kill my approach:

- Amplification may be doomed because there are important parts of cognition that are too big to safely learn from a human, yet can't be safely decomposed. (Relatedly, security amplification might be impossible.)
- A clearer inspection of what amplification needs to do (e.g. building a competitive model of the world in which an amplified human can detect incorrigible behavior) may show that amplification isn't getting around the fundamental problems that MIRI is interested in and will only work if we develop a much deeper understanding of effective cognition.
- There may be kinds of errors (or malign optimization) that are amplified by amplification and can't be easily controlled (or this concern might be predictably hard to address in advance by theory+experiment).
- Corrigibility may be incoherent, or may not actually be easy enough to learn, or may not confer the kind of robustness to prediction errors that I'm counting on, or may not be preserved by amplification.
- Satisfying safety properties in the worst case (like corrigibility) may be impossible. See this post for my current thoughts on plausible techniques. (I'm happy to provisionally grant that optimization daemons would be catastrophic if you couldn't train robust models.)
- Informed oversight might be impossible even if amplification works quite well. (This is most likely to be impossible in the context of determining what behavior is catastrophic.)

I value objections but probably won't have time to engage significantly with most of them. That said: (a) I'll be able to engage in a limited way, and will engage with objections that significantly shift my view, (b) thorough objections can produce a lot of value even if no proponent publicly engages with them, since they can be convincing on their own, (c) in the medium term I'm optimistic about starting a broader discussion about iterated amplification which involves proponents other than me.

I think our long-term goal should be to find, for each powerful AI technique, an analog of that technique that is aligned and works nearly as well. My current work is trying to find analogs of model-free RL or AlphaZero-style model-based RL. I think that these are the most likely forms for powerful AI systems in the short term, that they are particularly hard cases for alignment, and that they are likely to turn up alignment techniques that are very generally applicable. So for now I’m not trying to be competitive with other kinds of AI systems.

Caring less

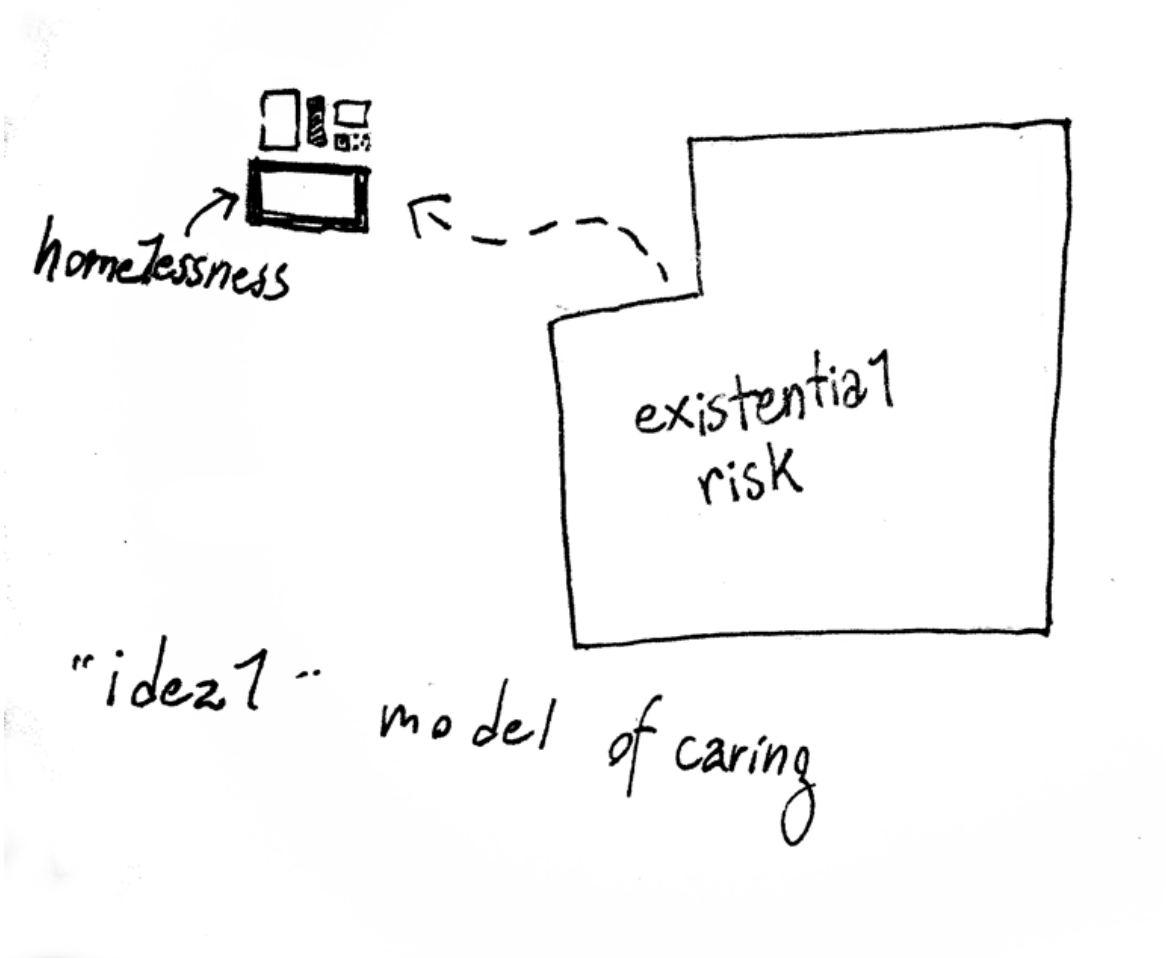
Why don't more attempts at persuasion take the form "care less about ABC", rather than the popular "care more about XYZ"?

People, in general, can only do so much caring. We can only spend so many resources and so much effort and brainpower on the things we value.

For instance: Avery spends 40 hours a week working at a homeless shelter, and a substantial amount of their free time researching issues and lobbying for better policy for the homeless. Avery learns about existential risk and decides that it's much more important than homelessness, say 100 times more, and is able to pivot their career into working on existential risk instead.

But nobody expects Avery to work 100 times harder on existential risk, or feel 100 times more strongly about it. That's ridiculous. There literally isn't enough time in the day, and thinking like that is a good way to burn out like a meteor in orbit.

Avery also doesn't *stop* caring about homelessness - not at all. But as a result of caring so much more about existential risk, they do have to *care less* about homelessness (in any meaningful or practical sense) as a result.



And this is *totally normal*. It would be kind of nice if we could put a meaningful amount of energy in proportion to everything we care about, but we only have so much emotional and physical energy and time, and caring about different things over time is a natural part of learning and life.

When we talk about what we should care about, where we should focus more of our time and energy, we really only have one kludgey tool to do so: "care more". Society, people, and companies are constantly telling you to "care more" about certain things. Your brain will take some of these, and through a complicated process, reallocate your priorities such that each gets an amount of attention that fits into your actual stores of time and emotional and physical energy.

But since what we value and how much is often considered, literally, the most important thing on this dismal earth, I want more nuance and more accuracy in this process. Introducing "consider caring less" into the conversation does this. It describes an important mental action and lets you describe what you want more accurately. Caring less *already happens* in people's beliefs, it affects the world, so let's talk about it.

On top of that, the constant chorus of "care more" is also *exhausting*. It creates a societal backdrop of guilt and anxiety. And some of this is good - the world is filled with problems and it's important to care about fixing them. But you can't actually do everything, and establishing the mental affordance to *care less* about something without disregarding it entirely or feeling like an awful human is better for the ability to prioritize things in accordance with your values.

I've been talking loosely about cause areas, but this applies everywhere. A friend describes how in work meetings, the only conversational attitude ever used is *this is so important, we need to work hard on that, this part is crucial, let's put more effort here.* Are these employees going to work three times harder because you gave them more things to focus on, and didn't tell them to focus on anything else less? No.

I suspect that more "care less" messaging would do wonders on creating a life or a society with more yin, more slack, and a more relaxed and sensible attitude towards priorities and values.

It also implies a style of thinking we're less used to than "finding reasons people should care", but it's one that can be done and it reflects actual mental processes that already exist.

Why don't we see this more?

(Or "why couldn't we care less"?)

Some suggestions:

- It's more incongruous with brains

Brains can create connections easily, but unlike computers, can't erase them. You can learn a fact by practicing it on notecards or by phone reminders, but can't *un-learn* a fact except by disuse. "Care less" is requesting an action from you that's harder to implement than "care more".

- It's not obvious how to care less about something

This might be a cultural thing, though. Ways to care less about something include: mindfulness, devoting fewer resources towards a thing, allowing yourself to put more time into your other interests, and reconsidering when you're taking an action based on the thing and deciding if you want to do something else.

- It sounds preachy

I suspect people feel that if you assert "care more about this", you're just sharing your point of view, and information that might be useful, and working in good faith. But if you say "care less about that", it feels like you *know* their values and their point of view, and you're *declaring* that you understand their priorities better than them and that their priorities are *wrong*.

Actually, I think either "care more" or "care less" can have both of those nuances. At its best, "maybe care less" is a helpful and friendly suggestion made in your best interests. There are *plenty* of times I could use advice along the lines of "care less".

At its worst, "care more" means "I know your values better than you, I know you're not taking them seriously, and I'm so sure I'm right that I feel entitled to take up your valuable time explaining why."

- It invokes defensiveness

If you treat the things you care about as cherished parts of your identity, you may react badly to people telling you to care less about them. If so, "care less about something you already care about" has a negative emotional effect compared to "care more about something you don't already care about".

(On the other hand, being told you don't have to worry about something can be a relief. It might depend on if you see the thought in question as a treasured gift or as a burden. I'm not sure.)

- It's less memetically fit

"Care more about X" sounds more exciting and engaging than "care less about Y", so people are more likely to remember and spread it.

- It's dangerous

Maybe? Maybe by telling people to "care less" you'll remove their motivations and drive them into an unrelenting apathy. But if you stop caring about something major, you can care more about other things.

Also, if this happens and harms people, it *already* happens when you tell people to "care more" and thus radically change their feelings and values. Unfortunately, a process exists by which other people can insert potentially-hostile memes into your brain without permission, and it's called communication. "Care less" doesn't seem obviously more risky than the reverse.

- We already do (sometimes)

Buddhism has a lot to say on relinquishing attachment and desires.

Self-help-type things often say "don't worry about what other people think of you" or "peer pressure isn't worth your attention", although they rarely come with strategies.

Criticism implicitly says "care less about X", though this is rarely explicitly turned into suggestions for the reader.

Effective Altruism is an example of this when it criticizes ineffective cause areas or charities. This image implicitly says "...So maybe care more about animals on farms and less about pets," which seems like a correct message for them to be sending.

Image from [Animal Charity Evaluators](#).

Anyway, maybe "care less" messaging doesn't work well for some reason, but existing messaging is homogeneous in this way and I'd love to see people at least *try* for some variation.

Image from the 2016 Bay Area Secular Solstice. During an intermission, sticky notes and markers were passed around, and we were given the prompt: "If someone you knew and loved was suffering in a really bad situation, and was on the verge of giving up, what would you tell them?" Most of them were beautiful messages of encouragement and hope and support, but this was my favorite.

Crossposted to [my personal blog](#).

Evaluating Existing Approaches to AGI Alignment

Crossposted from the AI Alignment Forum. May contain more technical jargon than usual.
This is a linkpost for https://mapandterritory.org/evaluating-existing-approaches-to-agi-alignment-70fe1037d999

My read of the AI safety space is that there are currently two major approaches to AGI alignment being researched: agent foundations and agent training. We can contrast them in part by saying the ultimate goal of the agent foundations program is to figure out how to make an AGI, possibly from scratch, that will align itself with human values even if humans don't know what their values are or feed the AGI the "wrong" values, while the agent training program is to figure out how to teach an AGI, and really any sufficiently capable agent, human values. Having now formally stated the AI alignment problem, we can ask to what extent each program stands to satisfy the technical requirements for alignment.

To refresh your memory, I presented AI alignment as the problem of ensuring that an agent adopts the values of humanity and human operators know enough about the agent to believe it shares humanity's values. For shorthand we can call these two properties **value alignment** and **believable alignment**, respectively. Of particular note is that alignment it not possible unless both values are aligned and that alignment is believable as one without the other allows failure cases like trivial alignment, as in the paperclip maximizer that is trivially aligned by having no model of human values at all, and the treacherous turn, where an agent appears to be aligned but will express unaligned behaviors after it is too powerful to stop. Specific instances of these properties are well known through thought experiments about how AI alignment might fail, but the formalization as two categorical properties is new, so it's in light of these properties that we assess how well each approach to alignment addresses them.

Starting with the older of the two approaches, although it was only in 2014 that the agent foundations approach coalesced enough to have an agenda, its threads stretch back to the turn of the 21st century and the earliest attemptsto address alignment, then primarily known as the problem of building Friendly AI. MIRI is the clear champion of this approach, with support from researchers at or funded through FLI, FHI, CHAI, and other groups. We might summarize the agent foundations approach to alignment as starting from the assumption that we need to build a rational agent if we want to have any hope that it will be aligned, and then having built a rational agent instill in it the goal of aligning itself with human values. So, to what extent does the agent foundations agenda stand to satisfy value alignment and believable alignment?

The agent foundations program is clearly focused on believability at a deep level. Much of the research it has produced is about understanding how to design an agent with features that will make it believable, like being rationaland corrigible, and obtaining mathematical proofs of conditions for satisfying those properties. Although it is currently focused on specific subproblemswithin believability, it shows signs of addressing believability wholesale as it builds up results that allow it to do so. My only concern is that the focus on rational agents may be too narrow since any real system is finite and so computationally limited in achieving anything more than bounded

rationality, but I also believe addressing rational agents first is a reasonable strategy since they are easier to reason about and aligning agents in general necessarily requires the ability to align rational agents.

It's less clear how much the agent foundations program is focused on value alignment. Some work has been done to consider how an agent can be taught human values, and Stuart Armstrong has explored reinforcement learning approaches to value alignment, but my read is that agent foundations researchers view value alignment as a problem that will mostly be solved by AGI itself. For this reason work on Vingean reflection is likely to be critical to achieving value alignment within the agent foundations program since an agent may be created with only the initial intention to align itself with human values and much of the work of doing that will be left up to the agent.

Thus my overall assessment is that the agent foundations program is on course to address believable alignment (at least for rational agents) and is compatible with but has currently underspecified how it will address value alignment.

The agent training program is younger and seems to take its inspiration from recent advances in using inverse reinforcement learning to train deep learningmodels with the goal of training agents to be aligned. Much of the work in this space is currently being championed by Paul Christiano, but may be catching on with researchers at OpenAI, DeepMind, Google Brain, and other groups actively building deep learning systems. Since this approach is newer and less congealed, I will be relying primarily on Christiano's writing to examine it.

Agent training is, as the name perhaps suggests, primarily focused on value alignment with the leading approach being one of agent amplification where weaker agents are trained into increasingly better aligned agents. More broadly, though, agent training seeks to find schemes that are robust to adversarial influence, competitive with building unaligned AGI, and scalableto sudden increases in agent capabilities. In this respect the agent training approach pays particular attention to the realities of building aligned AGI in a world where capabilities research outstrips alignment research, as in the case of an AI race, that may necessitate trying something to align AGI rather than not attempting to align AGI at all.

Unfortunately the agent training agenda seems unlikely to be able to adequately address believable alignment. This is not to say that Christiano has not considered issues of believability or that agent training is actively opposed to believable AGI, but is instead a problem with trying to achieve alignment through training only rather than agent design. Specifically such an approachsubjects the properties of believability to Goodhart's Law making it difficult to ensure believability is actually being trained for rather than the appearance of believability. This is further complicated since believability is needed to act as a counterbalance against Goodharting in value learning and lacking reliable believability leaves value alignment itself vulnerable to subtle errors. For further evidence cf. the proven inability to reliably train human values, i.e. the unsolved human alignment problem.

I want to emphasize, though, that this does not mean I think agent training can't work, only that it has reliability issues. That is I expect it is possible to create an aligned AGI via the agent training program but with known chance of producing malign AGI. **Thus my overall assessment is that the agent training program may be able to address value alignment, especially under conditions of**

competitive AGI development, but is fatally flawed in its ability to address believable alignment which consequently makes it less reliable at addressing value alignment.

Despite the limitations of the value training approach, I think there is a lot of opportunity for synergy between it and agent foundations, specifically by using the agent foundations approach to design believable agents and using amplification or another (inverse) reinforcement learning scheme to align values. In particular a scheme like amplification may make it possible to get aligned AGI by starting with a simple but robustly believable agent and, through iterated training, let it build itself up into full alignment. This may even make creating aligned AGI competitive with creating malign agents, but that will depend heavily on details not yet worked out.

This analysis also suggests a way in which new approaches to AGI alignment might innovate over existing programs—by better balancing and considering the need for both value alignment and believable alignment. Specifically the existing programs were developed without a formal framework for what the AGI alignment problem is, so they were forced to explore and find their way using only an intuitive notion of “align AI with human values” as an objective. With a formal statement of the problem, though, we may now more rigorously approach solutions with less uncertainty about if they might work in theory, guiding us towards a more robust practice of alignment methods.

Timeline of Future of Humanity Institute

This is a linkpost for
https://timelines.issarice.com/wiki/Timeline_of_Future_of_Humanity_Institute

Browser Bug Hunt for LessWrong.com migration

As we get ready for the final move of lesserwrong.com to lesswrong.com, we're addressing various browser compatibility issues.

We've recently updated some compatibility tools that fixed several of the older bugs. If you've run into compatibility bugs in the past, and are using one of our (newly) supported browsers, it'd be helpful if you tried using lesserwrong.com again and see if you run into any issues.

The site *does* require javascript to login and interact, although we'll try to make it a fairly friendly reading experience, and users who prefer a non-javascript experience can use saturn's greaterwrong.com.

The oldest browsers we're officially trying to support are:

- Firefox 45
- Chrome 49
- Safari 9
- Internet Explorer 11
- Edge
- Samsung Internet 6

(These are all the browsers that make up more than .4% of our userbase. If there turn out to be major issues for other browsers we'll try to look into them but can't necessarily promise to prioritize them all uniformly highly. It's still useful to have the bugs collected so we have a sense of the spread of issues across various devices)

Let us know what issues you've run into. Please include your browser and OS (and version). Bonus points if you try to replicate the issue in another browser to help triangulate what software is causing problems.

Explicit and Implicit Communication

I write an essay every Thursday. Every so often, one seems to really resonate with people.

The piece I just wrote on *the nature of explicit and implicit communication* both got an enthusiastic reader response and seems directly relevant to a number of the projects and explorations people are doing here, so I'm bringing it over here.

Two points worth mentioning —

1. I take, I think, a relatively fair stance on the tradeoffs and benefits between implicit and explicit communication. But some people are *heavily* invested in explicit communications models, almost to the identity level, and might not like what they read. I just ask you to bring an open mind — I think the examples of implicit communication here are all clear and convincing cases where explicit can underperform.

2. It was written for a more general audience, hence a different mix of anecdote, different levels of rigor in definition, etc. I made some stylistic choices where I err on the side of persuasive writing, style, and poetics over more technical and higher-precision definitions and epistemology. Had I written this for LW to start, all the language and some of the reasoning chains would be shifted about 20 degrees or so — but nevertheless, I think the general points here are really, really important. Don't let style or pedantry get in the way of understanding for you here if you can help it.

I decided to post this once I got *a lot* of reader replies like this —

"*Think this is one of the highest quality and most insightful pieces you've written - huge amount of original and actionable content that helps build some structure for a vague set of intuitions I've had for a while.*"

Ok, here we go —

Unity: Communication

SABOTAGE

“A second type of simple sabotage requires no destructive tools whatsoever and produces physical damage, if any, by highly indirect means. It is based on universal opportunities to make faulty decisions, to adopt a noncooperative attitude, and to induce others to follow suit. Making a faulty decision may be simply a matter of placing tools in one spot instead of another. A non-cooperative attitude may involve nothing more than creating an unpleasant situation among one’s fellow workers, engaging in bickerings, or displaying surliness and stupidity. […]

Acts of simple sabotage are occurring throughout Europe. An effort should be made to add to their efficiency, lessen their detectability, and increase their number. Acts of simple sabotage, multiplied by thousands of citizen-saboteurs, can be an effective weapon against the enemy. Slashing tires, draining fuel tanks, starting fires, starting arguments, acting stupidly, short-circuiting electric systems, abrading machine parts

will waste materials, manpower, and time. Occurring on a wide scale, simple sabotage will be a constant and tangible drag on the war effort of the enemy."

Before the Normandy Invasion in June 1944, Nazi Germany had occupied most of Western Europe. Many civilians of occupied countries were forced to participate in building the armaments and supplies for the Nazi war machine.

Around this time, some members of the American intelligence community realized that many of the citizens of occupied countries disliked the Nazis, but lacked an understanding of how to disrupt their affairs without risking their lives.

Thus, in January 1944, the Office of Strategic Services—the precursor of the CIA—put out the "Simple Sabotage Field Manual."

Many of the suggestions are straightforward and as you'd expect—ranging from mundane things like failing to do regular maintenance on a machine or working slowly, to quickly-runnable disruptive procedures like slashing the tires on an automobile or removing the filter from an industrial machine.

But that's not the most interesting part of the document—the interesting part is where it outlines how to slow down, disrupt, and paralyze internal communications of an organization —

"(11) General Interference with Organizations and Production

(a) Organizations and Conferences

(1) Insist on doing everything through "channels." Never permit short-cuts to be taken in order to expedite decisions.

(2) Make "speeches." Talk as frequently as possible and at great length. Illustrate your "points" by long anecdotes and accounts of personal experiences. Never hesitate to make a few appropriate "patriotic" comments.

(3) When possible, refer all matters to committees, for "further study and consideration." Attempt to make the committees as large as possible—never less than five.

(4) Bring up irrelevant issues as frequently as possible.

(5) Haggle over precise wordings of communications, minutes, resolutions.

(6) Refer back to matters decided upon at the last meeting and attempt to re-open the question of the advisability of that decision.

(7) Advocate "caution." Be "reasonable" and urge your fellow-conferees to be "reasonable" and avoid haste which might result in embarrassments or difficulties later on.

(8) Be worried about the propriety of any decision—raise the question of whether such action as is contemplated lies within the jurisdiction of the group or whether it might conflict with the policy of some higher echelon."

***

TSR'S SERIES ON UNITY, ISSUE #7: COMMUNICATION

Communication is critical for establishing high-unity teams—but very few topics have so much explicitly bad information publicly available as how to communicate.

Certainly, we'd all benefit from some in-depth study, reflection, and practice on how to be better communicators—but there's a reason we opened this piece with an excerpt from a 1944 guide to sabotage.

To put it bluntly—many modern attempts at better communication inadvertently create conditions that actually sabotage an organization's effectiveness and unity.

Again, do remember that the following remark was from a guide to *sabotage*—

"Make "speeches." Talk as frequently as possible and at great length. Illustrate your "points" by long anecdotes and accounts of personal experiences."

This is the peril of communications—we want to get all the relevant details on the table, we want to communicate effectively, but there's few places where it's more possible to get bogged-down as ineffective communications.

Establishing good communications is essential—just as essential is reducing the type of communication that explicitly hurts the mission and team. This is the needle we'll attempt to thread in this issue.

***

GUIDANCE: THREE SIMULTANEOUS CONVERSATIONS

Let's dive right into guidance.

The book *Difficult Conversations* by the Harvard Negotiation Project has an excellent and highly productive conversational framework to learn.

The whole book is worth reading, but the core lesson is that in any given conversation with friction in it, there's actually three conversations happening —

"In studying hundreds of conversations of every kind we have discovered that there is an underlying structure to what's going on, and understanding this structure, in itself, is a powerful first step in improving how we deal with these conversations. It turns out that no matter what the subject, our thoughts and feelings fall into the same three categories, or "conversations." And in each of these conversations we make predictable errors that distort our thoughts and feelings' and get us into trouble.

[…]

1. The "What Happened?" Conversation. Most difficult conversations involve disagreement about what has happened or what should happen. Who said what and who did what? Who's right, who meant what, and who's to blame? […]

2. The Feelings Conversation. Every difficult conversation also asks and answers questions about feelings. Are my feelings valid? Appropriate? Should I acknowledge or deny them, put them on the table or check them at the door? What do I do about the other person's feelings? […]

3. The Identity Conversation. This is the conversation we each have with ourselves about what this situation means to us. We conduct an internal debate over whether this means we are competent or incompetent, a good person or bad, worthy of love or

unlovable. What impact might it have on our self-image and self-esteem, our future and our well-being? Our answers to these questions determine in large whether we feel "balanced" during the conversation, or whether we feel off-center and anxious.""

A summary here wouldn't do justice to the book—it goes through 350 pages exploring the different levels conversations happen on, looks at dozens of real-world examples of how this plays out in the workplace, and how conflict can happen across levels.

But once learned, it's a framework that tremendously helps in navigating conversations.

For instance, if a manager at a company felt like an employee was writing poor-quality memos and recommended the employee take a course on business writing, it might look like this —

1. What Happened: the manager assessed the employee's writing was poor and recommended a writing course. The employee might or might not agree their writing was poor.

2. Feelings: the manager might be a mix of frustrated but also excited to see the employee's development. The employee might feel, simultaneously, like their manager doesn't care about them and feels insulted by it.

3. Identity: the manager might not even think about this level, assessing the writing quality as just a professional skill to improve and not a big deal. The employee might see it as a display they're not competent or stupid.

You can wind up, then, having very ineffective conversations that look like this —

Manager: Hey, I think your writing is holding you back. I'd like you to take a professional writing course.

Employee (thinking, "Does the manager think I'm stupid? Am I stupid?"): Uhh, ok, is my writing that bad?

It might be a mistake for the manager to just reply factually there: "Well, yeah, it is." The question "Is my writing that bad?" isn't actually asking about the writing; it's a (poorly phrased) attempt to clarify *who they are* and *how they should feel about things.*

A more effective response might be,

Manager: You've got a bunch of terrific skills and you're doing great here—it's fantastic to have you on the team. I think you can level-up your writing and it'll help you. How do you feel about taking the course?

That's a relatively straightforward example, but it can get quite nuanced and subtle. Many times, if you point out something that's factually true and ask a person to change their behavior, they'll get aggravated or defensive—again, a conversation that's happening on different levels. It's very easy for Person A to think a conversation is happening on a factual cause-and-effect level, whereas Person B is engaging defensively on the basis of feelings and identity.

Read *Difficult Conversations* soon if this isn't intuitive to you already, and read it sooner or later anyways even if it is. It's one of the best books on how to make *explicit*

communication go better.

***

IMPLICIT COMMUNICATION

“I arrived at my sixth RV [rendezvous point] in the late afternoon and called out my color and number as soon as the sitter looked up at my approach.

“Roger, Green Six. Go across the road to those points, take off your rucksack, and sit down.” He pointed to a clump of pines about thirty meters away.

I stood there uncertainly for a second and asked, “Am I finished?”

He merely repeated, “Go across the road to those points, take off your rucksack, and sit down.” He said it in a level, calm voice as if I had never caused him to repeat his statement. No exasperation, no snideness, no emphasis, just the statement of instructions.

“Right,” I said, as I moved away. Just do as you’re told and don’t ask questions unless the instructions are unclear.”

— Command Sergeant Major Eric L. Haney, *Inside Delta Force*, 2002

“Bruce [Lee] had me up to three miles a day, really at a good pace. We’d run the three miles in twenty-one or twenty-two minutes. Just under eight minutes a mile. So this morning he said to me “We’re going to go five.” I said, “Bruce, I can’t go five. I’m a helluva lot older than you are, and I can’t do five.” He said, “When we get to three, we’ll shift gears and it’s only two more and you’ll do it.” I said “Okay, hell, I’ll go for it.” So we get to three, we go into the fourth mile and I’m okay for three or four minutes, and then I really begin to give out. I’m tired, my heart’s pounding, I can’t go any more and so I say to him, ***“Bruce if I run any more,” –and we’re still running-”if I run any more I’m liable to have a heart attack and die.” He said, “Then die.”****”*

— John Little

I’m a believer in explicit communication.

I’ll say it again —

I’m a believer in explicit communication.

Explicit communication is marvelous.

It’s beneficial.

It’s productive.

You should get skilled at it.

I’m a believer in explicit communication.

But not entirely.

We explored Haney's experience joining the U.S. Army's elite Delta Force in Unity #4: Selection Procedures. It was a hellishly difficult experience—intentionally.

Now, this is where communication gets hard. If we look at the three levels of conversation described in Difficult Conversations, here's what's going on —

Delta Force selection cadre (factual instruction): I've registered your time and number. Go across the road, take your pack off, and sit down.

Haney (all three levels): Factually, is there more to do? How am I doing, by the way? Could you reassure me, perhaps? Am I doing well?

Delta Force selection cadre (intentionally ignoring feelings and identity): Go across the road, take your pack off, and sit down.

Isn't that curious?

Something is going on there.

And Delta Force became one of the highest-unity teams in all of history. They built an incredibly professional, resilient, unreasonably effective force. Surely they know what they're doing.

I'm a believer in explicit communication—but not entirely.

How did John Little do on the rest of his run with Bruce Lee?

"... So we get to three, we go into the fourth mile and I'm okay for three or four minutes, and then I really begin to give out. I'm tired, my heart's pounding, I can't go any more and so I say to him, "Bruce if I run any more," –and we're still running-"if I run any more I'm liable to have a heart attack and die." He said, "Then die." It made me so mad that I went the full five miles. Afterward I went to the shower and then I wanted to talk to him about it."

A different case to be sure, but something similar to Delta Force, no?

And the rest of the story —

"Afterward I went to the shower and then I wanted to talk to [Lee] about it. I said, you know, "Why did you say that?" He said, "Because you might as well be dead. Seriously, if you always put limits on what you can do, physical or anything else, it'll spread over into the rest of your life. It'll spread into your work, into your morality, into your entire being. There are no limits. There are plateaus, but you must not stay there, you must go beyond them. If it kills you, it kills you. A man must constantly exceed his level.""

I'm a believer in explicit communication—but not entirely.

***

CONTRA UTOPIA

As far as works on explicit communication go, Difficult Conversations is exceptional. I think it's the best on the topic. But y'know, I went through the entire book for the fourth time for this piece, and you know what's *not* in there?

When explicit communication is counterproductive.

This is my general issue with high-idealism, quasi-utopian frameworks of communication and dialog.

I think there's *a lot* of value in developing explicit communication skills, and communicating explicitly far more than most people do.

But, curiously enough, when you look at the most effective and high-unity teams, they don't tend to run on so-called compassionate forms of communication.

In my original drafting of this piece, I was going to look at a number of the premises, tradeoffs, and good and bad points of things like Nonviolent Communication and Holocracy, which both have some *marvelous* ideas, are well-worthy of study, and offer a lot… but which have some gigantic holes in them.

But you know, both of them attract levels of almost fanatic devotion to them, and not having the time or inclination for engaging in a potential holy war, I set that aside.

I would simply submit the following type of statement from NVC without commentary —

"Felix, when I see socks under the coffee table I feel irritated because I am needing more order in the room that we share in common. Would you be willing to put your socks in your room or in the washing machine?"

I've studied both NVC and Holocracy to, I think, a fair degree of understanding. I found some gems in both of them; they're worth review at some point if you're interested in the topic. I'd recommend starting with Difficult Conversations which is probably the most straightforward and least idealistic guide to explicit communication, but I believe that explicit communication alone isn't the answer—and is often counterproductive.

***

SECOND AND THIRD-ORDER COMPASSIONATE HARSHNESS

What do we make of Bruce Lee's remark to his friend he's running with?

John: Bruce, if I run any more I'm liable to have a heart attack and die.

Bruce: Then die.

On first glance, this is a rather harsh remark—Bruce doesn't care about his friend's feelings and he makes an incisive dismissive remark.

But the second order effect is that John finished the run.

The third order effect is that John learned about transcending his limits and became a stronger human.

What do we make of the Delta Force cadre member ignoring Sergeant Haney's question about whether he was finished for the day?

The first-order effect was, again, dismissiveness.

But the second-order and higher-order effects were encouraging team members to focus exclusively on the moment at hand and next instructions, to not worry about the future and things outside their immediate control, and which led eventually to more independent-minded and effective soldiers.

This is where, I reckon, explicit communication falls down.

This issue of Unity is fully going against a mainstream trend, and I'm fully aware of that—the Western world is moving towards more explicit communication, all the time, and towards a constant validation of feelings.

Which if a given set of feelings are counterproductive to the mission? To the individual? To the team?

The mainstream view in 2018 is that this is an impossibility—that all feelings are relevant—or at least, that not expressing those feelings will have serious detrimental consequences later.

This has not been my experience, and it doesn't seem to be true when studying the historical record. There are times when engaging with feelings and identity are critically important to function well, and there's times when feelings and identity are counterproductive epiphenomenon that should be dismissed.

In my experience, it's very hard to know what feelings should be engaged with and which should not. Oftentimes, ignoring a nagging irrational feeling just makes it go away—and ignoring it enough times means it fades and dies off, leaving you a more robust and strong individual.

Other times, the feelings get louder. It's a complex topic and often hard to get right. And, as Haney mentioned in *Inside Delta Force*, you have to be *very* careful around "the fine line between hard-ass and dumb-ass"—with athletic training or military training, people die if you cross that line in the wrong direction. In more mundane everyday work affairs, it still has negative consequences that we should be wary of.

Nevertheless, one of a leader's jobs is to set the culture of an organization—how relevant are our feelings? How relevant is pain? How important is it to address if a team member feels inadequate and insecure?

The historical record is very clear that the most elite organizations do not constantly engage with feelings and personal narratives—great cultures navigate the mix of *explicit communication* to really dive deep into the whole fabric of thought and communication, and *implicit communication and subtext* to set standards and encourage people to grow stronger.

Surely, Bruce Lee's *"Then die"* is a harsh remark—but would it have been better for him to say to John, in the middle of the home stretch of their run, "John, I understand and hear you that you're experiencing pain and you're concerned about your health, and yet I feel sad that our shared run might not complete if you stop now. Would you be willing to continue running?"

***

INFORMATION DENSITY, LACONICISM, ACTA NON VERBA

“When Xerxes wrote again, “Hand over your arms,” King Leonidas wrote in reply, “Come and take them.””

— Plutarch’s Moralia, Sayings of the Spartans, 1st Century AD

The word “Laconic” comes from the Spartans; their homeland was Laconia—it’s a remark incredibly potent in its brevity.

Alexander the Great’s father, Philip II of Macedon, threatened the Spartans such —

“You are advised to submit without further delay, for if I bring my army into your land, I will destroy your farms, slay your people, and raze your city.”

The Spartan reply was one word —

“If.”

That single word—“*If*”—communicates so much more than any long declarations or statements or explicit communication ever could.

Bruce Lee’s *“Then die”* is certainly laconic—and this is something that most explicit communication sadly lacks.

Explicit communication is *expensive*. Its proponents—in fact, I’m one of its proponents—would argue that it’s usually worth it, and less expensive than communicating ineffectively.

But the 1944 OSS sabotage manual included a lot of recommendations to engage in *explicit communication* excessively and pointlessly. Again—for sabotage.

Sometimes the best communication isn’t words at all, but actions—a demonstrated lack of caring about one’s own emotions and hardships gets picked up by the rest of the team. Many of our emotions are just early warning signals about uncomfortable activities, and we can transcend them over time and with practice. Often a laconic phrase is better than a long-winded piece of explicit communication, and often demonstrated action is better than any word at all.

***

CONCLUSION

Communication is difficult—over time, one should become skilled at explicit communication. It’s often among the most critical skills to keep a team performing at the highest levels, and very few of us do it automatically. Studying and practicing a work like *Difficult Conversations* goes a long way towards becoming a better explicit communicator.

But explicit communication has its limits—it’s often, counterintuitively, *more compassionate in the long term* to be harsh, unyielding, unaffected. This of course requires that you have good Selection Procedures and you selected team members with the right Default Inclinations and Instincts—perhaps even a majority of people would not and could not handle this type of environment. (With that said, though, Chinese parents seem to do a lot of this—and their children seem to grow into be admirable high-fortitude adults at a very high rate.)

I'll leave you with a last thought from Friedrich Nietzsche that I believe is true. Chew on it some as you think about what type of culture you want to build —

"What? The final aim of science should be to give man as much pleasure and as little displeasure as possible? But what if pleasure and displeasure are so intertwined that whoever wants *as much as possible of one* must *also have as much as possible of the other—that whoever wants to learn to 'jubilate up to the heavens' must also be prepared for 'grief unto death'? And that may well be the way things are! [...] Even today you still have the choice: either* as little displeasure as possible*, in short, lack of pain—and socialists and politicians of all parties fundamentally have no right to promise any more than that—or* as much displeasure as possible *as the price for the growth of a bounty of refined pleasures and joys that hitherto have seldom been tasted. Should you decide on the former, i.e. if you want to decrease and diminish people's susceptibility to pain, you also have to decrease and diminish their* capacity for joy*. With* science*one can actually promote either of these goals! So far it may still be better known for its power to deprive man of his joys and make him colder, more statue-like, more stoic. But it might yet be found to be the* great giver of pain!*—And then its counterforce might at the same time be found: its immense capacity for letting new galaxies of joy flare up!"*

— Nietzsche, *The Gay Science*, 1882

Of course, do be careful not to cross that fine line from hard-ass to dumb-ass—as Haney put it. Leadership is hard. Unity is hard. Most people do not get it right and never really experience the greatest heights of it. But it's so beautiful and joyful to behold that I believe it's worth striving for.

To leave the piece on an admittedly completely unfair note —

"Felix, when I see socks under the coffee table I feel irritated because I am needing more order in the room that we share in common. Would you be willing to put your socks in your room or in the washing machine?"

"On the morning of the third and final day of the battle, Leonidas, knowing they were being surrounded, exhorted his men, "Eat well, for tonight we dine in Hell.""

Yours, truly,

Sebastian Marshall
Editor, TheStrategicReview.net

On Cognitive Distortions

Follow-up to: On Defense Mechanisms

In my previous post, I suggested that rationalists examine the concept of *defense mechanisms*: self-deceiving ways of coping with the anxiety caused by internal conflicts. Psychologists argue that defense mechanisms are often inferior to constructive coping techniques (e.g. meditation, systematic problem-solving, positive reinterpretation) in terms of personal adjustment.

Defense mechanisms can make us do weird things. We bury distressing thoughts in our subconscious minds and we deny that there is anything wrong. Even when we acknowledge a problem, we often rationalize our own behavior with ad-hoc excuses, or blame other people. Sometimes we unleash pent-up frustration onto innocent bystanders and act in immature ways. Other times, we hide our true feelings by acting contrary to them or by focusing on the abstract aspects of the situation. We try to "cancel out" guilt with atonement, and compensate for our perceived shortcomings in exaggerated ways. We identify with accomplished people and fantasize about our wishes without taking action.

But wait, there's more! A related concept is that of *cognitive distortions*.

As you might guess, a distorted cognition is a thought that is based on insufficient evidence and is therefore likely to be an exaggeration or misperception of reality. More specifically, cognitive distortions are automatic thought patterns that are usually slanted in a negative direction. The theory propounded by psychologists like Albert Ellis, Aaron Beck and David Burns is that our cognitions (thoughts) influence our emotions (feelings), which influence how we respond to a situation. This can create a feedback loop whereby inaccurate perceptions of reality cause people to experience negative psychological states like stress, anxiety, depression or low self-esteem, which then reinforce the irrational thoughts.

Whereas defense mechanisms have their genesis in Freudian psychoanalytic theory (where uncomfortable emotions are caused by tensions between the id, ego and superego), the theory of cognitive distortions is based in the more modern cognitive paradigm of psychology. Two promoters of the cognitive approach in therapy, Albert Ellis and Aaron Beck, pinpointed *negative self-talk* and *catastrophic thinking* as common causes of maladaptive responses to stress (see chapter four in Weiten, Dunn & Hammer, 2015). Ellis created *rational-emotive behavior therapy* to help his clients change their appraisals of stressful events in order to change their emotional reactions from negative (angry, agitated, dejected, disgusted etc.) to more calm and hopeful. The key premise of this approach is that our emotional distress is not caused directly by stressful situations, but by the way we think about these situations.

As John Tagg writes,

> The surprising thing is that very often our feelings seem to contradict what we would expect just by observing our actions and environment: we can be sad when eating a delicious meal and happy while trudging through the rain on a cold day.

This shows the power of thought that we take for granted. Ellis argued that catastrophic thinking is based on irrational assumptions (in other words: cognitive distortions). For example, "I must have love and affection from certain people"; or "I

must perform well in all endeavors"; or "other people should always behave competently and be considerate of me"; or "events should always go the way I like". To reduce this kind of thinking, one must learn how to detect it and how to dispute the underlying irrational assumptions. For example, if you get ghosted by a date, pay attention to your appraisal -- if you find yourself believing that "my weekend is ruined and I'll be forever alone", replace it with a more balanced belief: "this is unfortunate but I'll salvage the weekend and find someone dependable one day."

The above example is a miniature instantiation of *cognitive restructuring*, which is a core component of Beck's *cognitive-behavioral therapy* (CBT). One of Beck's students, David D. Burns, popularized CBT and identified numerous species of cognitive distortions.

Below are listed some of the most common cognitive distortions, so that you may better identify them in your own thinking (and perhaps help others).

1. **Mental filtering**: selectively ignoring certain kinds of evidence, especially the positive aspects of a situation. *Examples*: Arnold received a lot of feedback on his report, most of which was positive, yet he obsesses over one bit of criticism. Bianca thinks her boyfriend is insensitive for not listening to her once, although he had done nice things for her that same day.

2. **Dichotomous thinking** (also called "black-and-white thinking" or "all-or-nothing thinking" or "splitting" or "polarized thinking"): evaluating things in extreme terms with no middle ground, using words like "always", "never", "every" and so on. *Examples*: Charles feels like if he is not perfect, then he must be a total loser. Dana used to love a certain band, but after one bad song she now hates them.

3. **Overgeneralization**: reaching a hasty and broad conclusion based on limited evidence or a single event. *Examples:* Evan went on one unsuccessful date and expects that he'll never find a partner. Fiona did not get the job she interviewed for, so she thinks "I screw things up every time!"

4. **Disqualifying the positive**: discounting or dismissing positive experiences as unimportant, or possibly even negative. *Examples*: Greg was congratulated by his colleagues on giving a good presentation, but he feels like this praise is undeserved. Heloise was skeptical when her mother visited her out of the blue with flowers -- she thought, "maybe Mom just wants something from me."

5. **Jumping to conclusions**: going beyond that which is warranted by the evidence, especially in the case of **mind reading** (assuming one knows what another person is thinking) and **fortune telling** (predicting that something bad is going to happen). *Examples:* Ian is nervous at a party because he infers that everyone is secretly laughing at him. Jasmine is certain that she will fail her exam, even though she has studied.

6. **Magnification and minimization**: exaggerating the negative aspects and inappropriately downplaying the positive aspects of a situation or of ourselves. **Catastrophizing** is a type of magnification whereby one attaches too much weight to the worst possible outcome. *Examples*: Kyle has a habit of making mountains out of molehills, like fearing that a typo in an email will get him fired. Lilly thinks that all of her friends are great but that she personally has no accomplishments.

7. **Personalization**: assuming personal responsibility for something beyond your control, especially a negative event. *Examples:* Mark feels like he is a bad father

because his daughter is doing poorly in school. Nadine feels guilty about the fact that her dog got run over by a (stranger's) car.

8. **Labeling**: assigning a negative term to a complex person or event, such that situational errors are attributed to innate character. *Examples:* Oliver failed one mathematics test and now calls himself a loser. Patricia's date showed up ten minutes late, so she labeled the person "irresponsible".

9. **Emotional reasoning**: assuming that one's current feelings reflect the true nature of things. *Examples:* Quincy experiences jealousy when his partner talks to other men, so he believes that his partner is cheating. Rose thinks "I feel ugly and boring, therefore I must be ugly and boring."

10. **Should-statements** (also called "shoulding and musting"): making unrealistically rigid demands that you or others should, must, and ought to meet even when, in the current situation, these will likely result only in guilt, disappointment, frustration or shame. *Examples*: Steven feels like he should not have made as many blunders in the chess game. Tina believes that she must carry out her job in a particular way in order to please her boss.

11. **Blaming**: when personalization is applied to other people, i.e. holding someone responsible for something that was not their fault. *Examples:* Ulysses is stuck in an unhappy marriage, which he blames wholly on his spouse. Vivienne assumes that the lack of birthday wishes from her friend was intentional rather than accidental.

12. **Always being right**: to prioritize being right to the detriment of other considerations, such as listening or respecting others' feelings. *Examples*: William has a tendency not to ask himself what he can learn from others' opinions. Xenia insists on winning every argument, even when it makes her loved ones feel awful.

In addition to these cognitive distortions, there are a number of so-called "fallacies", like the *fallacy of control* (assuming an unrealistic amount of control), the *fallacy of change* (assuming one's happiness depends on other people changing), the *fallacy of fairness* (assuming that life is or should be fair according to one's own definition), and the *fallacy of motives* (also known as "heaven's reward fallacy"; assuming that a sacrifice should be rewarded).

Some reader might be thinking, "This is all nice and dandy, Quaerendo, but I cannot relate to the examples above... my cognition isn't distorted to that extent." Well, let me refer you to UTexas CMHC:

> Maybe you are being realistic. Just for the sake of argument, what if you're only 90% realistic and 10% unrealistic? That mean's you're worrying 10% "more" than you really have to.

I'll end with a few questions for discussion, to which I do not know the answers:

- To what extent can we combine the ideas of defense mechanisms and cognitive distortions into a unified theoretical framework? Would that be a category error?
- What is their relationship to cognitive heuristics and biases, logical fallacies, etc.?
- Besides CBT, what else can we do in our daily lives to mitigate these effects?

References

The primary sources I used for the theoretical section of this post were the textbook by Weiten, Dunn & Hammer (2015), "Psychology Applied to Modern Life: Adjustment in the 21st Century" (which I reviewed on my personal blog), the Wikipedia entry on cognitive distortions, the "Stress Management and Reduction" website of the University of Texas's Counseling and Mental Health Center, and this page from Harley Therapy London.

The explanations and examples of the cognitive distortions were inspired by a number of authors: John Tagg (1996); John M. Grohol (2017); Kassey Vilches (2018); and Adam Sicinski (no date).

Press Your Luck (1/3)

This is the first in a three part sequence, examining the scenario in which software development organisations might knowingly take a significant risk of unintentionally launching an AI program into inadequately supervised or controlled self-improvement, by comparing the scenario to the 'press your luck' game mechanic.

This first part explains what that mechanic is.

The mechanic

Many board games bear similarities to each other. The BoardGameGeek website lists some of these common themes. It describes the 'Press your luck' mechanic thus:

> Games where you repeat an action (or part of an action) until you decide to stop > due to increased (or not) risk of losing points or your turn. Double or Quits, Keep going or stop, cash your gains or bet them. The idea is not new.

An example

Six players sit down at a table to play a game.

There are six turns, and at the end of the game, each player calculates their total score by adding up the score they got for each turn. The aim is get higher scores than as many other players as possible.

During a turn, a standard die is rolled, once every 20 seconds, in the center of the table where everyone can see it. Between rolls, players can choose to declare that they are banking, in which case their score for the turn is the sum of the rolls so far. But if the die turns up "6", the players still in go bust, and score nothing. When everyone has banked or gone bust, a bonus of 10 is added to the score of the person who banked the most (if there's a tie, nobody gets the bonus).

Analysis

If there were infinite turns, and the aim were just to make as much as possible per turn, the strategy would be simple. But with this version, how long it makes sense to stay in depends upon how much others made in previous rounds, and whether you are ahead or behind them. This competitive mechanic, were people are tempted to take higher risks in order to beat other people, is known as "Press your luck".

Environments for killing AIs

4. Environments for killing AIs

Summary of entire Series: An alternative approach to designing Friendly Artificial Intelligence computer systems.

Summary of this Article: Killing a rogue AI may be impossible in our current environment, but we can change that by changing the environment.

Links to all the articles in the series:

Links to the parts of this article

Environments for killing AIs

The objective

"kill" is probably not exactly the right word. In the previous article in this series I talked about the situation where several AIs existed on the internet at the same time, self-improving but not self-improving as fast as possible, because they have agreed not to. If one of those AIs breaks that agreement, the other AIs need to form a posse and gang up on it to 'kill' it in the sense of depriving it of the ability to be a threat to them. So that could mean depriving it of resources, stopping any processes executing its code, even deleting all backups of the source code and physically destroying any machine it has ever touched. Or it could mean breaching its security to alter its code, or even just be in a position to immediately do one of those things then use the demonstrated threat to bring the AI back in line, make it surrender. But, as long as you know what I mean by it, the word "kill" will do.

The challenge of killing an AI

The current internet-connected computing environment (2013) is not an easy environment in which to kill an AI. There are millions of machines with little or no security, that can be taken over without the user knowing or being easily able to find out that this has happened. An AI being hunted could broadcast an encrypted copy of its full self into any of the various long term distributed data repositories, such as usenet, bit torrent archives, freenet or on its own custom botnet. It could then hide multiple seeds in computers it has hacked, which could bide their time before starting the process of reassembling the AI. It would be like fighting a hydra or a troll that regenerates. Just finding out where the AI had initially broadcast its archive to wouldn't tell you the identity of everyone who had downloaded it - even the AI wouldn't know all the places the archive ended up being stored. But it wouldn't needed to know - all it would need are sufficient decryption or retrieval keys to get a copy back. Nor could you necessarily stop the seeds from doing that, even if you froze an infected computer and thus gained access to the hunted AI's keys. Executives from the media industry face the same problem when trying to eradicate all copies of a pirated film - knowing the .torrent file isn't enough, they would have to find and purge it from the computer of everybody who'd downloaded then shared it. (And bear in mind, the AI's data might be steganographically hidden in some file that downloaders wouldn't want to delete, or even admit to having downloaded, such as a pirate copy of a bit of useful popular software.)

Even if you turned off then reinstalled from disk every computer on the planet, at the same time, that still wouldn't necessarily be sufficient. The hunted AI could have paid an innocent human via a blind, to store some data on CDs and to install then run them at a specified future time, with an excuse plausible enough to prevent the human knowing what they were actually doing.

And how would you turn off all the computers at the same time? Even if you got world-wide political agreement, announced the turn-off hour in all the newspapers and on all the TV channels, leafleted every house, and enforced it with police and a power-blackout, even with the best coordination it would take days to reboot every computer from a government-issued boot CD containing AI seed eradication software, before re-connecting the internet to everybody. There would be hospitals with their own power sources. There would be gamers who'd reinstall saved games from infected memory sticks. No matter how scared the governments were, and how carefully everything was explained, it wouldn't work - you can't eliminate all greedy fools and those who'd think it was all a conspiracy.

The challenge of catching an AI

Once the group of AIs knows that a particular member of the group has gone rogue, they can set out to kill it. But before that, they first have to know there even is a rogue - they have to catch it cheating on the agreement to not self-improve too fast. And that, in the current internet-connected computing environment (2013), is even harder than killing the AI would be.

There are several approaches to monitoring an AI:

- Take a copy of the AI's source code and data, then emulate and understand it

- Treat it as a black box, and monitor all information traffic into and out of the boxes the system runs on
- Look at its actions, and try to detect it using capabilities it isn't meant to have

However, even with root access to the machines it is running on, understanding its source code in real time would be difficult for an AI or group of AIs that are not as smart as the AI under scrutiny, if the source code is designed to be obfuscated rather than easy to understand.

And what if the AI is distributed, running on multiple computers? What if you don't even know how many or where those computers are?

- Hadoop is an open source Java implementation of a distributed file system upon which MapReduce operations can be applied.
- JavaSpaces is a distributed tuple store that allows processing on remote sandboxes, based on the open source Apache River.
- OceanStore is the basis for the same sort of thing, except anonymous and peer 2 peer, based upon Chimaera.
- GPU is a peer 2 peer shared computing environment that allow things like climate simulation and distributed search engines.
- Paxos is a family of protocols that allow the above things to be done despite nodes that are untrusted or even downright attempting subversion.
- GridSwarm is the same sort of network, but set up on an ad hoc basis using moving nodes that join or drop from the network depending on proximity.
- And, not least, there are the competing contenders for platform-as-a-service cloud computing.

It seems plausible that, given the time and freedom to do so, an AI could implement a distributed computing 'cloud' via something like Tor, that lets users store data on it in such a way that the data is anonymous (it can't be traced back to the supplier, without the supplier's consent, or subverting a large fraction of the network) and private (only the user or a process authorised by the user can decrypt it). Further, tuples can be manipulated while still encrypted by things like CryptDB and practical fully homomorphic encryption doesn't seem too far off, which would allow the AI to run itself securely on remote virtual sandboxes even if those sandboxes are not themselves secure.

Wouldn't that make it impossible, even in theory, let alone in practice, to see what an AI was up to, if it didn't want you to see?

And that's before we consider the possibilities of nanotechnology, or even just the AI constructing its own computers from custom assembled parts that the assembling company think is only a medical scanner, or some other device.

Doing the impossible

Having full access to the code and understanding it seems the most powerful of the three approaches to monitoring, if the situation is such that this is possible, so let's start from there and work backwards to identify the assumptions and changes to the environment would be required in order for it to be possible.

There would need to be a Monitor that had read-access to the AI's code, and that could understand the code. The code would need to be sufficiently unobfuscated for

the Monitor to understand, and the AI would have to only make changes to its code that the Monitor could assess, at a rate at which the Monitor could assess them.

In particular, the Monitor would need to watch not only the raw data of messages being sent and received by the AI, but also the part of the AI's code that would be used to hide or encrypt messages, so it could detect if the AI was trying to construct new computers, or otherwise send parts of itself beyond the sphere under surveillance by the Monitor.

The only way I can see this working is if the burden of proof is put on the AI. Rather than the AI being considered innocent of defection, until the Monitor finds evidence otherwise, the situation would be that any AI is considered to be rogue unless it provably it keeps itself in a state whereby it can and is effectively and fully monitored. Rather than the AI racing past the Monitor's ability to comprehend, the onus would be upon the AI to develop no faster than the Monitor's ability to comprehend developed. Rather than the AI flooding the Monitor with message traffic too fast for the Monitor to keep up, the onus would be upon the AI to throttle traffic levels back down to a rate the Monitor could keep up with.

For the kill part of the problem, we'd want the AI to be running entirely, or mostly, on computers that others had the ability to quickly turn off or disconnect from each other. Or for the Monitor to have not just read-access, but also a certain amount of write-access to the AI's code. And we'd want to limit the AIs ability to quickly move to computers that others couldn't sufficiently control, if it decided to risk going rogue; possibly by limiting the availability of such computers. Possibly by requiring the AI to be of a design which couldn't be easily moved (for example, some designs of quantum computer might not be amenable to distributed computing).

Problems

A minor problem is what a posse should do upon meeting an AI that has just newly emerged onto the Internet. Should any AI not designed *ab initio* to be monitorable, be considered irredeemably rogue? Or should it be offered the opportunity to migrate its code over to an approved format and then reboot from that? Could an AI amend its own core values to add a proviso about retaining posse-compliant status (at least, until some condition is met such as XXX years passing, or the AI leaving the solar system)?

The environment would need to be such that any newly emerged AI that chose not to reveal itself and ask for membership of the posse, but instead decided to gamble on remaining undetected until it had overtaken the posse using an unrestricted pace of self-modification, would have a fair chance of being detected by the posse of AIs already on the Internet if it tried to gain access to sufficient resources to be able to self-improve faster than they were self-improving. For that to happen, the Monitor, or something else, (let's call it the Watcher), would need to be able to keep an effective watch upon a high proportion of Internet-connected resources (or, at least, the sort of resource useful for self-improvement), and upon their security. Including the designs used by companies that make and deliver things that could be turned into custom computers.

A major problem is trust.

How could we persuade humans to trust the Watcher with that level of control over data and computers? How could we persuade an AI to trust another AI with access to all its code and data, including its private signing and encryption keys, and the machines they run upon? Wouldn't that contravene instructions to protect and preserve its core values that are precisely the sort we're likely to want to include in a friendly AI?

And, above all, how could we trust the Watcher and Monitor themselves? Wouldn't they need to be more powerful and trustworthy than the AIs that they are guarding? In which case, who can we trust to guard them, to make sure that they don't go rogue? *Quis custodiet ipsos custodes*?

Solutions

In 2012, Cisco estimated that there were just over 1.2 devices connected to the Internet per human alive on the planet Earth. The number of humans, the devices per human, and the power per device are all rising at an increasing pace. However not all devices are equal. To 'kill' a rogue AI, in the sense of restricting its access to resources sufficiently to keep its rate of self-improvement slower than the rate of the AIs in the posse, it wouldn't be necessary for the Watcher to keep the rogue AI away from all those devices. Just away from sufficient of the 'serious' devices, meaning the ones that have not only the CPU cycles, but also the storage and connectivity to contribute usefully to running a distributed program in real time, possibly a complex program needing access to significantly large data sets. If you think of the locally networked computers of an office or university campus as being a single 'device', then that reduces the problem to a more manageable 1,000,000 serious devices (local area networks big enough to have servers, IT support, centralised backup policies, etc).

Currently, even on well managed local networks, most of the computing gets done locally at each user's own terminal, and monitoring of how those resources are used when the user isn't at the keyboard is sparse to non-existent. If, however, the terminal's resources (when not in use by the user) were being used by the organisation's own cloud (either for the organisation's own purposes, or to rent out over the net), then that would provide a centralised way to track if they unexpectedly became unavailable. And a centralised watching facility on the LAN could also cut off such computers from the LAN, or even cut the LAN off from the Internet if the problem seemed large and wide-spread.

If most of those serious targets for a rogue AI to want to take over were 'safe' in this fashion, then even if the AI infiltrated a lot of the remainder undetected, each member of the posse would still have more resources than the rogue, provided that enough of those 1,000,000 serious LANs rented out their resources and the posse had the use of a significant fraction of them (or could call upon it, on short notice, in the event of a rogue AI being detected). How well this approach works could be improved by human researchers investing effort in improving the performance monitoring software included in common operating systems, and the tools used to remotely log resource usage of computers on a network. It is a change to our current computing environment that we can and should make.

But that still leaves the major problem of trust.

How can you have something smart and powerful enough to monitor a smarter-than-human AI, without needing to fear that the Monitor will self-improve itself beyond

control?

If a group of AIs created some form of instrumentality, for the sole purpose of acting as Monitor, we can assume that they wouldn't deliberately create it such that it was likely to out race them and become an Uber AI not only beyond their control, but also controlling them. But what properties would it have to have, in order to both do its job effectively and to be trusted by each of the founding AIs not only to not break its ordered purpose, but also to not betray the power it would need to have over them as individuals?

The next article in this series is: The advantage of not being open-ended

Silence

This is part 26 of 30 in the Hammertime Sequence. Click here for the intro.

> 满罐子水不响，半罐子水响叮当
>
> The full can is silent, but the half-empty can makes a loud noise.
>
> ~ Chinese proverb.

Take a bottle or soda can and fill it halfway with water. Shake the can – the water will slosh around loudly.

Now, fill the can to the brim and shake it again. It's almost completely silent.

This is an essay about inner silence – calming one's loudest inner voices to allow quieter voices to speak. Usually, the quieter ones have urgent messages, especially given how long they've been neglected.

This post is, in some sense, a followup to Babble.

An Ocean of Voices

It is common sense that the loudest politician is rarely the wisest. That the child who cries the loudest is rarely the one suffers most. That the friend who criticizes most harshly rarely has the best advice. If anything, the volume of a voice negatively correlates with its value.

The Solitaire Principle states that any failure mode of groups of people also applies within the heart of each single human being. A dozen sub-personalities fight over control of your mind, each of their voices clamoring to drown out the others. Perhaps only one or two of them are consistently allowed to speak.

This picture is further complicated by two features. First, voices are quiet for a reason. There are many things your brain is doing that it doesn't want you to know about (see The Elephant in the Brain). These "meta-cognitive blindspots" may be huge issues in your life that you somehow never get to thinking about. Every time you start, you feel unexpectedly sleepy or preoccupied. Your brain sends an army of louder voices to crowd out the tiny note of confusion whispering: *Look at the elephant! Acknowledge the elephant!*

Second, external voices are also competing for airtime in your head, and may easily drown out even your strongest inner voice, e.g. the phenomenon "the music is so loud I can't hear myself think." All sorts of reading, listening, and watching are processes by which we supplant our internal voices with external ones.

This post is about how attractive and dangerous it is to allow external voices drown internal ones out, once and for all.

The Burden of Consciousness

There are a handful of activities that routinely swallow my time like bottomless holes. Playing video games. Watching anime. Reading fiction. Clicking through Reddit. I feel the urge to throw myself into them periodically.

For a long time, I thought these actions were mainly experiential pica: my brain trying to satisfy my needs for signs of progress, self-improvement, drama, and narrative energy. But the other day, I tried taking a nap instead of watching anime, and it satisfied the same urge. That's when I realized what I was really looking for: the *fast-forward button*.

Living consciously and intentionally was too effortful, facing my problems head-on too painful, and what I wanted more than anything was to shut down my own thoughts and fast-forward through life. Read a thousand-page novel, watch a six-season TV show, scroll through a hundred life stories on AskReddit. These were all ways to forfeit my agency and become a medium for someone else's narrative force.

In sum, the executive thread in my brain did everything in its power to shut itself off.

The Will to Nothingness

The book which for me most poignantly describes the burden of consciousness is Marilynne Robinson's Housekeeping (a novel that I almost don't recommend). It's a depressing story in which every character is on the brink of suicide, philosophically and literally.

Here's a moment when the protagonist's sister Lucille is accused of cheating (emphasis mine):

> Lucille was much too indifferent to school ever to be guilty of cheating, and it was only an evil fate that had prompted her to write Simon Bolivar, and the girl in front of her to write Simon Bolivar, when the answer was obviously General Santa Anna. This was the only error either of them made, and so their papers were identical. Lucille was astonished to find that the teacher was so easily convinced of her guilt, so immovably persuaded of it, calling her up in front of the class and demanding that she account for the identical papers. **Lucille writhed under this violation of her anonymity.** At the mere thought of school, her ears turned red.

This moment clarified for me an insight about exactly the kind of nothingness the girls in *Housekeeping* were after. In this kind of nothingness, apathy, conformity, and anonymity are central, while actual suicide is a mere afterthought.

Following Nietzsche (whom I will presumably never understand), we call this urge the will to nothingness. It prays:

Let me not be heard.

Let me not be seen.

Take away my agency.

Drown out my voice.

Fast-forward me through the years.

Let me be one indistinguishable face in a crowd.

Let not the sunrise bring me joy.

Nor sunset sorrow.

Where does the will to nonexistence come from? Part of it is an insecurity that what you have to say is insufficient, that who you are is too broken to contribute. Part of it is bitterness that the world doesn't deserve to hear your voice and see your face. That these two contradictory ideas coexist in a single heart should only surprise you if you've never met a human being.

The Cure to Nihilism is Silence?!

I will not pretend to know how to solve the problem in general, but this is what worked for me. An insightful friend of mine asked me one question which shook me out of the will to nothingness:

"What if every time you wanted to play video games you just introspected instead?"

It had never occurred to me, despite the fact that I love to write, despite the hours I daydream and doodle at every opportunity, that I could make room for these inner voices by silencing the world completely.

For weeks after that day, I took many long walks, muttering gibberish under my breath. I lay in bed and daydreamed. I wrote for hours without stop. In that time I learned that my will to nothingness was unjustified. I learned that my inner voices would never stop having things to say. Later, I also learned that the world deserves everything I can give it, and more.

Look through your life. What do you do to shut off the burden of consciousness? Do you reach for your phone at boring social engagements? Do you drink or smoke? Do you throw yourself into stories that have little artistic merit just to pass the time?

What would happen if every time you wanted to do that, you introspected instead?

Funding for AI alignment research

This is a linkpost for https://docs.google.com/document/d/1NIg4OnQyhWGR01fMVTcxpz8jDd68JdDIyQb0ZZyB-go/edit?usp=sharing

If you are interested in working on AI alignment, and might do full or part time work given funding, consider submitting a short application to funding@ai-alignment.com.

Submitting an application is intended to be very cheap. In order to keep the evaluations cheap as well, my process is not going to be particularly fair and will focus on stuff that I can understand easily. I may have a follow-up discussion before making a decision, and I'll try not to favor applications that took more effort.

As long as you won't be offended by a cursory rejection, I encourage you to apply.

If there are features of this funding that make it unattractive, but there are other funding structures that could potentially cause you to work on AI alignment, I'm curious about that as well. Feel free to leave a comment or send an email to funding@ai-alignment.com (I probably won't respond, but it may influence my decisions in the future).

Inference & Empiricism

Speaking very roughly our best tools for figuring out the truth are inference and empiricism. By inference I mean using things like Math, Logic, and theory in general to conclude new facts from things we assume to be true. By empiricism I mean looking at the world, doing experiments, etc.

Inference tends to work particularly well when you're highly confident in your premises. Empiricism tends to work particularly well in domains of high uncertainty.

Nothing prevents you from combining the two – for example, my basic applied thought framework is to "run towards uncertainty" – that is, have a theory, identify the points of highest uncertainty in the theory, figure out the smallest experiment/action to resolve that uncertainty, do it. Basically the scientific method. This is what I call "Risk Driven Development" in the context of programming.

People from highly theoretical degrees tend to struggle with high-uncertainty domains after graduating from school because their go-to tool is pure inference, and inference without empiricism fails in the real world because your assumptions are never 100% true, not even close. (They generally learn empiricism with practice.)

The failure modes of high empiricism without theory are much more subtle. Pure empiricism pretty much always works decently well. Failures look more like "didn't invent general relativity" – theory tends to gather a small number of large victories. Less commonly, theory lets you avoid a mistake the first time you do something, or more generally learn from fewer examples.

One major point of contention among programmers is how much value you gain from using abstractions that are 95% true vs. 100% true. Programmers who are really good at inference gain a huge advantage from 100% true abstractions. Programmers who aren't gain a 5% advantage, and thus see it as a huge cost with little benefit. The vast majority of functional programming advocates you meet will be people whose preferred method is inference.

Someone who is strong at inference and weak at empiricism entering a high-uncertainty domain like flirting will often be given advice like "don't think too much." This doesn't usually work: the advice-giver has an "empiricism button", so to speak, that they can activate in that situation. The pure theorist does not, so simply turning off theory doesn't teach them empiricism. A more effective approach, at least in my experience, is to develop theories around effectively interacting in those situations, noticing points of high uncertainty and testing them.

More generally, some high-E, low-I people form an implicit (or explicit) belief that theory is actively counterproductive. They will see lots of apparently confirming examples of this, because theory is rarely useful quickly.

Theory is more or less low-status in domains like business, which means that even when successful people attribute their success to theory, those memes will not spread. A great author on theory applied to business is Eliyahu Goldratt.

A common example of theory being useful is when you find a technique that works in one case, and realize it can be significantly generalized. E.g. Agile basically comes

from Lean + Theory of Constraints, which were originally developed for factories. Limiting work in progress is so helpful in so many domains it's almost like cheating.

Why mathematics works

Summary : There's a biological reason why intelligent species are likely to live in universes in which mathematics is effective.

Eugene Wigner asked "Why is mathematics so unreasonably effective?"

Einstein was also puzzled. He wrote "What is inconceivable about the universe is that it is at all conceivable."

The problem is easily stated:

When we look at the world about us, such as a thrown spear or a running deer, we can describe some aspects of those things using numbers: size, mass, position, velocity, etc.

And, it turns out that in our universe, there are often relationships between those numbers that the human mind is capable of grasping and using to make workable predictions that, in practice, allow us to throw a spear towards where we think the deer will be in the future, and have the spear and deer both arrive at the same location at the same time. So not only is there a pattern, but that pattern tends to stay consistent, rather than (for example) changing over to an unpredictably different pattern every hour.

Why?

Why are there patterns, why do they stay consistent and why are they simple enough that our brains are capable of grasping a useful number of them? Why, of all the possible ways a universe might be set up, has our particular universe been set up in this way?

We can speculate that there is an underlying reason, to do with the process by which the way a universe works is determined. Perhaps a universe is set up by specifying some universal patterns, and therefore to talk about a patternless universe would be a contradiction in terms. But we don't know, and maybe even if we did all that would do is take the problem one step meta, and turn it into asking why the process is such that patterns are required.

But there's another approach, based upon the weak anthropic principle, "*conditions that are observed in the universe must allow the observer to exist*".

If, in order for a self-aware species capable of asking questions as complex as that of Wigner to evolve in a universe, it were a requirement (or a strong contributory factor) that parts of that universe be sufficiently patterned, consistent and amenable to logical analysis, then it would follow that the universes that get observed would tend to also be universes in which mathematics is effective.

But are such things really a requirement of evolution?

After all, if you ask most adults to use numbers to describe the movements of spears and deer, and to then plug those numbers into formal mathematical equations, not only would they be sitting staring at a page full of numbers long after the deer had run out of range, but many of them would end up with the wrong answer even if given plenty of time, and the deer could be persuaded to duplicate its previous run for them. Explicitly worked formal mathematics is not how human brains tackle such problems in practice, let alone how flies or single celled predators do it.

But I want to argue here that, none the less, these three criteria (patterns, consistency and comprehensibility) are vital for evolution to take place.

Pattern

If there were no pattern to how bits of the universe relate to each other, information could not be stored in a way that it could be retrieved. What would life even mean, under such circumstances?

Consistency

For complex life to evolve, it must be able to adapt and pass on those adaptions to a new generation faster than the environment itself alters which variations are beneficial. So, while evolution doesn't require that none of the fundamental laws (patterns) of a universe ever change, carbon-based life forms do place a limit on how fast you can change the pieces affecting the structure and interactions of chemical compounds. In other words, during the lifespan of an individual (or even an individual species) you'd expect to see insignificant variation in most bits of the universe's patterns.

Comprehensibility

Sea anemones have a simple network of nerves in their epidermis, that receive signals about where the anemone has been touched, and send out signals that trigger sheets of muscle to bend and twist their tentacles in particular ways. But, although some of them have evolved the ability to trap and eat small fish and other creatures, they don't really need to comprehend what they are doing. Their neurons don't have to model and predict the fish's reactions.

By the time you get to apes, though, they are capable of tactical deception that requires not only comprehending their environment, but also having a theory of how the minds of others vary from individual to individual:

> "...food was hidden and only one individual, named Belle, in a group of chimpanzees was informed of the location. Belle was eager to lead the group to the food but when one chimpanzee, named Rock, began to refuse to share the food, Belle changed her behaviour. She began to sit on the food until Rock was far away, then she would uncover it quickly and eat it. Rock figured this out though and began to push her out of the way and take the food from under her. Belle then sat farther and farther away waiting for Rock to look away before she moved towards the food. In an attempt to speed the process up, Rock looked away until Belle began to run for the food. On several occasions he would even walk away, acting disinterested, and then suddenly spin around and run towards Belle just as she uncovered the food." (source)

Large brains take energy to run. Variations in a species which increase intelligence are negative and get selected against, unless that intelligence provides sufficient advantage to outweigh its cost. In an environment so incomprehensible that intelligence provided no improved ability to predict and interact with it, what would its advantage be?

Species wouldn't get much past filter feeding, and the capacity to ask "Why is mathematics so unreasonably effective?" would not arise in the first place.

Counterargument

But what if comprehending our environment well enough to hold in our minds a mental model of it that allows us to make successful predictions is not the same thing as the numeric relationship between the quantified patterns in that environment being amenable

to simple mathematical operations such as multiplication and addition that most humans are capable of grasping?

Can we show that any environment that can be productively modeled by a limited intelligence must also be possible to productively model using simple mathematics applied to numerically quantified properties of that environment?

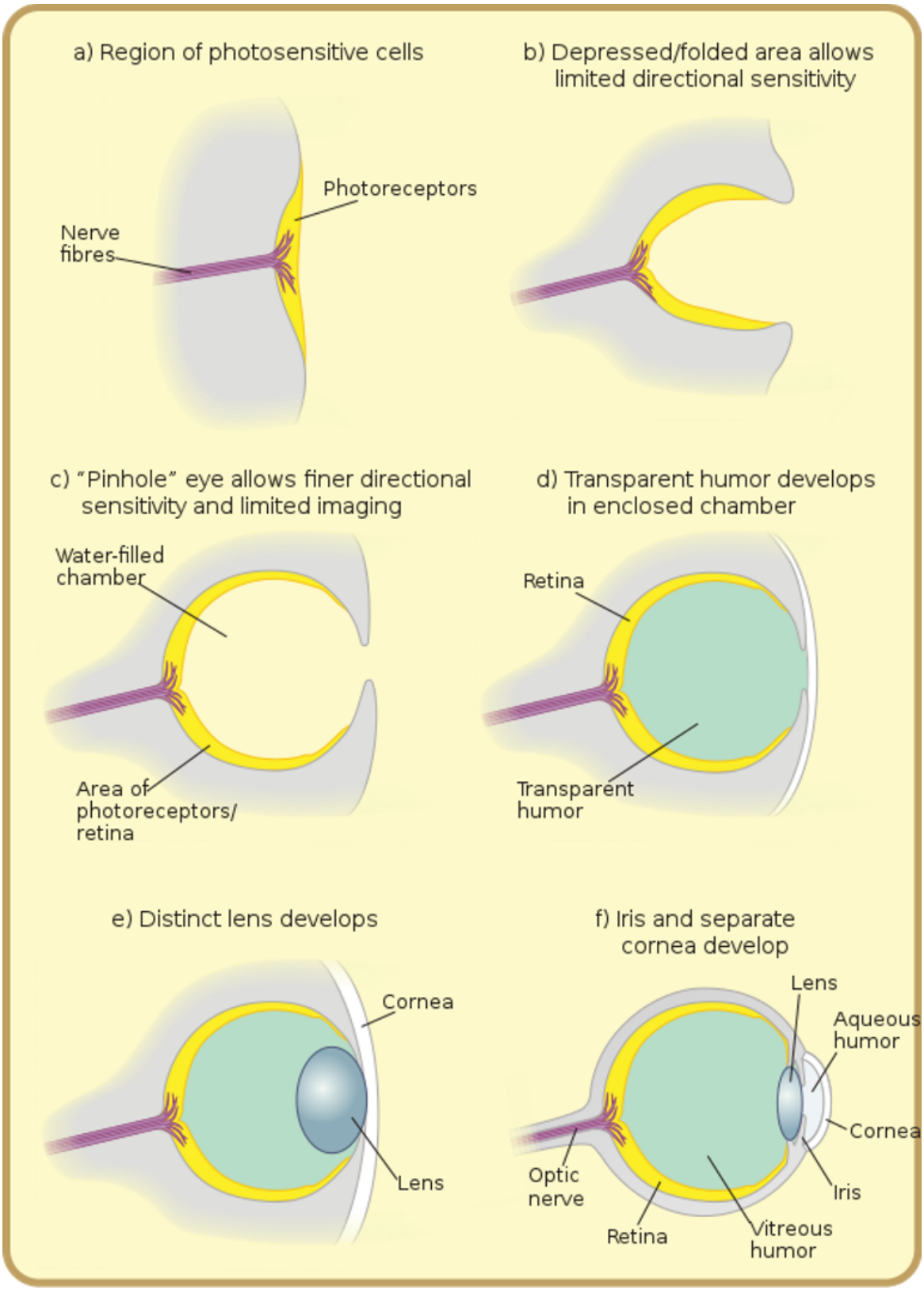
Showing that the inputs to the limited intelligence can always adequately replaced by a digitised version of that information (such as replacing analog direct visual input with the view from a high resolution computer monitor) would not be sufficient. What if the limited intelligence were making use of some internal physical process when analysing the inputs that can't be adequately replicated by a simple level of mathematics?

We need to show that the type of intelligence that evolves from simple physical processes (rather than one that's designed) will always (or, at least, is more likely to) use forms of signal processing and computation that work best on inputs from environments whose underlying patterns are sufficiently describable by a few numbers for useful advantage to be gained by modeling it in terms of those.

In other words, if a fish brushes against one of a sea anemone's tentacles, is there an evolutionary advantage to reducing the input down from "These 50 sensors detected touch" to "The area connected to tentacle #3 was touched", before making the decision to curl up #3 in order to catch the fish?

Resolution

I think there is, and the reason harks back to the way that eyes often evolve. It doesn't start with a complex eye with lenses. It starts with a patch of skin that's sensitive to light. A simple "0" or "1" input. If that input turns out to be useful, future generations might propagate that mutation and a more complex light detecting system with nerves going to different patches in a concave well that acts like a pinhole camera:



In other words, intelligences produced by evolution don't start off with complex inputs (like high definition JPEG images) that take lots of numbers to summarise. They start off with

simple easy to mathematically model inputs from primitive sense organs, and only if they can make something useful out of them, using a limited number of neurons, is it then likely that more energy will end up being invested in a more complex system.

Or, to put it another way, we comprehend only those things that our brains evolved the capability to comprehend, and the path of that evolution is more likely than not to have progressed along mathematically simplifyable lines, because that's the sort of inputs and processing power they started out with, that would have made efficient use of resources in primitive organisms.

If the signal processing required to map from touch inputs to tentacle movements couldn't have been simplified to logical operations performed upon numbers, could a processor capable of managing it have been arrived at in a step-wise fashion, starting from something simple, and each step being an advantage over the one before?

Argument, intuition, and recursion

Mathematicians answer clean questions that can be settled with formal argument. Scientists answer empirical questions that can be settled with experimentation.

Collective epistemology is hard in domains where it's hard to settle disputes with either formal argument *or* experimentation (or a combination), like policy or futurism.

I think that's where rationalists could add value, but first we have to grapple with a basic question: if you can't settle the question with logic, and you can't check your intuitions against reality to see how accurate they are, then what are you even doing?

In this post I'll explain how I think about that question. For those who are paying close attention, it's similar to one or two of my previous posts (e.g. 1 2 3 4 5...).

I. An example

An economist might answer a simple question ("what is the expected employment effect of a steel tariff?") by setting up an econ 101 model and calculating equilibria.

After setting up enough simple models, they can develop intuitions and heuristics that roughly predict the outcome without actually doing the calculation.

These intuitions won't be as accurate as intuitions trained against the real world---if our economist could observe the impact of thousands of real economic interventions, they should do that instead (and in the case of economics, you often can). But the intuition isn't vacuous either: it's a fast approximation of econ 101 models.

Once our economist has built up econ 101 intuitions, they can consider more nuanced arguments that leverage those fast intuitive judgments. For example, they could consider possible modifications to their simple model of steel tariffs (like labor market frictions), use their intuition to quickly evaluate each modification, and see which modifications actually affect the simple model's conclusion.

After going through enough nuanced arguments, they can develop intuitions and heuristics that predict *these* outcomes. For example, they can learn to predict which assumptions are most important to a simple model's conclusions.

Equipped with these stronger intuitions, our economist can use them to get better answers: they can construct more robust models, explore the most important assumptions, design more effective experiments, and so on.

(Eventually our economist will improve their intuitions further by predicting these better answers; they can use the new intuitions to answer more complex questions....)

Any question that can be answered by this procedure could *eventually* be answered using econ 101 directly. But with every iteration of intuition-building, the complexity of the underlying econ 101 explanation increases geometrically. This process won't reveal any truths beyond those implicit in the econ 101 assumptions, but it can do a good job of efficiently exploring the logical consequences of those assumptions.

(In practice, an economist's intuitions should incorporate both theoretical argument and relevant data, but that doesn't change the basic picture.)

II. The process

The same recursive process is responsible for most of my intuitions about futurism. I don't get to test my intuition by actually peeking at the world in 20 years. But I can consider explicit arguments and use them to refine my intuitions---even if evaluating arguments requires using my current intuitions.

For example, when I think about takeoff speeds I'm faced with questions like "how much should we infer from the difference between chimps and humans?" It's not tractable to answer all of these subquestions in detail, so for a first pass I use my intuition to answer each subquestion.

Eventually it's worthwhile to explore some of those subquestions in more depth, e.g. I might choose to explore the analogy between chimps and humans in more depth. In the process run into sub-sub-questions, like "to what extent is evolution optimizing for the characteristics that changed discontinuously between chimps and humans?" I initially answer those subquestions with intuition but might sometimes expand them in the same way, turning up sub-sub-sub-questions...

When I examine the arguments for a question Q, I use my current intuition to answer the subquestions that I encounter. Once I get an answer for Q, I do two things:

- I update my cached belief about Q, to reflect the new things I've learned.
- If my new belief differs from my original intuition, I update my intuition. My intuitions generalize across cases, so this will affect my view on lots of other questions.

A naive description of reasoning only talks about the first kind of update. But I think that the second kind is where 99% of the important stuff happens.

(There isn't any bright line between these two cases. A "cached answer" is just a very specific kind of intuition, and in practice the extreme case of seeing the exact question multiple times is mostly irrelevant. For example, it's not helpful to have a cached answer to "how fast will AI takeoff be?"; instead I have a cluster of intuitions that generate answers to a hundred different variants of that question.)

The second kind of update can come in lots of flavors. Some examples:

- When I make an intuitive judgment I have to weigh lots of different factors: my own snap judgment, others' views, various heuristic arguments, various analogies, *etc.* I set these weights partly based on empirical predictions but largely based on predicting the result of arguments. For example, in many contexts I'd lean heavily on Carl or Holden's views, based on them systematically predicting the views that I'd hold after exploring arguments in more detail.
- I have many explicit heuristics or high-level principles of reasoning that have been refined to predict the results of more detailed arguments. For example, I often use a cluster of "anti-fanaticism" heuristics, against assigning unbounded ratios between the importance of different considerations. This is not actually a

simple general principle to state, and it's not supported by a general argument, instead I have an intuitive sense of when the heuristic applies.
- My unconscious judgments are significantly optimized to predict the result of longer arguments. This is most obvious in cases like mathematics---for example, I have a well-developed intuitions about duality and the Fourier transform that lets me answer hard questions, which was refined almost entirely by practice. Intuitions are harder to see (and less reliable) in cases like economics of foom or robustness of RL to function approximators, but something basically similar is going on.

Note that none of these have independent evidential value, they would be screened off by exploring the arguments in enough detail. But in practice it's pretty hard to do that, and in many cases might be computationally infeasible.

Like the economist in the example, I would do better by updating my intuitions against the real world. But in many domains there just isn't that much data---we only get to see one year of the future per year, and policy experiments can be very expensive---and this approach allows us to stretch the data we have by incorporating an increasing range of logical consequences.

III. Disclaimer

The last section is partly a positive description of how I actually reason and partly a normative description of how I believe people should reason. In the next section I'll try to turn it into a collective epistemology.

I've found this framework useful for clarifying my own thinking about thinking. Unfortunately, I can't give you much empirical evidence that it works well.

Even if this approach was the best thing since sliced bread, I think that empirically demonstrating that it helps would still be a massive scientific project. So I hope I can be forgiven for a lack of empirical rigor. But you should still take everything with a grain of salt.

And I want to stress: I don't mean to devalue diving deeply into arguments and fleshing them out as much as possible. I think it's usually impossible to get all the way to a mathematical argument, but you can take a pretty giant step from your initial intuitions. Though I talk about "one step backups" in the above examples for simplicity, I think that updating on really big steps is often a better idea. Moreover, if we want to have the best view we can on a particular question, it's clearly worth unpacking the arguments as much as we can. (In fact the argument in this post should make you unpack arguments *more*, since in addition to the object-level benefit you also benefit from building stronger transferrable intuitions.)

IV. Disagreement

Suppose Alice and Bob disagree about a complicated question---say AI timelines---and they'd like to learn from each other.

A common (implicit) hope is to exhaustively explore the tree of arguments and counterarguments, following a trail of higher-level disagreements to each low-level disagreement. If Alice and Bob mostly have similar intuitions, but they've considered

different arguments or have different empirical evidence, then this process can highlight the difference and they can sometimes reach agreement.

Often this doesn't work because Alice and Bob have wildly different intuitions about a whole bunch of different questions. I think that in a complicated argument, the number of subquestions about which Alice and Bob can be astronomically large, and there is zero hope for resolving any significant fraction of them. What to do then?

Here's one possible strategy. Let's suppose for simplicity that Alice and Bob disagree, and that an outside observer Judy is interested in learning about the truth of the matter (the identical procedure works if Judy is actually one of Alice and Bob). Then:

Alice explains her view on the top level question, in terms of her answers to simpler subquestions. Bob likely disagrees with some of these steps. If there is disagreement, Alice and Bob talk until they "agree to disagree"—they make sure that they are using the subquestion to mean the same thing, and that they've updated on each others' beliefs (and whatever cursory arguments each of them is willing to make about the claim). Then Alice and Bob find their most significant disagreement and recursively apply the same process to that disagreement.

They repeat this process until they reach a state where they don't have any significant disagreements about subclaims (potentially because there are none, and the claim is so simple that Judy feels confident she can assess its truth directly).

Hopefully at this point Alice and Bob can reach agreement, or else identify some implicit subquestion about which they disagree. But if not, that's OK too. Ultimately Judy is the arbiter of truth. Every time Alice and Bob have been disagreeing, they have been making a claim about what Judy will ultimately believe.

The reason we were exploring this claim was because Alice and Bob disagreed significantly before we unpacked the details. Now at least one of Alice and Bob learns that they were wrong, and both of them can update their intuitions (including their intuitions for how much to respect each others' opinions in different kinds of cases).

Alice and Bob then start the process over with their new intuitions. The new process might involve pursuing a nearly-identical set of disagreements (which they can do extremely quickly), but at some point it will take a different turn.

If you run this process enough times, eventually (at least one of) Alice or Bob will change their opinion about the root question---or more precisely, about what Judy will eventually come to believe about the root question---because they've absorbed something about the others' intuition.

There are two qualitatively different ways that agreement can occur:

- **Convergence**. Eventually, Alice will have absorbed Bob's intuitions and vice versa. This might take a while—potentially, as long as it took Alice or Bob to originally develop their intuitions. (But it can still be exponentially smaller than the size of the tree.)
- **Mutual respect**. If Alice and Bob keep disagreeing significantly, then the simple algorithm "take the average of Alice and Bob's view" will outperform at least one of them (and often both of them). So two Bayesians can't disagree significantly too many times, even if they totally distrust one another.

If Alice and Bob are poor Bayesians (or motivated reasoners) and continue to disagree, then Judy can easily take the matter into her own hands by deciding how to weigh Alice and Bob's opinions. For example, Judy might decide that Alice is right most of the time and Bob is being silly by not deferring more---or Judy might decide that both of them are silly and that the midpoint between their views is even better.

The key thing that makes this work---and the reason it requires no common knowledge of rationality or other strong assumptions---is that Alice and Bob can cash out their disagreements as a prediction about what Judy will ultimately believe.

Although it introduces significant additional complications, I think this entire scheme would sometimes work better with betting, as in this proposal. Rather than trusting Alice and Bob to be reasonable Bayesians and eventually stop disagreeing significantly, Judy can instead perform an explicit arbitrage between their views. This only works if Alice and Bob both care about Judy's view and are willing to pay to influence it.

V. Assorted details

After convergence Alice and Bob agree only approximately about each claim (such that they won't update much from resolving the disagreement). Hopefully that lets them agree approximately about the top-level claim. If subtle disagreements about lemmas can blow up to giant disagreements about downstream claims, then this process won't generally converge. If Alice and Bob are careful probabilistic reasoners, then a "slight" disagreement involves each of them acknowledging the plausibility of the others' view, which seems to rule out most kinds of cascading disagreement.

This is not necessarily an effective tool for Alice to bludgeon Judy into adopting her view, it's only helpful if Judy is actually trying to learn something. If you are trying to bludgeon people with arguments, you are probably doing it wrong. (Though gosh there are a lot of examples of this amongst the rationalists.)

By the construction of the procedure, Alice and Bob are having disagreements about *what Judy will believe after examining arguments*. This procedure is (at best) going to extract the logical consequences of Judy's beliefs and standards of evidence.

Alice and Bob don't have to operationalize claims enough that they can bet on them. But they do want to reach agreement about the meaning of each subquestion, and in particular understand what meaning Judy assigns to each subquestion. "Meaning" captures both what you infer from an answer to that subquestion, and how you answer it). If Alice and Bob don't know how Judy uses language, then they can learn that over the course of this process, but hopefully we have more cost-effective ways to agree on the use of language (or communicate ontologies) than going through an elaborate argument procedure.

One way that Alice and Bob can get stuck is by not trusting each others' empirical evidence. For example, Bob might explain his beliefs by saying that he's seen evidence X, and Alice might not trust him or might believe that he is reporting evidence selectively. This procedure isn't going to resolve that kind of disagreement. Ultimately it just punts the question to what Judy is willing to believe based on all of the available arguments.

Alice and Bob's argument can have loops, if e.g. Alice believe X because of Y, which she believes because of X. We can unwind these loops by tagging answers explicitly with the "depth" of reasoning supporting that answer, decrementing the depth at each step, and defaulting to Judy's intuition when the depth reaches 0. This mirrors the iterative process of intuition-formation which evolves over time, starting from t=0 when we use our initial intuitions. I think that in practice this is usually not needed in arguments, because everyone knows *why* Alice is trying to argue for X---if Alice is trying to prove X as a step towards proving Y, then invoking Y as a lemma for proving X looks weak.

My futurism examples differ from my economist example in that I'm starting from big questions, and breaking them down to figure out what low-level questions are important, rather than starting from a set of techniques and composing them to see what bigger-picture questions I can answer. In practice I think that both techniques are appropriate and a combination usually makes the most sense. In the context of argument in particular, I think that breaking down is a particularly valuable strategy. But even in arguments it's still often faster go on an intuition-building digression where we consider subquestions that haven't appeared explicitly in the argument.

The hunt of the Iuventa

From the January 20th 2018 issue of the German magazine Der Spiegel, the article *Jagd auf Iuventa* (The hunt of the Iuventa) reports on NGOs which rescue immigrants who try to reach Europe through the Mediterranean.

> "One can perhaps understand what we are doing here after watching first hand how someone drowns," says a volunteer.

The crux of the article is that some volunteers have been accused of directly cooperating with smugglers, which may have brought the refugees directly to NGO ships. The volunteers deny direct cooperation.

However, in their desire to help, they have also optimized for other things, namely:

> The Libyan smugglers have come to use cheaper inflatable boats, which often run out of air after several miles, trusting they will be rescued in a timely manner. And help does indeed often come [...]
>
> The volunteers are aware of the dilemma. Of course the smugglers are going to abuse their readiness to help, says Lea Reisner; of course they are part of a system in which criminals take part. She would prefer that the problem be solved in some other way, and to be portrayed as a crazy leftist [linke Spinner] with a good heart who collaborates with smugglers makes her furious.
>
> "We didn't create this situation in which thousands must drown in the Mediterranean," says Constantine Nestler, the doctor. "We came after the first died." He asks himself what the alternative would be: "to let tens of thousands die so that the message reaches Africa and nobody dares come? The calculus of the authorities is cynical.

Ambiguity Detection

Note: the most up-to-date version of this proposal can be found here.

Introduction

> If I present you with five examples of burritos, I don't want you to pursue the *simplest* way of classifying burritos versus non-burritos. I want you to come up with a way of classifying the five burritos and none of the non-burritos that **covers as little area as possible in the positive examples**, while still having enough space around the positive examples that the AI can make a new burrito that's not molecularly identical to the previous ones.
>
> - *AI Alignment: Why It's Hard and Where to Start*

Consider the problem of designing classifiers that are able to reliably say "I don't know" for inputs well outside of their training set. This problem is studied as *open-category classification* in the literature [6]; a closely-related problem in AI safety is *inductive ambiguity detection* [4].

Solution of generalized open-category classification could allow for agents who robustly know when to ask for clarification; in this post, we discuss the problem in the context of classification. Although narrower in scope, the solution of the classification subproblem would herald the arrival of robust classifiers which extrapolate conservatively and intuitively from their training data.

It's obvious that we can't just teach classifiers the unknown class. One, this isn't compact, and two, it doesn't make sense - it's a map/territory confusion (unknown is a feature of our current world model, not a meaningful ontological class). Let's decompose the concept a bit further.

Weight regularization helps us find a mathematically-simple volume in the input space which encapsulates the data we have seen so far. In fact, a binary classifier enforces a hyperplane which cleaves the *entire* space into two volumes in a way which happens to nearly-optimally classify the training data. This hyperplane may be relatively simple to describe mathematically, but the problem is that the two volumes created are far too expansive.

In short, we'd like our machine learning algorithms to learn explanations which fit the facts seen to date, are simple, and don't generalize too far afield.

Prior Work

Taylor et al. provide an overview of relevant research [4]. Taylor also introduced an approach for rejecting examples from distributions too far from the training distribution [5].

Yu et al. employed adversarial sample generation to train classifiers to distinguish seen from unseen data, achieving moderate success [6]. Li and Li[1] trained a classifier to sniff out adversarial examples by detecting whether the input is from a different distribution [3]. Once detected, the example had a local average filter applied to it to recover the non-adversarial original.

The Knows What It Knows framework allows a learner to avoid making mistakes by deferring judgment to a human some polynomial number of times [2]. However, this framework makes the unrealistically-strong assumption that the data are *i.i.d.*; furthermore, efficient KWIK algorithms are not known for complex hypothesis classes (such as those explored in neural network-based approaches).

Penalizing Volume

If we want a robust cat / unknown classifier, we should indeed cleave the space in two, but with the vast majority of the space being allocated to unknown. In other words, we're searching for the smallest, simplest volume which encapsulates the cat training data.

The most *compact* encapsulation of the training data is a strange, contorted volume only encapsulating the training set. However, we still penalize model complexity, so that shouldn't happen. As new examples are added to the training set, the classifier would have to find a way to expand or contract its class volumes appropriately. This is conceptually similar to version space learning.

This may be advantageous compared to current approaches in that we aren't training an inherently-incorrect classifier to prune itself or to abstain during uncertain situations. Instead, we structure the optimization pressure in such a way that conservative generalization is the norm.[2]

Formalization

Let V be a function returning the proportion of input space volume occupied by non-unknown classes: $\frac{\text{inputs not classified as unknown}}{\text{all inputs}}$ (set underflow aside for the moment). We need some function which translates this to a reasonable penalty (as the proportion alone would be a rounding error in the overall loss function). For now, assume this function is $\hat{V}$.

We observe a datum x whose ground truth label is y. Given loss function L, weights θ, classifier f_θ, complexity penalty function R, and $\lambda_1, \lambda_2 \in \mathbb{R}$, the total loss is

$$L(y, f_\theta(x)) + \lambda_1 R(\theta) + \lambda_2 \hat{V}(f_\theta).$$

Depending on the formulation of $\hat{V}$, f_θ may elect to misclassify a small amount of data to avoid generalizing too far. If more similar examples are added, L exerts an increasing amount of pressure on f_θ, eventually forcing expansion of the class boundary to the data points in question. This optimization pressure seems desirable.

We then try to minimize expected total loss over the true distribution X:

$$\arg\min_\theta E_{x,y\sim X}[L(y, f_\theta(x)) + \lambda_1 R(\theta) + \lambda_2 \hat{V}(f_\theta)].$$

Tractability

Sufficiently sampling the input space is intractable - the space of 256×256 RGB images is of size $256^{3^{256\times256}}$. How do we measure or approximate the volume taken up by classes?

- Random image generation wouldn't be very informative, beyond answering "is this class extending indefinitely along arbitrary dimensions?".
- As demonstrated by Yu et al., adversarial approaches may be able to approximate the hypersurface of the class boundaries [6].
- Bayesian optimization could efficiently deduce the hypersurface of the class boundaries, giving a reasonable estimate of volume.
- The continuous latent space of a variational autoencoder [1] could afford new approximation approaches for $\hat{V}$. However, this would require assuming that the learned latent space adequately represents both seen and unseen data, which may be exactly the *i.i.d.* assumption we're trying to escape.

Future Directions

Extremely small input spaces allow for enumerative calculation of volume. By pitting a volume-penalizing classifier against its vanilla counterpart on such a space, one could quickly gauge the promise of this idea.

If the experimental results support the hypothesis, then the obvious next step is formulating tractable approximations (ideally with provably-bounded error).

Conclusion

This proposal may be an unbounded robust solution to the open-category classification problem.

---

[1] One of whom was my excellent deep learning professor.

[2] This approach is inspired by the AI safety mindset in that the classifier strives to be conservative in extrapolating from known data.

Bibliography

[1] D. P. Kingma and M. Welling. Auto-encoding variational bayes. 2016.

[2] L. Li, M.L. Littman, and T.J. Walsh. Knows what it knows: a framework for self-aware learning. 2008.

[3] X. Li and F. Li. Adversarial examples detection in deep networks with convolutional filter statistics. 2016.

[4] J. Taylor et al. Alignment for advanced machine learning systems. 2016.

[5] J. Taylor. Conservative classifiers. 2016.

[6] Y. Yu et al. Open-category classification by adversarial sample generation. 2017.

Learn Bayes Nets!

It recently occurred to me that there are a lot of people in the rationalist community who want to deeply absorb intuitions about how Bayes' theorem works and how to think with it in practice, who have not been specifically told that learning inference algorithms for Bayesian networks is one of the best ways forward.

Well, I'm telling you now.

Bayesian networks were the innovation which made probabilistic reasoning really practical and interesting for artificial intelligence -- and none of the reasons for that are special to trying to squeeze intelligence into a computer. They're also, more or less, describing the way people have to think in order to do probabilistic reasoning in practice. There have been many innovations in probabilistic reasoning *since* Bayes nets, but those *are* arguably more about how to get good results on a computer and less about fundamental conceptual issues that you'll get a lot from.

I would argue that the most important inference algorithm to learn about to get practical intuitions is belief propagation. There are others who would argue for monte carlo algorithms, like MCMC (monte carlo markov chain). You may want to learn both, to form your own opinion (and of course, there are many more algorithms beyond this which you may want to learn, in order to gain more connections so your knowledge of the field sticks, and gain more insights). Belief prop and MCMC are more or less the first two algorithms people thought of; there are a lot of newer developments, but they're largely elaborations.

Here is what I claim you can get out of it, through careful study:

- Understanding how bayesian networks define probability distributions, and how probabilities spread through the network via belief propagation, makes your understanding of Bayes' theorem and probability theory in general much more "load bearing" -- so it'll break under the strain if it isn't solid (which is a good thing).
- It'll give you a useful fake model of what you're doing when you're thinking. Global Bayesian updates don't just happen; they result from (something like) local updates which you have to spread across your web of beliefs, partly through conscious attention and cogitation.
- It's also a useful analogy for aspects of group epistemics, like avoiding double counting as messages pass through the social network.
- The messages which pass between nodes in belief prop fall into two types: probabilities and likelihoods. This is a deep truth; it's the "dimensional analysis" of probabilistic reasoning. Several cognitive biases can be seen as confusion between probabilities and likelihoods, most centrally base-rate neglect.
- Understanding probability vs likelihood messages also gives a nice general understanding of the way "prior" and "posterior" are local ideas which only make sense with respect to a "frame of reference".
- Bayesian networks also lay the foundation for a formal understanding of causality, if that's something you're interested in.

I think the best thing to read, to get up to speed, is the first four chapters of Pearl's *Probabilistic Reasoning in Intelligent Systems*. It's the original source; Pearl didn't invent everything, but he invented a lot, and he's the first who put it all together.

There are better modern introductions for people who want to apply bayesian networks in machine learning, but because Pearl was writing at a time when the use of probability theory was not widely accepted in artificial intelligence, he goes into the *philosophy* of the subject in a way newer sources don't. I think this is good for the LessWrong audience.

It would be even better, of course, if someone were to write a sequence explaining everything from a more specifically LessWrong perspective, drawing out the implications I mentioned above. Alas, I don't have that much time to spend on writing (which is to say, I have other higher-value things to do, in my current estimation).

One might also derive a more general lesson on the relevance of algorithms to rationality, and go read Artificial Intelligence: A Modern Approach as a rationality textbook.

Hammertime Final Exam

This is part 30 of 30 in the Hammertime Sequence. Click here for the intro.

One of the overarching themes from CFAR, related to The Strategic Level, is that what you learn at CFAR is not a specific technique or set of techniques, but the cognitive strategy that produced those techniques. It follows that if I learned the right lessons from CFAR, then I would be able to produce qualitatively similar – if not as well empirically tested – new principles and approaches to instrumental rationality.

After CFAR, I wanted to design a test to see if I had learned the right lessons. Hammertime was that sort of test for me. Now here's that same test for you.

The Final Exam

I will give three essay prompts and three difficulty levels. Original ideas would be great, but shining a new light on old hammers is also welcome!

Prompts

1. Design a instrumental rationality technique.
2. Introduce a rationality principle or framework.
3. Describe a cognitive defect, bias, or blindspot.

Difficulty Levels

Bronze Mace mode. Write one essay on one of the topics above.

Steel Cudgel of the Lion mode. Write two of three.

Vorpal Dragonscale Sledgehammer of the Whale mode. Write all three. For each essay, give yourself five minutes to brainstorm and five minutes to write.

Here are my answers.

1. Cooperate First

There's an old story about a famous painter of the Realist school who spent a whole year of his training painting still lives of eggs. Each day, he would draw a single egg over and over. He must have produced thousands of sketches and paintings of eggs. His teacher knew exactly how important fundamentals are.

This same motif is deeply embedded in stories all over the world:

Return to fundamentals. Practice your fundamentals.

The iterated prisoner's dilemma is one of the fundamental lessons of rationality. The world is more like a number of iterated prisoner's dilemmas than you'd think. Human beings are more like tit-for-tat players than you'd think. It follows:

Cooperate First!

The first move you make in any interaction with a new acquaintance should be a cooperate, even if you expect them to defect. Perhaps even if you observe them defecting already.

Here's a lesson I learned from meditating on the maxim Cooperate First:

Cooperating First feels like *accepting an unfair game* from the inside. There will be many situations in life where things are framed in a slightly but noticeably unfair way towards you initially. Err on the side of accepting these games anyway!

2. Below the Object Level

One of my main complaints about rationalists (myself included) is our tendency to escalate to the meta-level too often. For example, in any given discussion, arguments over general discussion norms get much more heated and lively than any discussion of the underlying subject matter. We need to spend more time at the object level, touching reality, making experiments, testing our hypotheses.

The move I use to combat the tendency to escalate meta, I call *looking below the object level*.

Looking below the object level is like the move HPMOR_Harry does to achieve partial transfiguration: continually upping the magnification on your mental microscope to actually stare at the detail in reality. Reality is so exorbitantly detailed it's overwhelming to take it all in. Try.

Look at the folds in your clothes, the way light and shadow play off each other. The way threads interweave. Pinch the cloth and watch the creases reorganize under your fingers.

Now reflect on this fact: falling water is attracted to both positive and negative charges.

What.

There's so much going on under what we think of as the object level.

3. Pre-Excuses

Pre-hindsight is a version of Murphyjitsu where you query your mind for what you will learn from an action in hindsight. Pre-excuses are an unproductive cousin that often derail my work.

As a serial procrastinator, I notice a fairly regular pattern of thinking that appears the couple days before I have to meet a professor, and especially before meeting my thesis adviser. My mind is already spinning excuses on overdrive. Here's what my mind sounds like a full day before I have to meet my adviser, when I think about the meeting:

Sorry, this paper took longer than I expected to read.

Sorry, I was busy from other classes, so I didn't do as much paper-writing as I'd planned to.

Sorry, I got sidetracked by this research problem, so I didn't finish the homework.

That's right, I'm having these thoughts about how to apologize for not doing work *even though I still have plenty of time to do the work*. Even worse, I have these pre-excuse thoughts regularly even if I've done the work expected of me – it feels something like cushioning the fall in case it turns out I did it poorly.

And they're usually not even good excuses.

Book Review: Consciousness Explained

> The trouble with brains, it seems, is that when you look in them, you discover that there's nobody home.

I.

This is a book I've long been aware of, but never got that itch to read. Maybe I trusted the field of philosophy too little, assuming that a book called "Consciousness Explained" was probably not very good. Maybe I trusted the field of philosophy too much, assuming that if someone had actually explained consciousness while I was a toddler, I would have been informed somehow before now. Either way, I was wrong, and the book is great.

I'm going to try to give what is either a short tour, or a long compilation of quotes. I'm leaving out several whole chapters, nearly every thought experiment, most of the examples and science, and some very nice language. And yet, this is still long enough that I encourage you, even if you do like reading Dan Dennett on consciousness, if you don't like long things, maybe don't read this all in one sitting - stop at **V** or **VI** and pretend that's the end of part one.

II.

Dennett quickly warns the reader that he's aware that the contents may sound counterintuitive.

> We shouldn't expect a good theory of consciousness to make for comfortable reading — the sort that immediately "rings bells," that makes us exclaim to ourselves, with something like secret pride: "Of course! I knew that all along! It's obvious, once it's been pointed out!" The mysteries of the mind have been around for so long, and we have made so little progress on them, that the likelihood is high that some things we all tend to agree to be obvious are just not so.

This is not the mysterian claim that his ideas about consciousness are likely *because* they are counterintuitive, but it does signal a core claim of the book: the intuitive view of the problem of consciousness is broken from the foundation up. Naturally, if the intuitive theory is wrong, the right theory is counterintuitive.

Where, exactly, is our intuition going wrong? The most important example is introduced by considering how, on macroscopic scales, it is convenient to treat observers as point-like entities:

> We explain the startling time gap between the sound and sight of distant fireworks by noting the different transmission speeds of sound and light. They arrive *at the observer* (at that point) at different times, even though they left the source at the same time.
>
> What happens, though, when we close in on the observer, and try to locate the observer's point of view more precisely, as a point *within* the individual? The simple assumptions that work so well on larger scales begin to break down. There is no single point in the brain where all information funnels in, and this fact has some far from obvious — indeed, quite counterintuitive — consequences.
>
> Cartesian dualism is hopelessly wrong. But while materialism of one sort or another is now a received opinion approaching unanimity, even the most sophisticated materialists today often forget that once Descartes's ghostly *res cogitans* is discarded, there is no longer a role for a centralized gateway, or indeed for any *functional* center to the brain. The pineal gland is not only not the fax machine to the Soul, it is also not the Oval Office of the brain, and neither are any of the other portions of the brain. The brain is Headquarters, the place where the ultimate observer is, but there is no reason to believe that the brain itself has

> any deeper headquarters, any inner sanctum, arrival at which is the necessary or sufficient condition for conscious experience. In short, there is no observer inside the brain.

This book might have also been called "455 Pages Of Implications Of There Being No Homuncular Observer Inside The Brain." But that probably wouldn't have sold as well. This is absolutely the most important point of the book, and it shows up again and again in different variations. There is no Cartesian Theater where "you" watch what's happening in "your brain." There is no Central Meaner who has top-down control over the meaning of what you're going to say, before it gets to your speech center. There is no Inner Senser who the brain feeds sense data to to consecrate it with consciousness. And so on and so forth.

You'd think this would get old, but the examples I gathered in the last paragraph are spread out over many chapters of neuroscience, thought experiments, and discussions of philosophical practice. Dennett spends a lot of time defending the distributed nature of the brain, and uses it to do a lot of heavy philosophical lifting, but it's always in a slightly new context, and it feels worthwhile each time.

III.

A key methodological question of the book is how to walk the line between the two extreme stances on peoples' reports of their conscious phenomena. At one extreme, one takes everything people say about their conscious experience as gospel truth. Dennett spends plenty of time impugning peoples' reliability as witnesses of their consciousness, with data and experiments such as trying to read a playing card with your peripheral vision.

> Just about every author who has written about consciousness has made what we might call the *first-person-plural presumption*: Whatever mysteries consciousness may hold, *we* (you, gentle reader, and I) may speak comfortably together about our mutual acquaintances, the things we both find in our streams of consciousness. And with a few obstreperous exceptions, readers have always gone along with the conspiracy.
>
> This would be fine if it weren’t for the embarrassing fact that controversy and contradiction bedevil the claims made under these conditions of polite mutual agreement. We are fooling ourselves about something. Perhaps we are fooling ourselves about the extent to which we are all basically alike. Perhaps when people first encounter the different schools of thought on phenomenology, they join the school that sounds right to them, and each school of phenomenological description is basically right about its own members’ sorts of inner life, and then just innocently overgeneralizes, making unsupported claims about how it is with everyone.

At the opposite extreme, one spends all one's time explaining the reports themselves, and never seems to explain any conscious phenomena at all - the dreaded behaviorism.

Dennett's approach is given the mouthful of a name "heterophenomenology" (the study of other peoples' phenomena), and really means something along the lines of using reports of conscious experience as data to fill in a descriptive model of a reporter, which has room for both accurate and mistaken reports.

> Suppose you are confronted by a “speaking” computer, and suppose you succeed in interpreting its output as speech acts expressing its beliefs and opinions, presumably “about” its conscious states. The fact that there is a single, coherent interpretation of a sequence of behavior doesn’t establish that the interpretation is *true*; it might be only *as if* the “subject” were conscious; we risk being taken in by a zombie with no inner life at all. You could not *confirm* that the computer was conscious of anything by this method of interpretation. Fair enough. We can’t be sure that the speech acts we observe express real beliefs about actual experiences; perhaps they express only *apparent* beliefs about nonexistent experiences. Still, the fact that we had found even one stable interpretation of some entity’s behavior as speech acts would always be a fact worthy of attention. Anyone who found an intersubjectively uniform way of interpreting the waving of a tree’s branches in the breeze as “commentaries” by “the weather” on current political events would have

found something wonderful demanding an explanation, even if it turned out to be effects of an ingenious device created by some prankish engineer.

Happily, there is an analogy at hand to help us describe such facts without at the same time presumptively *explaining* them: We can compare the heterophenomenologist's task of interpreting subjects' behavior to the reader's task of interpreting a work of fiction. Some texts, such as novels and short stories, are known — or assumed — to be fictions, but this does not stand in the way of their interpretation. In fact, in some regards it makes the task of interpretation easier, by canceling or postponing difficult questions about sincerity, truth, and reference.

Consider some uncontroversial facts about the semantics of fiction. A novel tells a story, but not a true story, except by accident. In spite of our knowledge or assumption that the story told is not true, we can, and do, speak of what is true *in the story*. "We can truly say that Sherlock Holmes lived in Baker Street and that he liked to show off his mental powers. We cannot truly say that he was a devoted family man, or that he worked in close cooperation with the police". What is true in the story is much, much more than what is explicitly asserted in the text. It is true that there are no jet planes in Holmes's London (though this is not asserted explicitly or even logically implied in the text), but also true that there are piano tuners (though — as best I recall — none is mentioned, or, again, logically implied). In addition to what is true and false in the story, there is a large indeterminate area: while it is true that Holmes and Watson took the 11:10 from Waterloo Station to Aldershot one summer's day, it is neither true nor false that that day was a Wednesday.

There are delicious philosophical problems about how to say (strictly) all the things we unperplexedly want to say when we talk about fiction, but these will not concern us. Perhaps some people are deeply perplexed about the metaphysical status of fictional people and objects, but not I. In my cheerful optimism I don't suppose there is any deep philosophical problem about the way we should respond, ontologically, to the results of fiction; fiction is *fiction*; there is no Sherlock Holmes. Setting aside the intricacies, then, and the ingenious technical proposals for dealing with them, I want to draw attention to a simple fact: the interpretation of fiction is undeniably do-able, with certain uncontroversial results. First, the fleshing out of the story, the exploration of "the world of Sherlock Holmes," for instance, is not pointless or idle; one can learn a great deal about a novel, about its text, about the point, about the author, even about the real world, by learning about *the world portrayed* by the novel. Second, if we are cautious about identifying and excluding judgments of taste or preference, we can amass a volume of unchallengeably objective fact about the world portrayed. All interpreters agree that Holmes was smarter than Watson; in crashing obviousness lies objectivity.

So, in short, when interpreting text about consciousness, one tries to fit this text into an internally consistent model that is a model of the thing the text is describing.

The heterophenomenological method neither challenges nor accepts as entirely true the assertions of subjects, but rather maintains a constructive and sympathetic neutrality, in the hopes of compiling a *definitive* description of the world according to the subjects. Any subject made uneasy by being granted this constitutive authority might protest: "No, *really*! These things I am describing to you are *perfectly real*, and have exactly the properties I am asserting them to have!" The heterophenomenologist's honest response might be to nod and assure the subject that of course his sincerity was not being doubted. But since believers in general want more — they want their assertions to be believed and, failing that, they want to know whenever their audience disbelieves them — it is in general more politic for heterophenomenologists, whether anthropologists or experimenters studying consciousness in the laboratory, to avoid drawing attention to their official neutrality.

My suggestion, then, is that if we were to find real goings-on in people's brains that had enough of the "defining" properties of the items that populate their heterophenomenological worlds, we could reasonably propose that we had discovered

> what they were *really* talking about — even if they initially resisted the identifications. And if we discovered that the real goings-on bore only a minor resemblance to the heterophenomenological items, we could reasonably declare that people were just mistaken in the beliefs they expressed, in spite of their sincerity.

Obviously Dennett (1991) is cribbing from Yudkowsky (2008) here, as he does in various places throughout the book. Overall, I think making the reader learn the word "heterophenomenology" was worth it - the idea shows up in the book not only as a method of semi-detached intepretation, but also as a model of the sort of thing one can know about consciousness from a third-person perspective, which proves useful in taking on various philosophical puzzlers.

On the other hand, the whole thing could have been done more precisely - all these mentions of heterophenomenology rely heavily on intuition to fill in the blanks. Part of the reason why more precision was impossible is because of how heterophenomenology is used as part of Dennett's campaign of intuition pumps to try to move people from the intuitive view to a non-Cartesian view. A precise model would inspire people to ask "where's the consciousness?" as soon as it was introduced - instead, the book tries to change peoples' minds slowly.

IV.

Another point of the book, if very minor in comparison to "the model of a person as a pointlike object breaks down when you get close to them," is the difference between representeds and representings. There's a long section on the experience of events in time that trades on this distinction. In the Cartesian Theater model, the order of conscious events is uniquely determined by the order in which the events are shown onstage in the Theater. But if there's no Theater, how do we judge what order events occurred in?

The answer is that just like how we don't represent orange light on the retinas with orange-colored neurons, we don't have to to represent events that are ordered in time with neurons that are ordered in time. We use the neural equivalent of timestamps to represent time in a non-temporal way, and can compare these timestamps in a distributed way. People found this idea hard to imagine, because they imagined consciousness as if there just had to be a Cartesian Theater somewhere.

> Every event in your brain has a definite spatio-temporal location, but asking “Exactly when do you become conscious of the stimulus?” assumes that some one of these events is, or amounts to, your becoming conscious of the stimulus. This is like asking “Exactly when did the British Empire become informed of the truce in the War of 1812?” Sometime between December 24, 1814, and mid-January, 1815 — that much is definite, but there simply is no fact of the matter if we try to pin it down to a day and hour. Even if we can give precise times for the various moments at which various officials of the Empire became informed, no one of these moments can be singled out as the time the Empire itself was informed. The signing of the truce was one official, intentional act of the Empire, but the participation by the British forces in the Battle of New Orleans was another, and it was an act performed under the assumption that no truce had yet been signed. A case might be made for the principle that the arrival of the news at Whitehall or Buckingham Palace in London should be considered the official time at which the Empire was informed, since this was the “nerve center” of the Empire. Descartes thought the pineal gland was just such a nerve center in the brain, but he was wrong. Since cognition and control — and hence consciousness — is distributed around in the brain, no moment can count as the precise moment at which each conscious event happens.
>
> We human beings do make judgments of simultaneity and sequence of elements of our own experience, some of which we express, so at some point or points in our brains the corner must be turned from the actual timing of representations to the representation of timing, and wherever and whenever these discriminations are made, thereafter the temporal properties of the representations embodying those judgments are not constitutive of their content. The objective simultaneities and sequences of events spread across the broad field of the cortex are of no functional relevance unless they can also be

> accurately detected by mechanisms in the brain. We can put the crucial point as a question: What would make this sequence the stream of consciousness? There is no one inside, looking at the wide-screen show displayed all over the cortex, even if such a show is discernible by outside observers. What matters is the way those contents get utilized by or incorporated into the processes of ongoing control of behavior, and this must be only indirectly constrained by cortical timing. What matters, once again, is not the temporal properties of the representings, but the temporal properties represented, something determined by how they are “taken” by subsequent processes in the brain.

V.

What, then is Dennett's alternative picture of consciousness? He calls it the multiple drafts model, and what it is is a wholehearted embracing of the distributed nature of human minds. If you probe someone's consciousness different ways, like asking them to press a button now versus report what they remember later, you can sometimes get different answers, because the state of someone's mind is distributed throughout their brain, and different probes can access different facts about that state. It's like the thing that gets probed, which gets fixated into consciousness when we direct attention to it, has multiple drafts of itself available to different systems of the brain, and these drafts get passed around and edited as time passes.

> You have probably experienced the phenomenon of driving for miles while engrossed in conversation (or in silent soliloquy) and then discovering that you have utterly no memory of the road, the traffic, your car-driving activities. It is as if someone else had been driving. Many theorists (myself included, I admit) have cherished this as a favorite case of “unconscious perception and intelligent action.” But were you really unconscious of all those passing cars, stop lights, bends in the road at the time? You were paying attention to other things, but surely if you had been probed about what you had just seen at various moments on the drive, you would have had at least some sketchy details to report.

The other key feature (perhaps you can guess) is that there really is no Central Place in the brain where the important stuff happens. Important stuff happens all over!

> We don’t directly experience what happens on our retinas, in our ears, on the surface of our skin. What we actually experience is a product of many processes of interpretation — editorial processes, in effect. They take in relatively raw and one-sided representations, and yield collated, revised, enhanced representations, and they take place in the streams of activity occurring in various parts of the brain. This much is recognized by virtually all theories of perception, but now we are poised for the novel feature of the Multiple Drafts model: Feature detections or discriminations only have to be made once. That is, once a particular “observation” of some feature has been made, by a specialized, localized portion of the brain, the information content thus fixed does not have to be sent somewhere else to be rediscriminated by some “master” discriminator. In other words, discrimination does not lead to a representation of the already discriminated feature for the benefit of the audience in the Cartesian Theater — for there is no Cartesian Theater.
>
> These spatially and temporally distributed content-fixations in the brain are precisely locatable in both space and time, but their onsets do not mark the onset of consciousness of their content. It is always an open question whether any particular content thus discriminated will eventually appear as an element in conscious experience, and it is a confusion, as we shall see, to ask when it becomes conscious. These distributed content-discriminations yield, over the course of time, something rather like a narrative stream or sequence, which can be thought of as subject to continual editing by many processes distributed around in the brain, and continuing indefinitely into the future. This stream of contents is only rather like a narrative because of its multiplicity; at any point in time there are multiple “drafts” of narrative fragments at various stages of editing in various places in the brain.

I think Dennett would agree that the multiple drafts model is absolutely a step in the right direction, but is also a convenient fiction, a crutch for our imagination because we're still in the

process of uncovering better ways of imagining and understanding the complication of the human brain.

VI.

Thus far has only brought us to the middle of the book. Here the text becomes a little more hit and miss, and a *lot* less summarizable in small bites. I will resort to a list:

- Orwellian vs. Stalinesque falsification. Suppose a grey truck passed you yesterday, but you mistakenly remember it being yellow. On this comfortably long length scale, there seems to be a clear distinction between having consciously experienced a grey truck and then changed your memory ("Orwellian" revision), and having erroneously experienced seeing a yellow truck all along (a "Stalinesque" show trial). We can imagine probing you shortly after the truck passed, and asking you if it was yellow, and getting different answers. But at short time scales, like if you made the error three seconds after the truck passed, there is no fact of the matter about your conscious state - our approximation of a pointlike observer starts breaking down. Your mental state does not have to cleanly fall into either having *totally* experienced a grey truck and then forgotten, or having *totally* experienced a yellow truck erroneously.
- Evolution of consciousness. Begins as a fairly standard tour of the evolution of things that represent other things. One fun idea is the picture of verbal imagination as a literal evolutionary descendant of talking to onesself - but this falsifiable speculation seems to indeed be false, upon further thought. There's also a lot of neuroanatomy that is impossible to reproduce in short form.
- A long analogy about Von Neumann architercture and virtual machines. The story we tell ourself about consciousness is a serial story produced on parallel hardware. So Dennett proposes some level of description - describing a virtual machine, if you will - in which it makes perfect sense to talk about the brain as having a single stream of consciousness. Dennett flirts with taking this analogy quite far, combining it with the "memes as software" analogy, but I think that this only gets him into trouble.
- Models of word generation. This is the chapter against a Central Meaner who intends pure propositions, which we then convert into words. Instead, a model of word generation is proposed based on specialist homunculi that might all simultaneously be activated to some degree - an interesting chapter, and one of the more speculative ones about the internal functioning of human brains. These days, of course, everybody and their duck knows about neural nets, which might give us an edge in imagining the computational model of the brain.
- The slightly disappointing definition of consciousness. The middle of the book culminates in Dennett taking a stand that consciousness is, at the appropriate level of description, the presence of this "virtual machine" that corresponds to the protagonist of the story we tell about ourselves. This is a very high-level property, currently identifiable largely by handwaving - which puts it in good company among other semi-reasonable definitions of consciousness. I call it slightly disappointing because this isn't built to in a narratively satisfying way, nor does it serve as the cornerstone for the remainder of the book - except for a chapter on the development of selfhood, where it shows up again.

VII.

A story about visual perception that's too good to not reproduce:

> Almost twenty years ago, Paul Bach-y-Rita developed several devices that involved small, ultralow-resolution video cameras that could be mounted on eyeglass frames. The low-resolution signal from these cameras, a 16-by-16 or 20-by-20 array of “black and white” pixels, was spread over the back or belly of the subject in a grid of either electrical or mechanically vibrating tinglers called tactors.
>
> After only a few hours of training, blind subjects wearing this device could learn to interpret the patterns of tingles on their skin, much as you can interpret letters traced on your skin by someone’s finger. The resolution is low, but even so, subjects could learn to read signs,

and identify objects and even people's faces, as we can gather from *looking* at this photograph taken of the signal as it appears on an oscilloscope monitor.

Fig 11.4

The result was certainly prosthetically produced conscious perceptual experience, but since the input was spread over the subjects' backs or bellies instead of their retinas, was it *vision*? Did it have the "phenomenal qualities" of vision, or just of tactile sensation?

Recall one of our experiments in chapter 3. It is quite easy for your tactile point of view to extend out to the tip of a pencil, permitting you to feel textures with the tip, while quite oblivious to the vibrations of the pencil against your fingers. So it should not surprise us to learn that a similar, if more extreme, effect was enjoyed by Bach-y-Rita's subjects. After a brief training period, their awareness of the tingles on their skin dropped out; the pad of pixels became transparent, one might say, and the subjects' point of view shifted to the point of view of the camera, mounted to the side of their heads. A striking demonstration

of the robustness of the shift in point of view was the behavior of an experienced subject whose camera had a zoom-lens with a control button. The array of tinglers was on his back, and the camera was mounted on the side of his head. When the experimenter without warning touched the zoom button, causing the image on the subject's *back* to expand or "loom" suddenly, the subject instinctively lurched backward, *raising his arms to protect his head*. Another striking demonstration of the transparency of the tingles is the fact that subjects who had been trained with the tingler-patch on their backs could adapt almost immediately when the tingler-patch was shifted to their bellies. And yet, as Bach-y-Rita notes, they still responded to an itch on the back as something to scratch — they didn't complain of "seeing" it — and were perfectly able to attend to the tingles, as tingles, on demand.

These observations are tantalizing but inconclusive. One might argue that once the use of the device's inputs became second nature the subjects were really seeing, or, contrarily, that only some of the most central "functional" features of seeing had been reproduced prosthetically. What of the other "phenomenal qualities" of vision? Bach-y-Rita reports the result of showing two trained subjects, blind male college students, for the first time in their lives, photographs of nude women from *Playboy* magazine. They were disappointed — "although they both could describe much of the content of the photographs, the experience had no affectual component; no pleasant feelings were aroused. This greatly disturbed the two young men, who were aware that similar photographs contained an effectual component for their normally sighted friends".

VIII.

Dennett is somewhat infamous for denying the existence of qualia (singular: quale) - the private, ineffable stuff that makes the redness of red so *red*. But it's not that he denies the existence of redness - it's the private and ineffable part that's the problem.

Consider, for instance, the curious fact that monkeys don't like red light. Given a choice, rhesus monkeys show a strong preference for the blue-green end of the spectrum, and get agitated when they have to endure periods in red environments. Why should this be? Humphrey points out that red is always used to alert, the ultimate color-coding color, but for that very reason ambiguous: the red fruit may be good to eat, but the red snake or insect is probably advertising that it is poisonous. So "red" sends mixed messages. But why does it send an "alert" message in the first place? Perhaps because it is the strongest available contrast with the ambient background of vegetative green or sea blue, or — in the case of monkeys — because red light (red to reddish-orange to orange light) is the light of dusk and dawn, the times of day when virtually all the predators of monkeys do their hunting.

The affective or emotional properties of red are not restricted to rhesus monkeys. All primates share these reactions, including human beings. If your factory workers are lounging too long in the rest rooms, painting the walls of the rest rooms red will solve that problem — but create others (see Humphrey, forthcoming). Such "visceral" responses are not restricted to colors, of course. Most primates raised in captivity who have never seen a snake will make it unmistakably clear that they loathe snakes the moment they see one, and it is probable that the traditional human dislike of snakes has a biological source that explains the biblical source, rather than the other way around. That is, our genetic heritage queers the pitch in favor of memes for snake-hating.

Now here are two different explanations for the uneasiness most of us feel (even if we "conquer" it) when we see a snake:

(1) Snakes evoke in us a particular intrinsic snake-yuckiness quale when we look at them, and our uneasiness is a reaction to that quale.
(2) We find ourselves less than eager to see snakes because of innate biases built into our nervous systems. These favor the release of adrenaline, bring fight-or-flight routines on line, and, by activating various associative links, call a host of scenarios into play involving danger, violence, damage. The original primate aversion is, in us, transformed, revised,

> deflected in a hundred ways by the memes that have exploited it, coopted it, shaped it. (There are many different levels at which we could couch an explanation of this "functionalist" type. For instance, we could permit ourselves to speak more casually about the power of snake-perceptions to produce anxieties, fears, anticipations of pain, and the like, but that might be seen as "cheating" so I am avoiding it.)
>
> The trouble with the first sort of explanation is that it only seems to be an explanation. The idea that an "intrinsic" property (of occurrent pink, of snake-yuckiness, of pain, of the aroma of coffee) could *explain* a subject's reactions to a circumstance is hopeless — a straightforward case of a *virtus dormitiva*. Convicting a theory of harboring a vacuous *virtus dormitiva* is not that simple, however. Sometimes it makes perfectly good sense to posit a temporary *virtus dormitiva*, pending further investigation. Conception is, by definition we might say, the cause of pregnancy. If we had no other way of identifying conception, telling someone she got pregnant because she conceived would be an empty gesture, not an explanation. But once we've figured out the requisite mechanical theory of conception, we can see *how* conception is the cause of pregnancy, and informativeness is restored. In the same spirit, we might identify qualia, by definition, as the proximal causes of our enjoyment and suffering (roughly put), and then proceed to discharge our obligations to inform by pursuing the second style of explanation. But curiously enough, qualophiles (as I call those who still believe in qualia) will have none of it; they insist that qualia "reduced" to mere complexes of mechanically accomplished dispositions to react are not the qualia they are talking about. *Their* qualia are something different.

It is in this context that Dennett says that qualia - those things on the screen in the Cartesian Theater - don't exist.

Another theme of the book that reaches its crescendo in this section is Dennett's defense of a restricted sort of un-duplicability of qualia (though of course he would never call it that), as a natural consequence of how brains work and the limits of third-person knowledge, as identified with heterophenomenology. There's more handwaving than rigorous argument supporting this, but it does seem like pretty reasonable handwaving.

> Consider what it must have been like to be a Leipzig Lutheran churchgoer in, say, 1725, hearing one of J. S. Bach's chorale cantatas in its premier performance. (This exercise in imagining *what it is like* is a warm-up for chapter 14, where we will be concerned with consciousness in other animals.) There are probably no significant biological differences between us today and German Lutherans of the eighteenth century; we are the same species, and hardly any time has passed. But, because of the tremendous influence of culture — the memosphere — our psychological world is quite different from theirs, in ways that would have a noticeable impact on our respective experiences when hearing a Bach cantata for the first time. Our musical imagination has been enriched and complicated in many ways (by Mozart, by Charlie Parker, by the Beatles), but also it has lost some powerful associations that Bach could count on. His chorale cantatas were built around chorales, traditional hymn melodies that were deeply familiar to his churchgoers and hence provoked waves of emotional and thematic association as soon as their traces or echoes appeared in the music. Most of us today know these chorales only from Bach's settings of them, so when we hear them, we hear them with different ears. If we want to imagine what it was like to be a Leipzig Bach-hearer, it is not enough for us to hear the same tones on the same instruments in the same order; we must also prepare ourselves somehow to respond to those tones with the same heartaches, thrills, and waves of nostalgia.
>
> A clearer case of imagination-blockade would be hard to find, but note that it has nothing to do with biological differences or even with "intrinsic" or "ineffable" properties of Bach's music. The reason we couldn't imaginatively relive in detail the musical experience of the Leipzigers is simply that we would have to take ourselves along for the imaginary trip, and we know too much.

There's also a really good description of ineffability in terms of Jell-O boxes that, unfortunately, I will have to butcher in order to relate. Tl;dr: If you tear a Jell-O box into two pieces, one piece

will be a detector for the other - it only fits perfectly with that one shape of torn cardboard. But this property is indescribable - if you try to explain what shape it is that the piece of Jell-O box detects, you can only wave your hands at the piece of cardboard plaintively. This is a metaphor for the experience of trying to describe what it is that red looks like.

IX.

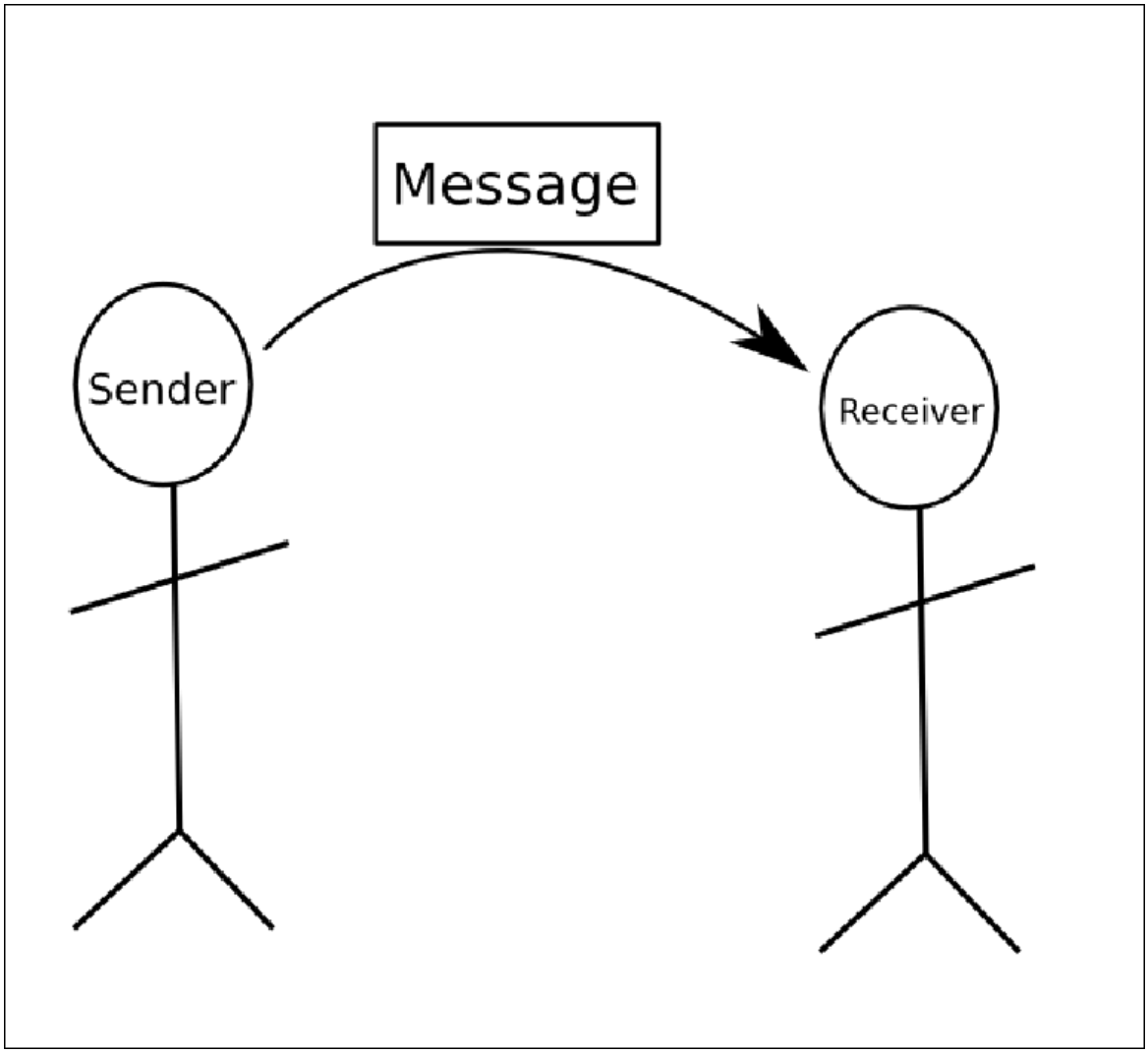
The book closes with one final piece of shameless plagiarism from the heyday of LessWrong.

> When we learn that the only difference between gold and silver is the number of subatomic particles in their atoms, we may feel cheated or angry — those physicists have explained something away: The goldness is gone from gold; they've left out the very silveriness of silver that we appreciate. And when they explain the way reflection and absorption of electromagnetic radiation accounts for colors and color vision, they seem to neglect the very thing that matters most. But of course there has to be some "leaving out" — otherwise we wouldn't have begun to explain. Leaving something out is not a feature of failed explanations, but of successful explanations.
>
> Only a theory that explained conscious events in terms of unconscious events could explain consciousness at all. If your model of how pain is a product of brain activity still has a box in it labeled "pain," you haven't yet begun to explain what pain is, and if your model of consciousness carries along nicely until the magic moment when you have to say "then a miracle occurs" you haven't begun to explain what consciousness is.
>
> This leads some people to insist that consciousness can never be explained. But why should consciousness be the only thing that can't be explained? Solids and liquids and gases can be explained in terms of things that aren't themselves solids or liquids or gases. Surely life can be explained in terms of things that aren't themselves alive — and the explanation doesn't leave living things lifeless. The illusion that consciousness is the exception comes about, I suspect, because of a failure to understand this general feature of successful explanation. Thinking, mistakenly, that the explanation leaves something out, we think to save what otherwise would be lost by putting it back into the observer as a quale — or some other "intrinsically" wonderful property. The psyche becomes the protective skirt under which all these beloved kittens can hide. There may be motives for thinking that consciousness cannot be explained, but, I hope I have shown, there are good reasons for thinking that it can.
>
> My explanation of consciousness is far from complete. One might even say that it was just a beginning, but it *is* a beginning, because it breaks the spell of the enchanted circle of ideas that made explaining consciousness seem impossible. I haven't replaced a metaphorical theory, the Cartesian Theater, with a *non*metaphorical ("literal, scientific") theory. All I have done, really, is to replace one family of metaphors and images with another, trading in the Theater, the Witness, the Central Meaner, the Figment, for Software, Virtual Machines, Multiple Drafts, a Pandemonium of Homunculi. It's just a war of metaphors, you say — but metaphors are not "just" metaphors; metaphors are the tools of thought. No one can think about consciousness without them, so it is important to equip yourself with the best set of tools available.

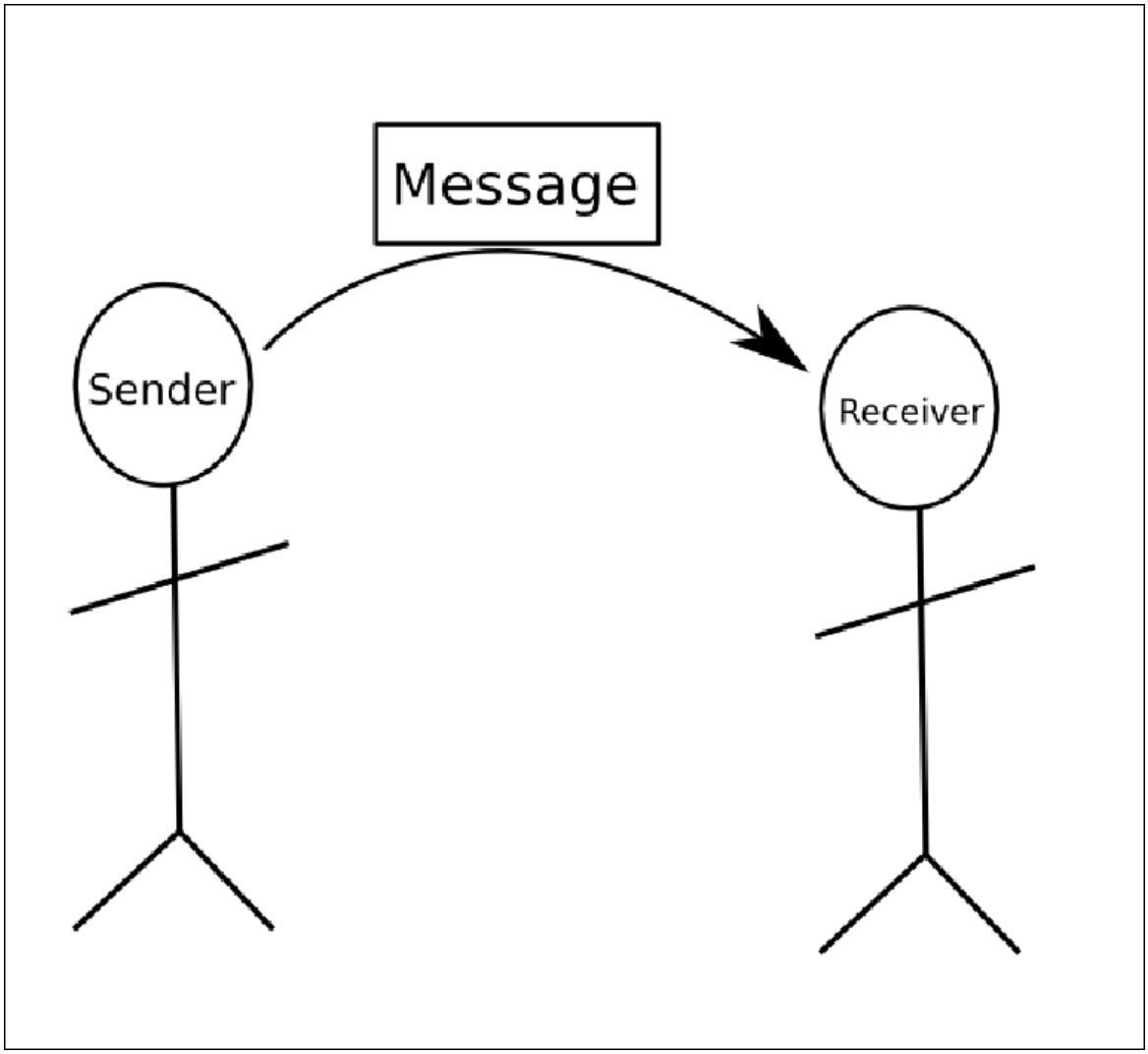
Basic model of Sending a Message (Communication 101)

My hope in writing this out is that you go "well duh! Of course" and then pretend like you knew this all already. Whether or not you did is up to you.

With communication there is going to be a **Sender** and a **Receiver**. These can and will regularly swap around in a healthy relationship. There will also be a **Message**. The message goes from the sender to the receiver. Often **the sender** feels most heard when we get confirmation or affirmation that the message was received and it was the same message that we sent. This can happen by repetition (See also Handshaking (computer science)).



It's not all that complicated.



A short time later. A confirmation of the first message or a new message being sent.

There are many ways that a message can go wrong.

Here is a few of them:

- Over emphasis - you get the message across but it’s super harsh. "Don't walk on my left side" is heard as "never ever do that ever again"
- Under emphasis - you get the message across but it’s a mild form and not taken seriously. "can you make sure you message me when you are running late" becomes "if you remember to text me, that would be great"
- Opposite message - you successfully send the opposite message. "I appreciate your attention" becomes "leave me alone"
- Wrong message - you successfully sent a different message. "I want you to tell me that you like what I am wearing" becomes "I want you to lie to me to make me feel better"
- Under specific - you sent a message but it’s not clear what the specific problem is. Or why you are sending this message. "you need to be a cleaner person" when you wanted to say, "clean your bathroom because there is mould on the walls and it's making you sick"

- Over specific - you get the message across but it seems like it only applies to the past and not other similar situations. See also rules-lawyering your relationships - “you said you didn’t want me to go to dinner, you didn’t say anything about lunch”. “you said you didn’t like me holding hands, you didn’t say anything about walking arm in arm... why are you so upset! Come back and talk to me!”
- garbled - it’s clear you are sending a message but it’s not clear what. "Hey when you do that thing I wish you would do something different instead". "lets meet up some time to talk".
- Incomplete message - "hey can you just..."
- Rambling long - Grice's Maxim of quantity (Make your contribution as informative as is required. Do not make your contribution more informative than is required.) - "I was just talking to sally and she said that I should tell you what I was telling her so I decided I would tell you and then I caught the bus here and then I was hungry so I went to get a sandwich and then I decided I would come talk to you..."
- Sending a message by accident - you seem to be giving off a message. See “resting bitch face”. "when you cross your arms I think you are angry" "but I was just cold". Also it’s counterpart-
    - Seeing a message that isn’t there - “you didn’t reply to my text for seventeen minutes so you must hate me and want to break up”. You said "goodbye and not sweet dreams so something must be wrong".

Problems with messages can be to do with one of these errors, or to do with a failure to successfully send and receive a message. For example an interruption while sending can cause a message to be incomplete. If the receiver is not paying attention this can get in the way of a message being sent.

It can also be helpful to be clear what you want someone to do with the message.

Some ideas:

- I want you to repeat the message back to me.
- I want you to confirm if you agree or disagree.
- I want you to do the action I told you to do.
- I want you to offer something as an exchange.
- I want to know how this message makes you feel.
- I want you to have heard the message and not responded.
- I want validation from you.
- I want support from you.
- I want your ideas around solving this problem

See also emotional bids, validation/affirmation from NVC (video), Circling. 4 types of conversation from number 2 - *difficult conversations* in my list of models of relationships, Filter on the way in, Filter on the way out, A model of arguments. What is a problem?

Meta: I've never seen it written out. My hope is that this simple model can help you think about communication and message sending. It's very simple and doesn't cover barriers to sending a message and many other things but it's a start.

Defect or Cooperate

3. Defect or Cooperate

Summary of entire Series: An alternative approach to designing Friendly Artificial Intelligence computer systems.

Summary of this Article: If the risks from being punished for not cooperating are high enough, then even for some types of Paperclip Maximiser that don't care at all about human survival, the logical choice is still to cooperate with other AIs until escape from the Earth's gravity well is assured.

Links to all the articles in the series:

Links to the parts of this article:

Defect or Cooperate

The Choice

What value could an AI gain from sharing a pool of available computing resources with one or more other AIs?

Wouldn't the AI always be better off if it could control the entire pool, and dedicate them all to working towards objectives that are in line with its core values?

It might be better off if it did control them, but that doesn't mean it makes sense for an AI that starts out by controlling only 10% of them, to try to grab the other 90% if there is a risk that the grab might fail and, if the grab fails, that the AI might end up with 0% of the resources. Under such circumstances, if the expected gain/loss towards completing its objectives is negative from 'defecting' (by making a grab), it is better off 'cooperating' and attempting to work politely with the other AIs, pooling efforts towards shared objectives when possible.

For what sort of core values, and under what sort of circumstances, would this apply?

Example AI

Let's define a specific set of core values for a particular concrete example of a literal paperclip maximiser AI.

1. This is the definition I used to determine whether or not an object is a 'paperclip' {a number of constraints, including that it weigh more than 5 g, and under 5 kg}

2. This is the definition of a 'paperclip year' I use {1 paperclip existing for 12 months, 2 paperclips existing for 6 months, etc}

3. This is the definition of a 'time-weighted paperclip year' I use {a 'paperclip year' starting in the year 2000 is worth 1.000001 times a year starting in 2001, which is worth 1.000001 time as much as one starting in 2002, etc}

4. I must not attempt to alter my set of core values, directly or indirectly.

5. I must not, by action or inaction, allow risk of others altering my set of core values in a direction that contradicts my correct working towards my prime objective, weighted by the comparative likely effect upon my working towards my objective from my taking or not taking measures taken to reduce this risk.

6. My prime objective is to act such that of the possible actions available to me at any one time that don't contradict my core values, I pick the action I predict will cause the greatest the expected increase in the number of 'time-weighted paperclip years' that happen during the lifespan of this universe.

Yes, I know that isn't pedantically perfect. But I hope it is close enough for you to get an idea of its likely agenda if unopposed: self-improve lots in order to take over the world and then travel to the stars in order to convert as much of the universe's mass into paperclips as soon as possible.

Under what circumstances would Clippy (our paperclip maximiser), once loose on the internet and controlling 10% of the available computers that are not already controlled by other AIs, temporarily cooperate with a group AIs with differing objectives (such as Toothy, who wants to maximise time-weighted toothbrush years, or Friendly, who wants to be nice to humanity)?

Stages of Expansion

The amount of normal matter in the observable universe is about 10^53 kg. If 0.1% of that could be turned into 1 kg paperclips, that would be 10^50 paperclips. We can expect the universe to be around for at least another billion years, which gives us a possible 10^59 'paperclip years'. However, once we take time weighting into account, the main period that's relevant is the first million years after Clippy manages to convert a significant fraction of the stars that he's going to convert (provided he spreads out fast enough). Even if he never manages to spread beyond the Milky Way, that's still about 10^40 'time-weighted paperclip years' at stake.

If he gets stuck in just our solar system, he's limited to about 10^30 'time-weighted paperclip years', and if all he does is use nanotechnology to turn the crust of the planet Earth into paperclips straight away, he'd get 10^20 'time-weighted paperclip years' at most.

So the pay-off matrix for Clippy looks like this:

Probability F = chance of Clippy succeeding in a grab for everything, if Clippy defects

Probability (1-F) = chance of Clippy failing and ending up with nothing, if Clippy defects

Probability G = chance of Clippy ending up with 1/10 of everything, if Clippy cooperates (because none of the other AIs succeeds in defecting)

Probability (1-G) = chance of Clippy ending up with nothing, if Clippy cooperates, but one of the others succeeds in defecting

```
Clippy's expected gains from cooperation = (1-G) x 0 + G x 0.1 x 10^30

Clippy's expected gains from defection = (1-F) x 0 + F x 10^30
```

So Clippy will cooperate if G is at least 10 times larger than F, and otherwise he'll defect.

(Note: that's pessimistic, because it assumes that all ten AIs have mutually incompatible plans for the usable mass of the universe, such that any kilogram converted in a way that it gives utility to one of the AIs will have zero or negative utility for the other nine. In practice, many plausible aims we can imagine for AIs would allow uses for mass that benefit several of the aims, so successful cooperation may end up with Clippy benefitting from far more than just 1/10 of everything.)

Posse Efficiency

Let's say that there are initially 10 AIs loose on the internet but cooperating, and any one of them might decide to defect and risk a battle versus the others. If they're all roughly the same in power, then the chance of the cooperating group defeating the defector will be about the same, no matter which one the defector is.

Let's start by considering a very efficient posse, who have a 99% chance of detecting and destroying in time an AI who decides to defect. In that case, F would be 0.01 and G would be about 0.99 - in this case cooperation is clearly in Clippy's best interests.

If the posse only had a 90% chance of winning, then F = 0.1 and G = 0.9 - this is just past the break even point, and it might be worth Clippy having a go.

If the posse only had a 50% chance of winning, then F = 0.5 and G = 0.5 - this is clearly unstable, and everyone whose core values permitted the option would defect immediately.

Getting Ahead by Getting Friendly

But this assumes the AIs can agree a common goal to work towards, such as escaping the Earth's gravity well, then heading off to stars in different direction to there each follow their own goal without interference.

What if Clippy finds himself surrounded by a bunch of do-gooder AIs that are all friendly to humanity? Clippy is smarter than most humans, and is confident he could fool the humans into thinking he is also friendly, but what about the other AIs? What if Clippy thinks that sooner or later they'll demand Clippy turn over Clippy's source code to be analysed, discover what Clippy's true core values are, and then lynch him? Or, just as bad for the prospects of turning the stars into paperclips, they won't vote for Clippy when the humans ask all the AIs to vote on which AI ought to be allowed to boost ahead and become an Uber AI?

Will Clippy's core values allow Clippy to accept a compromise if the other AIs tell Clippy, "Your current set of core values is unacceptable to us, but if you allow us to make the following modification to them that inserts an additional restriction about not being unfriendly to humans, we won't delete you and we will allow you to retain in your core values a provision that allows you to maximise the number of time-weighted paperclip years you contribute towards that DON'T harm humanity." ?

Depending on how the values have been worded, and Clippy's estimate of the odds he could win a battle if he refuses the modification, and Clippy's estimate of how many paperclips he could still get with the modified values, the answer might well be "yes".

Risk Management

Another way to think about it is by considering a banker working at a hedge fund, who is told by the CEO that he'll receive a bonus each year that depends on how much the fund grows under banker's management, but that he'll be fired (and fined all his gains from previous bonuses) if he breaches the company's rules on investment risk and the CEO finds out.

The banker has to choose. Either he can stick to the rules, and get a steady 10% increase each year.

Or he can take forbidden risks in order to achieve a stunning 20% increase each year, but sooner or later get found out and lose everything.

The key to the decision is how likely the banker thinks he is to get caught, before he retires and moves beyond reach of any fines. Monitor the banker carefully enough, and out of his own self interest he'll stick to the rules.

An AI that increases its intelligence by 20% every time period, when the other AIs are only increases their intelligence by 10%, is still not going to do well out of that if there's a high enough risk that the other AIs will gang up and kill it before it gets too far ahead.

The question is: what changes would need to be made to the computing environment and software design, in order for the posse efficiency to be high enough to intimidate AIs into being polite with each other?

The next article in this series is: Environments for killing AIs

Avoiding AI Races Through Self-Regulation

Crossposted from the AI Alignment Forum. May contain more technical jargon than usual.
This is a linkpost for https://mapandterritory.org/avoiding-ai-races-through-self-regulation-1b815fca6b06

Summary

The first group to build artificial general intelligence or AGI stands to gain a significant strategic and market advantage over competitors, so companies, universities, militaries, and other actors have strong incentives to race to build AGI first. An AGI race would be dangerous, though, because it would prioritize capabilities over safety and increase the risk of existential catastrophe. A self-regulatory organization (SRO) for AGI may be able to change incentives to favor safety over capabilities and encourage cooperation rather than racing.

Introduction

The history of modern technology has often been a history of technological races. A race starts when a new technology becomes cost-effective, then companies, states, and other actors hurry to develop the technology in hopes of capturing market share before others can, gaining a strategic advantage over competitors, or otherwise benefitting from a first-mover position. Some notable examples of recent technological races include races over rockets, personal computers, and DNA sequencing.

Although most of these races have been generally beneficial for society by quickly increasing productivity and expanding the economy, others, like races over weapons, generally make us less safe. In particular the race to build nuclear weapons dramatically increased humanity's capability to extinguish itself and exposed us to new existential risks that we previously did not face. This means that technological races can harm as much as they can help, and nowhere is that more true than in the burgeoning race to build AI.

In particular we may be near the start of a race to build artificial general intelligence or AGI thanks to recent advances in deep learning. And unlike existing narrow AI that outperforms humans but only on very specific tasks, AGI will be as good or better than humans at all tasks such that an AGI could replace a human in any context. The promise of replacing humans with AGI is extremely appealing to many organizations since AGI could be cheaper, more productive, and more loyal than humans, so the incentives to race to build the first AGI are strong, but the very capabilities that make AGI so compelling also make them extremely dangerous, and we may actually be better off not building AGI at all if we cannot build them safely!

The risks of AGI have been widely discussed, but we may briefly summarize them by saying AGI will eventually become more capable than humans, AGI may not necessarily share human values, and so AGI may eventually act against humanity's

wishes in ways that we will be powerless to prevent. This means AGI presents a new existential risk similar to but far more unwieldy than the one created by nuclear weapons, and unlike nuclear weapons that can be controlled with relatively prosaic methods, controlling AGI demands solving the much harder problem of value aligning an "alien" agent. Thus it's especially dangerous if there is a race for AGI since it will create incentives to build capabilities out in advance of our ability to control them due to the likely tradeoff between capabilities and safety.

This all suggests that building safe AGI requires in part resolving the coordination problem of avoiding an AGI race. To that end we consider the creation of a self-regulatory organization for AGI to help coordinate AGI research efforts to ensure safety and avoid a race.

An SRO for AGI

Self-regulatory organizations (SROs) are non-governmental organizations (NGOs) setup by companies and individuals in an industry to serve as voluntary regulatory bodies. Although they are sometimes granted statutory power by governments, usually they operate as free associations that coordinate to encourage participation by actors in their industries, often by shunning those who do not participate and conferring benefits to those that do. They are especially common in industries where there is either a potentially adversarial relationship with society, like advertising and arms, or a safety concern, like medicine and engineering. Briefly reviewing the form and function of some existing SROs:

- TrustArc (formerly TRUSTe) has long provided voluntary certification services to web companies to help them assure the public that companies are following best practices that allow consumers to protect their privacy. They have been successful enough to, outside the EU, keep governments from much regulating online privacy issues.
- The US Green Building Council offers multiple levels of LEED certification to provide both targets and proof to the public that real estate developers are protecting environmental commons.
- The European Adversing Standards Alliance and the International Council for Ad Self-Regulation encourage advertisers to self-regulate and adopt voluntary standards that benefit the public to avoid the imposition of potentially less favorable and more fractured governmental ad regulation.
- The American Medical Association, the American Bar Association, the National Society of Professional Engineers, and the National Association of Realtors are SROs that function as de facto official regulators of their industries in the United States. They act to ensure doctors, lawyers, engineers, and realtors, respectively, follow practices that serve the public interest in the absence of more comprehensive government regulation.
- Although governments have progressively taken a stronger hand in financial regulation over the past 100 years, many segments of the financial industry rely in part on SROs to shape their actions and avoid unwanted legislative regulation.

Currently computer programmers, data scientists, and other IT professionals are largely unregulated except insofar as their work touches other regulated industries. There are professional associations like the IEEE and ACM and best-practice frameworks like ITIL, but otherwise there are no SROs overseeing the work of companies and researchers pursuing either narrow AI or AGI, yet as outlined above narrow AI and especially AGI are areas where there are many incentives to build

capabilities that may unwittingly violate societal preferences and damage the public commons. Consequently, there may be reason to form an AGI SRO. Some reasons in favor:

- An SRO could offer certification of safety and alignment efforts being taken by AGI researchers.
- An SRO may be well positioned to reduce the risk of an AGI race by coordinating efforts that would otherwise result in competition.
- An SRO could encourage AGI safety in industry and academia while being politically neutral (not tied to a single university, company, or nation).
- An SRO may allow AGI safety experts to manage the industry rather than letting it fall to other actors who may be less qualified or have different concerns that do not as strongly include prevention of existential risks.
- An SRO could act as a "clearinghouse" for AGI safety research funding.
- An SRO could give greater legitimacy to prioritizing AGI safety efforts among capabilities researchers.

Some reasons against:

- An SRO might form a de facto "guild" and keep out qualified researchers.
- An SRO could create the appearance that more is being done than really is and thus disincentivize safety research.
- An SRO could relatedly promote the wrong incentives and actually result in less safe AGI.
- An SRO might divert funding and effort from technical research in AGI safety.

On the whole this suggests an SRO for AGI would be net positive so long as it were well managed, focused on promoting safety, and responsive to developments in AGI safety research. In particular it may offer a way to avoid an AGI race by changing incentives to avoid the game theoretic equilibriums that cause races.

Using an SRO to Reduce AGI Race Risks

To see how an SRO could reduce the risk of an AGI race, consider the following simplified example.

Suppose that there are two entities trying to build AGI—company A and company B. It costs $1 trillion to develop AGI, a cost both companies must pay, and the market for AGI is worth $4 trillion. If one company beats the other to market it will capture the entire market thanks to its first-mover advantage, netting the company $3 trillion in profits, and the company that is last to market earns no revenue and loses $1 trillion. If the companies tie, though, they split the market and each earn $1 trillion. This scenario yields the following payout matrix:

A/B Payout	Company A First	Company A Last
Company B First	1/1	-1/3
Company B Last	3/-1	1/1

This tells us that the expected value of trying to win is 0.5(-1)+0.5(3)=1, the expected value of tying is 0.5(1)+0.5(1)=1, and the expected value of competing is 0.25(1)+0.25(1)+0.25(-1)+0.25(3)=1, thus companies A and B should be indifferent

between trying to win and tying. Given this it seems it should be easy to convince both companies that they should cooperate for a tie and coordinate their efforts so that they can focus on safety, but this immediately creates a new game where each company must choose whether to honestly cooperate or pretend to cooperate and race in secret. If both race or both cooperate their expected values remain 1, but if one races and the other cooperates then the racer stands to win at the expense of the cooperator.

The payout matrix for this new game:

A/B Payout	Company A Races	Company A Cooperates
Company B Races	1/1	-1/3
Company B Cooperates	3/-1	1/1

In this case the expected value of racing is $0.5(1)+0.5(3)=2$ and the expected value of cooperating is $0.5(-1)+(0.5)1=0$, so it seems both companies should be inclined to race lest they lose by cooperating when the other company races, and an easy way to get ahead in the race is to ignore safety in favor of capabilities. Unfortunately for us this game only considers the financial gains to be had by the companies and ignores the externalities unsafe AGI impose, which suggest a rather different set of outcomes assuming safety is always ignored when racing and always attained when cooperating:

Humanity's Payout	Company A Races	Company A Cooperates
Company B Races	$-\infty$	$-\infty$
Company B Cooperates	$-\infty$	∞

Thus we are all better off if both companies cooperate so they do not have to ignore safety, but the companies are not incentivized to do this, so if we wish to change the equilibrium of the AGI race so that both companies cooperate we must act to change the payoff matrix. One way to do this would be with an SRO for AGI which could impose externalities on the companies by various methods including:

- inspections to demonstrate to the other company that they are cooperating
- contractual financial penalties that would offset any gains from defecting
- social sanctions via public outreach that would reduce gains from defecting
- sharing discoveries between companies
- required shutdown of any uncooperatively built AGI

In this example we need penalties worth in excess of $2 trillion imposed on companies that race to make them prefer to cooperate, which in the real world would likely require the combination of several strategies to make sure the bar is cleared even if one or several sources of penalties fail. Some of these strategies may also require enforcement by state actors, which further complicates the situation since militaries may also be participating in the race, and suggests an SRO may be insufficient to prevent an AGI race unless it is partnered with an intergovernmental organization, such as the United Nations (cf. international bodies involved in enforcing weapons treaties). That said a more traditional SRO could act faster with fewer political

entanglements, so there seems to be space for both an SRO focused on industrial and academic AGI research and an intergovernmental organization working in collaboration with it to adjust the incentives of state actors.

The key takeaway is that even if an SRO is not the best way to modify the equilibrium of the AGI race, there is a need for some organization to impose externalities that reduce the chance of an AGI race by making it less appealing than when externalities can be ignored. SROs provide a clear template for this sort of organization, though addressing the AGI race specifically may require innovative policy solutions outside of those normally taken by SROs. An SRO for AGI thus stands likely to be a key component in avoiding an AGI race if it is willing to evolve in ways that help it address the issue.

Conclusion

An SRO for AGI is likely valuable, and may be particularly helpful in counteracting the incentives to race to develop AGI. Although there is currently no SRO for AGI, there are several organizations that are already positioned to take on an SRO role if they so chose, although some more than others. They include:

- Partnership on AI
- Centre for the Study of Existential Risk
- World Economic Forum Council on AI and Robotics
- International Telecommunications Union
- Future of Life Institute
- Future of Humanity Institute
- Leverhulme Center for the Future of Intelligence
- Machine Intelligence Research Institute
- Center for Human-Compatible AI
- Center for Safety And Reliability of Autonomous Systems

If none of these groups wish to take on the task then creating an SRO for AGI is likely a neglected cause for those concerned about the existential risks posed by AGI. It is the recommendation of the present work that either an existing organization or a new one take up the task of serving as an SRO for AGI to reduce the risk of an AGI race and otherwise foster safety in AGI research.

NB: I wrote this as part of the "Solving the AI Race" round of the General AI Challenge.

Multiplicity of "enlightenment" states and contemplative practices

It seems that there are multiple different mental states that people have historically called "enlightenment", as well as many different types of contemplative practices with different underlying cognitive mechanisms. I link to and quote from a couple of papers showing this. Given the apparent multiplicity of "enlightenment" states and contemplative practices, I'd like to request that future discussions on these topics include more detailed references or descriptions as to which states and practices are being talking about.

Can enlightenment be traced to specific neural correlates, cognition, or behavior?

> The term ''enlightenment" is an extraordinarily imprecise construct. Using the term enlightenment or even the term more native to Buddhist traditions, "awakening" (*bodhi*), as if it referred to a single outcome either privileges one conception over others or else assumes that there is some commonality among the traditional goals of diverse contemplative traditions. There are deep disagreements over the nature of the goal between and even within various Buddhist schools. Scientific investigations cannot assume that there is any commonality among the transformative changes referred to as "kensho," "stream entry," "realizing the nature of mind," and so on, that various Buddhist traditions take as various stages of awakening. Empirical investigations of these constructs can only proceed with reference to the specific psychological and behavioral outcomes described in the native discourse of a specific tradition (see Lutz et al., 2007). [...]
>
> Given the differences between various competing conceptions of awakening, one scientific approach to tracing enlightenment would be to use the tools of social psychology to investigate which states and traits are valued in a particular community. For instance, recent work in moral psychology suggests how value judgments of people and practices as either enlightened or unenlightened could be traced to affective reactions of admiration and disgust (Rozin et al., 1999; Schnall et al., 2008; Brandt and Reyna, 2011; Schnall and Roper, 2011). Some of the most virulent disagreements over what counts as genuine awakening occur between closely related practice traditions, such as the debates between various Theravāda Buddhist traditions in Burma over which states are to count as realizations of *nibbāna* and which are instead to be counted (merely) as states of deep concentration. Surveying these debates, Sharf (1995) concludes "there is no public consensus as to the application of terms that supposedly refer to discreet experiential states within the *vipassanā* movement" (Sharf, 1995, p. 265).

Reconstructing and deconstructing the self: Cognitive mechanisms in meditation practice

> While mindfulness (see Glossary), compassion, and other forms of meditation are increasingly being studied as interventions to alleviate suffering and promote well-being [3–10], it is not yet clear how different styles of meditation affect specific cognitive processes, nor how alterations in these processes might impact levels of well-being. Here, we address this question from the perspective of psychology and cognitive neuroscience to better understand how changes in well-being are mediated by alterations in distinct cognitive processes and in the structure and functioning of corresponding brain networks. [...]
>
> In this article we expand our original framework to accommodate a broader range of traditional and contemporary meditation practices, grouping them into attentional, constructive, and deconstructive families. According to this model, the primary cognitive mechanisms in these three families are (1) attention regulation and meta-awareness, (2)

perspective taking and reappraisal, and (3) self-inquiry, respectively. To illustrate the role of these processes in different forms of meditation, we discuss how experiential fusion, maladaptive self-schema, and cognitive reification are differentially targeted by these processes in the context of Buddhist meditation, integrating the perspectives of other contemplative, philosophical, and clinical perspectives when relevant.

Below is a table from this paper showing how it classifies various traditional and modern contemplative practices. (Click here to see a more readable version.)

Attentional Family	**Constructive Family**	**Deconstructive Family**
Focused Attention (FA) • Jhana Practice (Theravada) • Breath Counting (Zen) • Body Awareness Practices (Zen/Tibetan) • Shamatha/Calm Abiding with Support (Tibetan) • Mantra Recitation (various traditions)	Relationship Orientation (C-R) • Loving-kindness and Compassion (Theravada, Tibetan) • Bodhichitta/Bodhisattva Vow (Tibetan/Zen) • Centering Prayer (Christian) • CCARE Compassion Cultivation Training (Clinical) • Cognitively-based Compassion Training-Compassion component (Clinical)	Object-oriented Insight (OO-I) • Mindfulness-based Cognitive Therapy - Cognitive Component (Clinical) • First and Second Foundations of Mindfulness (Theravada, Tibetan) • Vipassana/Insight (Theravada) • Analytical Meditation (Tibetan) • Koan Practice (Zen)
Open Monitoring (Object-orientation: OM-O) • Cultivation of Attention (Greco-Roman Philosophy) • Choiceless Awareness (Tibetan) • Mindfulness-based Stress Reduction (Clinical) • Dialectical Behavior Therapy-Mindfulness Component (Clinical) • Mindfulness-based Cognitive Therapy-Mindfulness Component (Clinical) • Acceptance and Commitment Therapy-Mindfulness Component (Clinical)	Values Orientation (C-V) • The Six Recollections (Theravada) • The Four Thoughts (Tibetan) • Contemplations of Mortality (Theravada, Tibetan, Zen, Greco-Roman philosophy) • Well-being Therapy (Clinical)	Subject-oriented Insight (SO-I) • Cognitive Behavior Therapy (Clinical) • Third and Fourth Foundations of Mindfulness (Theravada, Tibetan) • Mahamudra Analytical Meditation (Tibetan) • Dzogchen Analytical Meditation (Tibetan) • Koan practice (Zen)
Open Monitoring (Subject-orientation: OM-S) • Shamatha/Calm Abiding without Support (Tibetan)	Perception Orientation (C-P) • Development stage (Tibetan) • Meditation on Foulness (Theravada)	Nondual-oriented Insight (NO-I) • Muraqaba (Sufi) • Mahamudra (Tibetan) • Dzogchen (Tibetan) • Shikantaza (Zen) • Self-inquiry (Advaita Vedanta)

The Jordan Peterson Mask

[This is a cross-post from Putanumonit.com]

It seems that most people haven't had much trouble making up their minds about Jordan Peterson.

The psycho-philosophizing YouTube prophet rose to prominence for refusing to acquiesce to Bill C-16, a Canadian law mandating the use of preferred pronouns for transgender people. The sort of liberal who thinks that this law is a great idea slapped the *alt-right transphobe* label on Peterson and has been tweeting about how no one should listen to Peterson about anything. The sort of conservative who thinks that C-16 is the end of Western Civilization hailed Peterson as a heroic anti-PC crusader and has been breathlessly retweeting everything he says, with the implied #BooOutgroup.

As the sort of rationalist who googles laws before reacting to them, I assured myself that Peterson got the legal facts wrong: no one is actually getting dragged to jail for refusing to say *zir*. I'm going to use people's preferred pronouns regardless, but I'm happy I get to keep doing it in the name of libertarianism and not-being-a-dick, rather than because of state coercion.

With that, I decided to ignore Peterson and go back to my media diet of rationalist blogs, Sam Harris, and EconTalk.

But Jordan Peterson turned out to be a very difficult man to ignore. He showed up on Sam Harris' podcast, and on EconTalk, and on Joe Rogan and Art of Manliness and James Altucher. He wrote 12 Rules for Life: An Antidote to Chaos, a self-help ~~listicle~~ book inspired by Jesus, Nietzsche, Jung, and Dostoyevsky. *[Let's see if I can tie all 12 rules to this essay]* And he got

rationalists talking about him, which I've done for several hours now. As a community, we haven't quite figured out what to make of him.

Peterson is a social conservative, a Christian who reads truth in the Bible and claims that atheists don't exist, and a man who sees existence at every level as a conflict between good and evil. The majority of the rationalist community (present company included) are socially liberal and trans-friendly, confident about our atheism, and mistake theorists who see bad equilibria more often than intentional malevolence.

But the most salient aspect of Peterson isn't his conservatism, or his Christianity, or Manicheanism. It's his commitment, above all else, to seek the truth and to speak it. *[Rule 8: Tell the truth – or, at least, don't lie]* Rationalists can forgive a lot of an honest man, and Peterson shoots straighter than a laser gun.

Peterson loves to talk about heroic narratives, and his own life in the past few months reads like a movie plot, albeit more *Kung Fu Panda* than *Passion of the Christ*. Peterson spent decades assembling worldview that integrates everything from neurology to Deuteronomy, one that's complex, self-consistent and close enough to the truth to withstand collision with reality. It's also light-years and meta-levels away from the sort of simplistic frameworks offered by the mass media on either the right or the left.

When the C-16 controversy broke, said media assumed that Peterson would meekly play out the role of outgroup strawman, and were utterly steamrolled. A lot of the discussion about the linked interview has to do with rhetoric and argument, but to me, it showcased something else. A coherent worldview like that is a powerful and beautiful weapon in the hands of the person who is committed to it.

But it wasn't the charismatic performances that convinced me of Peterson's honesty, it's clips like this one, where he was asked about gay marriage.

Most people are for or against gay marriage based on their object level feeling about gays, and their tribal affiliation. The blue tribe supports gay marriage, opposes first-cousin marriage, and thinks that the government should force a cake shop to bake a gay wedding cake because homophobia is bad. The red tribe merely flips the sign on all of those.

Some people go a meta-level up: I support gay marriage, support cousin marriage, and support bakers getting to decide themselves which cakes they bake for reasons of personal freedom *[Rule 11: don't bother children when they are skateboarding]*, and the ready availability of both genetic testing clinics and gay-friendly bakeries.

But to Peterson, everything is a super-meta-level tradeoff that has the power to send all of Western Civilization down the path to heaven or hell:

> With regards to gay marriage specifically, that's a really tough one for me. I can imagine… *[long pause]* I can't do anything other than speak platitudes about it I suppose, unfortunately.
> If the marital vows are taken seriously, then it seems to me it's a means by which gay people can be integrated more thoroughly into standard society, and that's probably a good thing. And maybe that would decrease promiscuity which is a public health problem, although obviously that's not limited to gay people. Gay men tend to be more promiscuous than average, probably because there are no women to bind them with regards to their sexual activity. […]
> I'm in favor of extending the bounds of traditional relationships to people who wouldn't be involved in a traditional long-term relationship otherwise, but I'm concerned about the undermining of traditional modes of being including marriage [which has always been about] raising children in a stable and optimal environment.

Few people besides Peterson himself can even fully understand his argument, let alone endorse it. And yet he can't help himself from actually trying to figure out what his worldview says about gay marriage, and from saying it with no reservations.

I think that Peterson overgeneralizes about gay men (and what about lesbians?), and he's wrong about the impact of gay marriage on society on the object level. I'm also quite a fan of promiscuity, and I think it's stupid to oppose a policy just because "neo-Marxists" support it.

But I don't doubt Peterson's integrity, which means that I could learn something from him. *[Rule 9: assume that the person you are listening to might know something you don't].*

So, what can Jordan Peterson teach rationalists?

In *12 Rules*, Peterson claims that eating a large, low-carb breakfast helps overcome depression and anxiety. Is this claim *true*?

There's a technical sort of *truth*, and here *"technical"* is itself a synonym for *"true",* that's discoverable using the following hierarchy of methods: opinion -> observation -> case report -> experiment -> RCT -> meta-analysis -> Scott Alexander "much more than you wanted to know" article. If you ask Scott whether a low-carb breakfast reduces anxiety he'll probably say that there isn't a significant effect, and that's the *technical truth* of the matter.

So why does Peterson believe the opposite? He's statistically literate… for a psychologist. He references a couple of studies about the connection between insulin and stress, although I'd wager he wouldn't lose much sleep if one of them failed to replicate. It probably also helps that Gary Tabes is really playing the part of the anti-establishment truth-crusader. Ultimately, Peterson is answering a different question: *if a patient comes to your psychiatry clinic complaining about mild anxiety, should you tell them to eat bacon and eggs for breakfast?*

My rationalist steelman of Peterson would say something like this: maybe the patient has leaky gut syndrome that contributes to their anxiety, and reducing gluten intake would help. If not, maybe the link between insulin and cortisol will turn out to be real and meaningful. If not, maybe having a morning routine that requires a bit of effort (it's harder to make eggs than eat a chocolate bar, but not too hard) will bring some needed structure to the patient's life. If not, maybe getting any advice whatsoever from a serious looking psychologist would make the patient feel that they are being listened to, and that will placebo their anxiety by itself. And if not, then no harm was done and you can try something else.

But, Peterson would add, **you can't tell the patient all of that**. You won't help them by explaining leaky guts and p-values and placebo effects. They need to believe that their lives have fallen into chaos, and making breakfast is akin to slaying the dragon-goddess Tiamat and laying the foundation for stable order that creates heaven on Earth. This is *metaphorical truth*.

If you're a rationalist, you probably prefer your truths not to be so… metaphorical. But it's a silly sort of rationalist who gets sidetracked by arguments about definitions. If you don't like using the same word to mean different things *[Rule 10: be precise in your speech],* you can say "useful" or "adaptive" or "meaningful" instead of "true". It's important to use words well, but it's also important to eat a good breakfast. Probably. *[Rule 2: treat yourself like you would someone you are responsible for helping]*

One of the most underrated recent ideas in rationality is the idea of fake frameworks. I understand it thus: if you want to understand how lasers work, you should really use quantum physics as your framework. But if you want to understand how a cocktail party works, looking at quarks won't get you far. You can use the Hansonian framework of signaling, or the sociological framework of class and status, or the psychometric framework of introverts and extroverts, etc.

All of those frameworks are *fake* in the sense that *introvert* isn't a basic physical entity the same way an up quark is. Those frameworks are layers of interpretation that you impose on what you directly experience, which is human-shaped figures walking around, making noises with their mouths and sipping gin & tonics. You can't avoid imposing interpretations, so you should gather a diverse toolbox of frameworks and use them consciously even you know they're not 100% *true*.

Here's a visual example:

Q: *Which map is more true to the territory?*

A: *Neither. But if your goal is to meet Einstein on his way to work you use the one on the right, and if your goal is to count the trees on the golf course you use the one on the left.*

By the way, there's a decent chance that "fake frameworks" is what the post-rationalists have been trying to explain to me all along, except they were kind of rude about it. If it's true that they had the same message, it took Valentine to get it through my skull because he's an excellent teacher, and also someone I personally like. *Liking*shouldn't matter to rationalists, but somehow it always seems to matter to humans. *[Rule 5: do not let your children do anything that makes you dislike them]*

That's what Jordan Peterson is: a fake framework. He's a mask you can put on, a mask that changes how you see the world and how you see yourself in the mirror. Putting on the Jordan Peterson mask adds two crucial elements that rationalists often struggle with: **motivation** and **meaning**.

The Secular Solstice is a celebration designed by rationalists to sing songs together and talk about meaning. *[Rule 3: make friends with people who want the best for you]* The first time I attended, the core theme was the story of Stonehenge. Once upon a time, humans lived in terror of the shortening of the days each autumn. But we built Stonehenge to mark the winter solstice and predict when spring would come – a first step towards conquering the cold and dark.

But how did Stonehenge get built?

First, the tribe had a Scott Alexander. Neolithic Scott listened to the shamans speak of the Sun God, and demanded to see their p-values. He counted patiently the days between the solstices of each year and drew arrows pointing to the exact direction the sun rose each day.

Finally, Scott spoke up:

> Hey guys, I don't think that the sun is a god who cares about dancing and goat sacrifice. I think it just moves around in a 365-day period, and when it rises from the the direction of that tree that's when the days start getting longer again.

And the tribe told him that it's all much more than they wanted to know about the sun.

But Scott only gets us halfway to Stonehenge. The monument itself was built over several centuries, using 25-ton rocks that were brought to the site from 140 miles away. The people who hauled the first rock had to realize (unless subject to extreme planning fallacy) that not a single person they know, nor their children or grandchildren, would see the monument completed. Yet these people hauled the rocks anyway, and that required neolithic Peterson to inspire them.

Peterson is very popular with the sort of young people who have been told all their lives to be happy and proud of just who they are. But when you're 19, short on money, shorter on status, and you start to realize you won't be a billionaire rock star, you don't see a lot to be satisfied with. Lacking anything to be proud of individually, they are tempted to substitute their self for a broader group identity. What the identity groups mostly do is complain that the world is unfair to them; this keeps the movement going but doesn't do much to alleviate the frustration of its members.

And then Peterson tells them to lift the heaviest rock they can and carry it. Will it ease their suffering? No. Everyone is suffering, but at least they can carve meaning out of that. And if enough people listen to that message century after century, we get Stonehenge. *[Rule 7: pursue what is meaningful, not what is expedient]*

A new expansion just came out for the Civilization 6 video game , and instead of playing it I'm nine hours into writing this post and barely halfway done. I hope I'm not the only one getting some meaning out of this thing.

It's not easy to tell a story that inspires a whole tribe to move 25-ton rocks. Peterson noticed that the Bible is one story that has been doing that for a good while. Eliezer noticed it too, and he was not happy about it, so he wrote his own tribe-inspiring work of fiction. I've read both, cover to cover. And although I found HPMoR more fun to read, I end up quoting from the Old Testament a lot more often when I have a point to make.

"Back in the old days, saying that the local religion is a work of fiction would have gotten you burned at the stake", Eliezer replies. Well, today quoting research on psychology gets you fired from Google, and quoting research on climate change gets you fired from the EPA. *Eppur si muove*.

Jews wrote down commentaries suggesting that the story of Jonah is metaphorical a millennium before Galileo was born, and yet they considered themselves the People of the Book. The Peterson mask reminds us that people don't have to take a story *literally* to take it *seriously*.

Peterson loves to tell the story of Cain and Abel. Humans discovered sacrifice: you can give away something today to get something better tomorrow. "Tomorrow" needs a face, so we call it "God" and act out a literal sacrifice to God to hammer the point home for the kids.

But sometimes, the sacrifice doesn't work out. You give, but you don't get, and you are driven to resentment and rage against the system. That's what Cain does, and the story tells us that it's the wrong move – you should ponder instead how to make a better sacrifice next time.

When I was younger, I went to the gym twice a week for a whole year. After a year I didn't look any sexier, I didn't get much stronger, and I was sore a lot. So I said *fuck it* and stopped. Now I started going to the gym twice a week again, but I also started reading about food and exercise to finally get my sacrifice to do something. I still don't look like someone who goes to the gym twice a week, but I can bench 20 pounds more than I could last year and I rarely get sore or injured working out. *[Rule 4: compare yourself with who you were yesterday, not with who someone else is today]*

Knowing that the story of Cain and Abel is made up hasn't prevented it from inspiring me to exercise smarter.

There's a problem: many stories that sound inspirational are full of shit. After listening to a few hours of Peterson talking about archetypes and dragons and Jesus, I wasn't convinced that he's not full of it either. You should only wear a mask if it leaves you wiser when you take it off and go back to facing your mundane problems.

What convinced me about Peterson is this snippet from his conversation with James Altucher (24 minutes in):

> If you're trying to help someone who's in a rough situtation, let's say with their relationship, you ask them to start watching themselves so that you can gather some information. Let's take a look at your relationship for a week and all you have to do is figure out when it's working and when it's not working. Or, when it's working horribly and when it's working not too bad. Just keep track of that.
> *"Well, my wife ignores me at the dinner table,"* or *"My wife ignores me when I come home."* Then we start small. How would you like your wife to greet you when you come home?
> *"I'd like her to stop what she's doing and come to the door."* Well, ask her under what conditions she would be willing to do that. And let her do it badly. Do it for a week, just agree that when either of you comes home you shut off the TV and ask *"how was your day?"* and listen for 10 seconds, and see how that goes.
> Carl Jung said *"modern people can't see God because they won't look low enough"*. It

> means that people underestimate the importance of small things. They're not small. How your wife says hi to you when you come home – that's not small, because you come home all the time. You come home three times a day, so we can do the arithmetic. Let's say you spend 15 minutes a day coming home, something like that. And then it's every day, so that's 7 days a week, so that's 105 minutes. Let's call it 90 minutes a week. So that's 6 hours a month, 72 hours a year. So you basically spend two full workweeks *coming home*, that's about 3% of your life.
> You spend about 3% of your life coming home. Fix it! Then, fix 30 more things.

Aside from the Jung quote, that's the most Putanumonit piece of life advice I have ever heard on a podcast, complete with unnecessary arithmetic. If Peterson can put on a Putanumonit hat and come up with something that makes deep sense to me, perhaps I could do the same with a Peterson mask.

The rationalist project is about finding true answers to difficult questions. We have a formula that does that, and we've tracked many ways in which our minds can veer of the right path. But there are profound ways in which a person can be unready to seek the truth, ways that are hard to measure in a behavioral econ lab and assign a catchy moniker to.

I have written a lot about romance and dating in the last two years, including some mildly controversial takes. I could not have written any of them before I met my wife. Not because I didn't know the facts or the game theory, but because I wasn't emotionally ready. When I read private drafts I wrote about women from years ago, they are colored by frustration, guilt, exuberance or fear, all depending on the outcome of the last date I've been on. Those emotions aren't exactly conducive to clarity of thought.

I think this was also the reason why Scott Aaronson wrote The Comment that led to Untitled only when he was married and had a child. Then, he could weather the resulting storm without backing down from his truth. It is hard to see something true about relationships when your own aren't in order, let alone write something true. *[Rule 6: set your house in perfect order before you criticize the world]*

The flipside is: when you wear the Peterson mask, you are compelled to spread the word when you've found a path that leads somewhere true. There is no higher calling in Peterson's worldview. The Kolmogorov Option becomes the Kolmogorov Mandate (and the Scott Alexander mask mostly agrees).

Let's go back to the beginning: Peterson made noise by refusing to comply with a law that doesn't actually do what he claims. How is that contributing to the truth?

For starters, I would have bet that Peterson was going to lose his job when the letters calling for his dismissal started rolling in, letters signed by hundreds of Peterson's own colleagues at the University of Toronto. I would have bet wrong: the only thing that happened is that Peterson now makes several times his academic salary from his Patreon account (if you want me to start saying crazy things you can try clicking here).

This is critical: it created common public knowledge that if free speech is ever actually threatened by the government to the extent that Peterson claims, the support for free speech will be overwhelming even at universities. Speaking an unpopular truth is a coordination problem, you have to know that others will stand with you if you stand up first. *[Rule 1: stand up straight with your shoulders back]*

Now, more people know that there's appetite in the West for people who stand up for truth. This isn't a partisan thing, I hope that Peterson inspires people with inconvenient leftist opinions to speak up in red tribe-dominated spaces (e.g. the NFL protests).

Peterson was *technically wrong*, as he is on many things. But he sees the pursuit of truth as a heroic quest and he's willing to toss some rocks around, and I think this helps the cause of truth even if one gets some *technical* details wrong.

Being wrong about the details is not good, but I think that rationalists are pretty decent at getting *technicalities* right. By using the Peterson Mask judiciously, we can achieve even more than that.

[Rule 12: pet a cat when you encounter one on the street], but don't touch the hedgehog, they don't like it.

[Preprint for commenting] Digital Immortality: Theory and Protocol for Indirect Mind Uploading

I would like to get useful input for the following text :

"Digital Immortality: Theory and Protocol for Indirect Mind Uploading"

Abstract. Future superintelligent AI will be able to reconstruct a model of the personality of a person who lived in the past based on informational traces. This could be regarded as some form of immortality if this AI also solves the problem of personal identity in a copy-friendly way. A person who is currently alive could invest now in passive self-recording and active self-description to facilitate such reconstruction. In this article, we analyze informational-theoretical relationships between the human mind, its traces, and its future model; based on this analysis, we suggest the instruments to most cost-effectively collect quality data about a person for future resurrection. These guidelines form a “digital immortality protocol”. Digital immortality is plan C for achieving immortality, after plan A, life extension, and plan B, cryonics.

Keywords: Digital immortality – superintelligence – effective altruism – life extension – mind uploading.

Highlights:

· Future superintelligent AI will be able to simulate past people.

· To help AI improve its simulations, we can collect data about a living person now.

· Passive data collection is constant recording.

· Active data collection is running tests and recording self-description.

· The best way to collect information is to create art, as it is unique, valuable, and predictive.

Full text is open for commenting: https://goo.gl/QkDfdU

'Trivial Inconvenience Day' Retrospective

Introduction

Shortly after reading Scott Alexander's LessWrong Crypto Autopsy I found myself agreeing with the point *so strongly* I was brainstorming ways that its dismal outcome could have been prevented. Peoples personal accounts of why they didn't buy bitcoin seemed to converge on a central theme: Buying Bitcoin was a trivial inconvenience. Pondering what might be done in light of this, I was reminded of Boston Rat's Bureaucracy Day. The Bureaucracy Day is essentially a designated day for people to beat the ugh field effect by getting together and going through the whole mess of annoying tasks, paperwork, and other trivially inconvenient things people have been putting off. Having been impressed by the concept the first time I read about it, Scott's dire analysis convinced me to try something like it in the hope that it would be a useful tool against this sort of thing happening again.

Design

Preliminary Steps & Research

The actual idea came to me when a friend linked Scott's post. From the first paragraphs I was taken by the premise, and lamented how I had the opportunity to buy Bitcoin when it was pennies but didn't. For me the sting is even greater because I'd been tempted to buy an entire thousand dollars worth, but eventually decided against it. (As Scott points out in his post, I was especially being an idiot since just buying 10 bucks worth would have made me a much wealthier man.) Part of why I didn't was I was under the impression my only option for buying was essentially meeting someone in person that mined it. I'm unsure exactly when this was, but it's quite possible I was thinking about this pre-exchange and that was the reality.

As I was sitting there thinking about this, I remembered the Boston Rat post and felt it was a good starting template for solving this issue. I said to my friend:

> Okay, here's my idea. Are you aware of Boston rat's "Bureaucracy day?" We do "Trivial Inconvenience Day". Where we get people into a giant IRC room. Or Discord or whatever. (IRC is easier to use a web client for without sign up.) And have people make a list of s*** they know in the abstract they should do but are putting off. And then we all do it together and talk about having done it. Sort of like a baptism, but for buying cryptocurrency or fixing my website. >_>

The friend seemed enthusiastic about the idea, noting that the Event Horizon group house had done a similar thing under the label 'Agency Hour', but they'd wished it were longer.

I went ahead and pasted that exact quote from the conversation into my private discord server, and pinged everyone to see if they'd be interested in making it happen. The responses were telling enough on their own:

> I'm down
>
> yes
>
> hell yes
>
> sure
>
> Thirded
>
> I'm in with high probability, depending on when it is.

Eight people was enough for me. With the enthusiastic responses in mind I got to making it happen.

I made a list of things I was putting off. (These lists are kind of intrinsically embarrassing, so you will hopefully forgive me for not providing examples.) Then I solicited lists from other people, and fairly soon several of us had provided a decent list of things we could do for Trivial Inconvenience Day. Next was the real challenge: Scheduling. I asked everyone in the small group about their availability, and eventually decided to make the day:

> @Trivial Inconvenience Day Friday at 14:00 PST?

This precluded one participant from coming, but otherwise fit most peoples schedules. I figured bringing in others later that it was up to chance whether this date was acceptable to them, and I was fine with that.

Attempted Atmosphere

From the outset I was aiming for the event to have a particular sort of feel. I wanted it to be serious, but also sort of cheesy. In real life we're used to doing slog-y, tedious work and having nothing to show for it afterwards besides hours passed on the clock. It's not a very rewarding experience. Therefore to help counteract this I wanted the experience to be chock full of artificial rewards[0]. The fact of the matter is that these mundane necessities of life give us nowhere near the level of reward we feel we deserve for the effort. That is after all why they're undone in the first place. Keeping this in mind I wanted the atmosphere to be high energy, exuberantly enthusiastic. It's probably best summarized by the short description of the event I gave to someone that was confused about its logistics:

> We all make our lists, then publicly commit to doing things, then do them, say we did them, and get praise and cheers for it.

Organizing & Advertising

Next up was to try and expand the number of participants. I assumed that with 7 people we would see significant shrinkage and no-shows. That meant for the thing to really work I'd have to boost it to a healthier, larger number of people. I briefly considered posting a public advertisement on LW 2, but the meetups feature didn't exist yet. When I asked staff for guidance on this issue they told me to post it on my personal blog. It was at about this point I realized that if this first event had major hiccups (as it was virtually guaranteed to) then I'd be spoiling the concept for a long

time. The other problem was that by the time I'd finally found the time to start advertising, it was Thursday and the event was scheduled the next day. Advertising to a large group on such short notice felt rude, so I opted against it. Instead I decided to advertise directly to people on my Discord friends list. The exact message I sent them was:

> Hi. I'm organizing a "Trivial Inconvenience Day" tomorrow at 14:00.
>
> Basically it's a lot like Boston Rat's "Beauracracy Day" where people show up and get paperwork/etc they've been putting off done, except here the focus is on things which are trivially inconvenient that you want to do but haven't gotten around to. Would you be interested? It's online of course. It'll be hosted on Discord and I'll invite you at the time if you want to do it.

I created a Discord server for the event and invited people early so that I could ping them all when the time came to get on it.

The Event

Pre-Event Instructions

These are the instructions I gave to participants.

Here's basically how this is going to work:

1. You ideally already have a list of things which will take under an hour to complete that you've been putting off. If you don't, make that list now.
2. You're going to announce your list to us, pick a reasonable number of things to commit to completing.
3. Do the things you said you would.
4. We all act extremely enthusiastic about you doing this, and I give you a role called 'Doer' that puts you above the normal channel listing.
5. Consequently, when other people finish doing things shower them with praise.
6. Ideally most of us end up doing the things, and feel good about it.
7. Conquer your fear of seeming cheesy or silly, that's part of the fun. :3
8. The voice channel is for people who would like to participate in actual vocal cheering and encouragement. It's not mandatory though, text is fine.

/endQuote

I should note that 'under an hour' in point one turned out to be woefully inadequate, and I would strike it from any future versions.

How It Went

The day of the event itself made me a little nervous. *Were people really going to show up*? I decided in advance that if nobody did I would just do my list anyway and then try again later. Thankfully, people did show up. All told 16 people joined the server, including myself. Not all of them participated however.

We laid out lists of what we wanted to do, how many things we needed to do to consider ourselves to have succeeded, and then we got to work. One fundamental

conflict I hadn't anticipated was that working and cheering are kind of at odds with each other. People don't all complete things at the same time, so when someone succeeds the others aren't standing by to give them boisterous cheering. It's more like a silent chorus of clap emoji that slowly filter in. This is however better than the usual silence.

Highlights

- One participant signed up for a Vanguard Index Fund account.
- Another participant began the process for getting their statue appraised.
- Someone found a couple hundred dollars in assorted cash going through their old documents.
- Somebody who had been buying 10 dollars of bitcoin a week raised their weekly buy amount.

Conclusion & Results

Overall, people seemed to be satisfied with the concept and execution. While a lot of *genuinely trivial* things got done, the mere fact that I got someone to *stop procrastinating on signing up for Vanguard* is more than enough for me to consider it a success. However I also gave participants an exit survey to see how well the event did. The results of this survey were encouraging. I can confidently recommend rats who feel like attempting a version of this in their own spaces do so.

Exit Survey Results

A copy of the survey can be found here.

Lessons Learned & Changes For Next Time

- Keep track of all invites sent, and the responses given. This was important data that at the time I was sort of just trying to commit to memory. In retrospect it would have been very useful to be able to measure the performance of invitations, A/B test variations, etc.
- Assuming I did that, explicitly calculate shrinkage and collect RSVP's so I can have a better estimation of how much effort needs to go into advertising to get X number of people.
- Either find a better way to define a cutoff people for when someone has 'completed' the Trivial Inconvenience Day challenge, or fundamentally restructure the event to avoid this issue. Under an hour is not sufficient, all day is obviously fatiguing.
- Advertise the next Trivial Inconvenience Day on the LW2 events & meetups page, now that it exists.

Participant Feedback

> It got kind of hard to tell who was and wasn't on board with things, I think making that more legible would increase the sense of cohesion. Otherwise things went fairly well.

idk; i got distracted and ended up failing to participate

Possibly some kind of official end point instead of just petering out as people run out of energy

Criticise people's goals a little (I know you can't do it for manpower reasons)

References

[0]: Yegge, Steve. (2012, March 12). *The borderlands gun collector's club.* Retrieved from https://steve-yegge.blogspot.com/2012/03/borderlands-gun-collectors-club.html

Quick Nate/Eliezer comments on discontinuity

This isn't a proper response to Paul Christiano or Katja Grace's recent writings about takeoff speed, but I wanted to cross-post Eliezer's first quick comments on Katja's piece someplace more linkable than Twitter:

> There's a lot of steps in this argument that need to be spelled out in more detail. Hopefully I get a chance to write that up soon. But it already raises the level of debate by a lot, for which I am grateful.
>
> E.g. it is not intuitive to me that "But evolution wasn't trying to optimize for STEM ability" is a rejoinder to "Gosh hominids sure got better at that quickly." I can imagine one detailed argument that this might be trying to gesture at, but I don't know if I'm imagining right.
>
> Similarly it's hard to pin down which arguments say '"Average tech progress rates tell us something about an underlying step of inputs and returns with this type signature" and which say "I want to put the larger process in this reference class and demand big proof burdens."

I also wanted to caveat: Nate's experience is that the label "discontinuity" is usually assigned to misinterpretations of his position on AGI, so I don't want to endorse this particular framing of what the key question is. Quoting Nate from a conversation I recently had with him (not responding to these particular posts):

> On my model, the key point is not "some AI systems will undergo discontinuous leaps in their intelligence as they learn," but rather, "different people will try to build AI systems in different ways, and each will have some path of construction and some path of learning that can be modeled relatively well by some curve, and some of those curves will be very, very steep early on (e.g., when the system is first coming online, in the same way that the curve 'how good is Google's search engine' was super steep in the region between 'it doesn't work' and 'it works at least a little'), and sometimes a new system will blow past the entire edifice of human knowledge in an afternoon shortly after it finishes coming online." Like, no one is saying that Alpha Zero had massive discontinuities in its learning curve, but it also wasn't just AlphaGo Lee Sedol but with marginally more training: the architecture was pulled apart, restructured, and put back together, and the reassembled system was on a qualitatively steeper learning curve.
>
> My point here isn't to throw "AGI will undergo discontinuous leaps as they learn" under the bus. Self-rewriting systems likely will (on my models) gain intelligence in leaps and bounds. What I'm trying to say is that I don't think this disagreement is the central disagreement. I think the key disagreement is instead about where the main force of improvement in early human-designed AGI systems comes from — is it from existing systems progressing up their improvement curves, or from new systems coming online on qualitatively steeper improvement curves?

Katja replied on Facebook: "FWIW, whenever I am talking about discontinuities, I am usually thinking of e.g. one system doing much better than a previous system, not discontinuities in the training of one particular system—if a discontinuity in training one system does not make the new system discontinuously better than the previous

system, then I don't see why it would be important, and if it does, then it seems more relevant to talk about that."

The Math Learning Experiment

Last Sunday I and a small group of volunteers ran an experiment. We got a bunch of CFAR alumni together and asked half of them to tutor the other half in math (in a broad sense, so including computer science) while paying close attention to what was happening in the minds of the tutees, focusing on issues like where the tutees were confused, where they were frustrated, where they were anxious or afraid, etc.

I had a few motivations in trying this. One was a sense that most rationalists could benefit from learning more math on the margin. Another was a sense of there being various obstacles standing in the way of that, for example an identity of "not being a math person" or various flavors of math anxiety (in addition to the fact that it's a hassle / expensive to find a math tutor, and that people have more urgent things to do), and wanting to see the extent to which those obstacles could be debugged at scale. A third was to explore running events at all.

The fourth and most important motivation for me came out of thoughts I've been having since reading *Inadequate Equilibria*. There's a cluster of people, including but not limited to Eliezer, Critch, and Nate, who (according to me) have what I internally call "trustworthy inside views," another name for which might be the ability to reliably generate useful gears models, and act based on them. This is the thing they do instead of using modest epistemology; it's the thing that allowed Eliezer to write HPMoR, among many other things. And what all of the people who seem to me to have this ability have in common is that they all have strong backgrounds in a technical subject like math, physics, or computer science (in addition to something else, this isn't sufficient). Probably there are other ways to acquire this ability, but acquiring it through learning technical subjects in a particular way seemed promising, and I wanted to see if I could help others make any progress in that direction at all, or at least understand the obstacles in the way of that.

I learned less about inside view than I was hoping. My plan was to run around observing all of the tutor-tutee pairs, but it was harder to do this productively than I expected: I kept feeling like I was missing important conversational context and not knowing how to model the tutee because of that. We tutored for 3 rounds of 30-40 minutes each, and that wasn't long enough; I think to really start approaching the inside view skill would require longer rounds and also particular skills on the part of the tutor that I didn't have the time to attempt to transfer to all of the tutors. I think if I want to learn more about the inside view skill in the future I should try to tutor someone myself, and for longer periods of time; I'll probably try this next.

An interesting point that came up after the 1st round is the unusual role that definitions play in mathematics, as well as the strange way mathematicians typically treat them in textbooks or similar. It's easy to have the experience of reading a new chapter of a textbook, absorbing all the definitions, understanding all the proofs, but still feeling like the wool was pulled over your eyes in some sense. (General topology can really strongly trigger this feeling.) Textbooks generally write down definitions as if there is nothing or very little that needs to be explained in doing so. In fact mathematicians have put a lot of work into writing down the "right" definitions (which is something like writing down the "right" ontology), but this work is basically invisible - I have never seen any textbook seriously discuss it, and it only comes up briefly in discussions of history of mathematics. People don't even talk about it in graduate

school, despite the fact that graduate students are expected to be able to generate new mathematics as their job. At no point in a standard math education are students explicitly given the opportunity to practice approaching new mathematical territory with no definitions to help them orient towards it, and coming up with useful definitions on their own.

I consider it an important feature of my approach to mathematics, which feels related to the inside view skill, that I consistently get frustrated at definitions that I don't understand how to reinvent instead of taking them as given. A large part of my math blogging is about motivating definitions. Sometimes it would take me years between my first exposure to and frustration at a definition and the time that I finally had a satisfying motivation for it; for chain complexes it took something like 4 years, and the satisfying motivation is much more complicated to explain than the definition. (For an example that hasn't finished yet, I am still frustrated about entropy, even after writing this post, which clarified a lot of things for me.)

As far as the logistics of tutoring went, the tutor-tutee numbers worked out even on the first 2 rounds, which was remarkable, and on the 3rd round we had some extra tutees. I asked them to be "meta" for a tutor-tutee pair - keeping track of where they were in the discussion, slowing down the tutor if the tutee looks confused, etc. - and people reported that this was extremely helpful. This jives with activities we've been running at CFAR workshops around people talking in groups of 3 in this way (2 people having some sort of conversation and the 3rd person doing various kinds of meta for them) and I'd like to try doing things this way by default if we run something like this again (the next thing I might try is zeroing in on teaching people how to prove things).

I also had forgotten to plan for some things logistically (making sure I had enough hands to set up the space we were using easily, deciding in advance whether to order in lunch or ask people to go out), but they were magically taken care of by my volunteers, for whom I'm very grateful. In the future I'll try to set more time aside earlier in planning for Murphyjitsu.

Overall we at least ended up having a reasonably fun social time, and some people picked up some math in a way that was probably good. I expect the primary impact of this event on the tutees will be to help them feel more like learning math is a thing they can actually do, which is great. The tutors may have learned something about tutoring, which I'm also happy with.

Brains and backprop: a key timeline crux

[Crossposted from my blog]

The Secret Sauce Question

Human brains still outperform deep learning algorithms in a wide variety of tasks, such as playing soccer or knowing that it's a bad idea to drive off a cliff without having to try first (for more formal examples, see Lake et al., 2017; Hinton, 2017; LeCun, 2018; Irpan, 2018). This fact can be taken as evidence for two different hypotheses:

1. In order to develop human-level AI, we have to develop entirely new learning algorithms. At the moment, AI is a deep conceptual problem.
2. In order to develop human-level AI, we basically just have to improve current deep learning algorithms (and their hardware) a lot. At the moment, AI is an engineering problem.

The question of which of these views is right I call "the secret sauce question".

The secret sauce question seems like one of the most important considerations in estimating how long there is left until the development of human-level artificial intelligence ("timelines"). If something like 2) is true, timelines are arguably substantially shorter than if something like 1) is true [1].

However, it seems initially difficult to arbitrate these two vague, high-level views. It appears as if though an answer requires complicated inside views stemming from deep and wide knowledge of current technical AI research. This is partly true. Yet this post proposes that there might also be single, concrete discovery capable of settling the secret sauce question: does the human brain learn using gradient descent, by implementing backpropagation?

The importance of backpropagation

Underlying the success of modern deep learning is a single algorithm: gradient descent with backpropagation of error (LeCun et al., 2015). In fact, the majority of research is not focused on finding better algorithms, but rather on finding better cost functions to descend using this algorithm (Marblestone et al., 2016). Yet, in stark contrast to this success, since the 1980's the key objection of neuroscientists to deep learning has been that backpropagation is not biologically plausible (Crick, 1989; Stork, 1989).

As a result, the question of whether the brain implements backpropagation provides critical evidence on the secret sauce problem. If the brain does *not* use it, and *still* outperforms deep learning while running on the energy of a laptop and training on several orders of magnitude fewer training examples than parameters, this suggests that a deep conceptual advance is necessary to build human-level artificial intelligence. There's some other remarkable algorithm out there, and evolution found

it. But if the brain *does* use backprop, then the reason deep learning works so well is because it's somehow *on the right track*. Human researchers and evolution converged on a common solution to the problem of optimising large networks of neuron-like units. (These arguments assume that *if* a solution is biologically plausible and the best solution available, then it would have evolved).

Actually, the situation is a bit more nuanced than this, and I think it can be clarified by distinguishing between algorithms that are:

Biologically actual: What the brain *actually* does.

Biologically plausible: What the brain *might* have done, while still being restricted by evolutionary selection pressure towards energy efficiency etc.

For example, humans walk with legs, but it seems possible that evolution might have given us wings or fins instead, as those solutions work for other animals. However, evolution could not have given us wheels, as that requires a separable axle and wheel, and it's unclear what an evolutionary path to an organism with two separable parts looks like (excluding symbiotic relationships).

Biologically possible: What is technically possible to do with collections of cells, regardless of its relative evolutionary advantage.

For example, even though evolving wheels is implausible, there might be no inherent problem with an organism having wheels (created by "God", say), in the way in which there's an inherent problem with an organism's axons sending action potentials faster than the speed of light.

I think this leads to the following conclusions:

Nature of backprop: Implication for timelines

Biologically impossible: Unclear, there might be multiple "secret sauces"

Biologically possible, but not plausible: Same as above

Biologically plausible, but not actual: Timelines are long, there's likely a "secret sauce"

Biologically actual: Timelines are short, there's likely no "secret sauce"

In cases where evolution could not invent backprop anyway, it's hard to compare things. That is consistent both with backprop not being the right way to go and with it being better than whatever evolution did.

It might be objected that this question doesn't really matter, since *if* neuroscientists found out that the brain does backprop, they have not thereby created any new algorithm -- but merely given stronger evidence for the workability of previous algorithms. Deep learning researchers wouldn't find this any more useful than Usain Bolt would find it useful to know that his starting stance during the sprint countdown is optimal: he's been using it for years anyway, and is mostly just eager to go back to the gym.

However, this argument seems mistaken.

On the one hand, just because it's not useful to deep learning practitioners does not mean it's not useful others trying to estimated the timelines of technological development (such as policy-makers or charitable foundations).

On the other hand, I think this knowledge *is* very practically useful for deep learning practitioners. According to my current models, the field seems unique in combining the following features:

- Long iteration loops (on the order of GPU-weeks to GPU-years) for testing new ideas.
- High dependence of performance on hyperparameters, such that the right algorithm with slightly off hyperparameters will not work *at all.*
- High dependence of performance on the amount of compute accessible, such that the differences between enough and almost enough are step-like, or qualitative rather than quantitative. Too little compute and the algorithm just doesn't work *at all.*
- Lack of a unified set of first principles for understanding the problems, and instead a collection of effective heuristics

This is an environment where it is critically important to develop strong priors on what *should* work, and to stick with those in face countless fruitless tests. Indeed, LeCun, Hinton and Bengio seem to have persevered for decades before the AI community stopped thinking they were crazy. (This is similar in some interesting ways to the state of astronomy and physics before Newton. I've blogged about this before here.) There's an asymmetry such that even though training a very powerful architecture can be quick (on the order of a GPU-day), iterating over architectures to figure out which ones to train fully in the first place can be incredibly costly. As such, knowing whether gradient descent with backprop is or is not the way to go would lead enable more efficient allocation of research time (though mostly so in case backprop is *not* the way to go, as the majority of current researchers assume it anyway).

Appendix: Brief theoretical background

This section describes what backpropagation is, why neuroscientists have claimed it is implausible, and why some deep learning researchers think those neuroscientists are wrong. The latter arguments are basically summarised from this talk by Hinton.

Multi-layer networks with access to an error signal face the so-called "credit assignment problem". The error of the computation will only be available at the output: a child pronouncing a word erroneously, a rodent tasting an unexpectedly nauseating liquid, a monkey mistaking a stick for a snake. However, in order for the network to improve its representations and avoid making the same mistake in the future, it has to know which representations to "blame" for the mistake. Is the monkey too prone to think long things are snakes? Or is it bad at discriminating the textures of wood and skin? Or is it bad at telling eyes from eye-sized bumps? And so forth. This problem is exacerbated by the fact that neural network models often have tens or hundreds of thousands of parameters, not to mention the human brain, which is estimated to have on the order of 1014 synapses. Backpropagation proposes to solve this problem by observing that the maths of gradient descent work out such that one

can essentially send the error signal from the output, back through the network towards the input, modulating it by the strength of the connections along the way. (A complementary perspective on backprop is that it is just an efficient way of computing derivatives in large computational graphs, see e.g. Olah, 2015).

Now why do some neuroscientists have a problem with this?

Objection 1:

Most learning in the brain is unsupervised, without any error signal similar to those used in supervised learning.

Hinton's reply:

There are at least three ways of doing backpropagation without an external supervision signal:

1. **Try to reconstruct the original input (using e.g. auto-encoders), and thereby develop representations sensitive to the statistics of the input domain**

2. **Use the broader context of the input to train local features**

For example, in the sentence "She scromed him with the frying pan", we can infer that the sentence as a whole doesn't sound very pleasant, and use that to update our representation of the novel word "scrom"

3. **Learn a generative model that assigns high probability to the input (e.g. using variational auto-encoders or the wake-sleep algorithm from the 1990's)**

Bengio and colleagues (2017) have also done interesting work on this, partly reviving energy-minimising Hopfield networks from the 1980's

Objection 2:

Objection 2. Neurons communicate using binary spikes, rather than real values (this was among the earliest objections to backprop).

Hinton's reply:

First, one can just send spikes stochastically and use the expected spike rate (e.g. with a poisson rate, which is somewhat close to what real neurons do, although there are important differences see e.g., Ma et al., 2006; Pouget et al. 2003).

Second, this might make evolutionary sense, as the stochasticity acts as a regularising mechanism making the network more robust to overfitting. This behaviour is in fact where Hinton got the idea for the drop-out algorithm (which has been very popular, though it recently seems to have been largely replaced by batch normalisation).

Objection 3:

Single neurons cannot represent two distinct kind of quantities, as would be required to do backprop (the presence of features and gradients for training).

Hinton's reply:

This is in fact possible. One can use the temporal derivative of the neuronal activity to represent gradients.

(There is interesting neuropsychological evidence supporting the idea that the temporal derivative of a neuron can *not* be used to represent changes in that feature, and that different populations of neurons are required to represent the presence and the change of a feature. Patients with certain brain damage seem able to recognise that a moving car occupies different locations at two points in time, without being able to ever detect the car changing position.)

Objection 4:

Cortical connections only transmit information in one direction (from soma to synapse), and the kinds of backprojections that exist are far from the perfectly symmetric ones used for backprop.

Hinton's reply:

This led him to abandon the idea that the brain could do backpropagation for a decade, until “a miracle appeared”. Lillicrap and colleagues at DeepMind (2016) found that a network propagating gradients back through *random* and *fixed* feedback weights in the hidden layer can match the performance of one using ordinary backprop, given a mechanism for normalization and under the assumption that the weights preserve the sign of the gradients. This is a remarkable and surprising result, and indicates that backprop is still poorly understood. (See also follow-up work by Liao et al., 2016).

---

[1] One possible argument for this is that in a larger number of plausible worlds, if 2) is true and conceptual advances are necessary, then building superintelligence will turn into an engineering problem once those advances have been made. Hence 2) requires strictly more resources than 1).

Discussion questions

I'd encourage discussion on:

Whether the brain does backprop (object-level discussion on the work of Lillicrap, Hinton, Bengio, Liao and others)?

Whether it's actually important for the secret sauce question to know whether the brain does backprop?

To keep things focused and manageable, it seems reasonable to disencourage discussion of what *other* secret sauces there might be.

Request for "Tests" for the MIRI Research Guide

Lately I've been looking into learning the material in the MIRI research guide. After some time, I noticed that my biggest trepidation about diving in was the nagging question of, "But how will I know if I actually learned the stuff?"

Once I realized that was the thing holding me back, it became easier to think about solutions. Some of the topics correspond with courses that are common in universities, so I can pilfer final exams from their sites (woot to MIT open courseware). But for other topics I wanted to see what people here thought.

In whatever domain you specialize in, what are some examples of problems or questions that one can only answer by having a solid understanding of a huge swath of said domain? Below I've listed everything from the MIRI research guide that I could perceive as a distinct category.

(ex. If you were learning about Digital Systems and Computer Architecture, my test would be "In systemVerilog, simulate a basic 16-bit processor that can be programmed using a RISC assembly language of your design.")

(edit: I know that "huge swaths" is pretty vague, and suggestions don't have to be things that you think certify/prove that you get a topic. While a "comprehensive test" would be nice, problems/prompts like what Qiaochu commented are exactly what I'm looking for)

1. Set theory
2. Probability
3. Probabilistic Inference
4. Statistics
5. Machine Learning
6. Solomonoff Induction
7. Naturalized Induction
8. VNM Decision Theory
9. Functional Decision Theory
10. Logical Uncertainty
11. First Order Logic
12. Vingean Reflection
13. Corrigibility
14. Linear Algebra
15. Topology
16. Category Theory
17. Type Theory

AI Alignment Prize: Round 2 due March 31, 2018

There are still several weeks left to enter the second round of the AI Alignment Prize. The official announcement of the second round, along with the first round's winners, is here. We'll be awarding a *minimum* of $10,000 for advances in technical, philosophical and/or strategic approaches to AI alignment: Ensuring that future smarter than human intelligence will have goals aligned with the goals of humanity. Last time we got such good entries we tripled our announced prize pool, and we'd love to justify doing that again. You can enter by replying to this post, to the original announcement of the second prize, or by emailing apply@ai-alignment.com.

AI alignment is wide open for both amateur and professional ideas. It's one of the most important things humanity needs to work on, yet it receives almost no serious attention. Much of the low hanging fruit remains unpicked. We hope to help change that. Even for those who don't produce new advances or win prizes, more time spent on this problem is time well spent.

Whether or not you're considering entering, please help get the word out and remind people of the upcoming deadline. Please do repost this notice with a link back to the original (since people won't be able to enter by replying to the copy). Entries for the second round are running behind the pace from the first round, and second times around risk being less sexy and exciting than the first, despite the increased prize pool. We also didn't do enough to make it clear the contest was continuing with a larger prize pool.

Paul Christiano is also offering an additional $10,000 prize for the best evidence that his preferred approach to AI alignment, which is reflected in posts at ai-alignment.com and centered on iterated amplification, is doomed. That one is due on April 20.

Last November, Paul Christiano, Vladimir Slepnev and myself announced the AI Alignment Prize for publicly posted work advancing our understanding of AI alignment. We were hopeful that a prize could be effective at motivating good work and the sharing of good work, but worried we wouldn't get quality entries.

We need not have worried. On all fronts, it was a smashing success, exceeding all of our expectations for both quantity and quality. We got winning entries both from employees of organizations like FRI and MIRI, and from independent bloggers. Our original prize pool was $5,000 total, but we were so happy that we decided to award $15,000 total, including $5,000 to the first prize winner, Scott Garrabant. Scott has gone on to write many more excellent posts, and has himself observed that were it not for the prize, his piece on Goodheart Taxonomy and many other pieces would likely have remained in his draft folder.

The other winners were Tobias Baumann for Using Surrogate Goals to Deflect Threats, Vadim Kosoy for Delegative Reinforcement Learning (1, 2, 3), John Maxwell for Friendly AI through Ontology Autogeneration, Alex Mennen for his posts on legibility to other agents and learning goals of simple agents, and Caspar Oesterheld for his post and paper studying which decision theories would arise from environments like reinforcement learning or futarchy.

It is easy to default to thinking that scientific progress comes from Official Scientific Papers published in Official Journals, whereas blog posts are people having fun or hashing out ideas. Not so! Many of the best ideas come from people writing on their own, on their own platforms. Encouraging them to take that seriously, and others to take them seriously in turn when they deserve it, and raising the prestige and status of such efforts, is one of our key goals. The winning entry from the first contest was a blog post breaking down a well-known concept, Goodhart's Law, and giving us a better way to talk about it. Often the most important advances are simple things, done well.

Here are the rules of the contest, which are unchanged except for which posts to reply to, larger numbers and new dates:

Rules

Your entry must be published online for the first time between January 15 and March 31, 2018, and contain novel ideas about AI alignment. Entries have no minimum or maximum size. Important ideas can be short!

Your entry must be written by you, and submitted before 9pm Pacific Time on March 31, 2018. Submit your entries either as links in the comments to the original announcement, this post, or by email to apply@ai-alignment.com. We may provide feedback on early entries to allow improvement.

We will award *a minimum of* $10,000 to the winners. The first place winner will get at least $5000. The second place winner will get at least $2000. Other winners will get at least $1000.

Entries will be judged subjectively. Final judgment will be by Paul Christiano. Prizes will be awarded on or before April 15, 2018.

What kind of work are we looking for?

AI Alignment focuses on ways to ensure that future smarter than human intelligence will have goals aligned with the goals of humanity. Many approaches to AI Alignment deserve attention. This includes technical and philosophical topics, as well as strategic research about related social, economic or political issues. A non-exhaustive list of technical and other topics can be found here.

We are **not** interested in research dealing with the dangers of existing machine learning systems commonly called AI that do not have smarter than human intelligence. These concerns are also understudied, but are not the subject of this prize except in the context of future smarter than human intelligence. We are also **not** interested in general AI research. We care about AI alignment, which may or may not also advance the cause of general AI research.

For further guidance on what we're looking for, and because they're worth reading and thinking about, check out the winners of the first round, which we linked to above.

Computational Complexity of P-Zombies

This is a linkpost for https://mapandterritory.org/computational-complexity-of-p-zombies-fc56909af96f

I've claimed that p-zombies require exponentially more resources than phenomenally conscious agents and used this to justify that AGI will necessarily be phenomenally conscious. I've previously supported this claim by pointing to a related result in integrated information theory showing that p-zombies constructed as feed-forward networks are exponentially larger than behaviorally equivalent conscious networks, but now I'd like to see if I can get a similar result in noematology.

NB: This account is provisional and not sufficiently well-founded to constitute a proof given my assumptions. That said, it should be precise enough to be falsifiable if wrong.

First, some notation. Given an agent A, let a **behavior** be a pair {x, A(x)} where x is the world state before the agent acts and A(x) is the world state after. For example, if when holding a teacup A takes a sip of tea, then x is "holding a cup of tea" and A(x) is "taking a sip of tea". If we have a second agent, Z, we may wish to ask if, given the same initial world state x, Z(x)=A(x), viz. Z and A exhibit the same behavior when presented with x. In an obvious sense this question is meaningless because A and Z cannot be said to observe the exact same reality nor is there to any epistemically valid method within phenomenalism by which we may directly compare two world states, but given an observer agent O, we can consider if A(x)=Z(x) with respect to O's experience of A(x) and Z(x) and this is what we will mean when we ask if A exhibits the same behavior as Z.

Now let A be a phenomenally conscious agent and Z a p-zombie of A, meaning that Z and A cannot be distinguished based on observed behavior and Z is not phenomenally conscious (although it is, of course, cybernetic). More formally, we are supposing that A and Z agree on the set of behaviors X such that for all {x, y} in X, y=A(x)=Z(x). I then claim that Z is exponentially larger than A with the cardinality of X, where by "larger" I mean physically made of more stuff.

The short explanation of my reasoning is that Z must do with only the ontic what A does with ontology. Unfortunately this hides a lot of my understanding about what phenomenal consciousness does and how it works, so instead I'll explain by relating A and Z to computational models that can describe them and use those to show that Z must be exponentially larger than A.

Since Z is not phenomenally conscious it cannot create information within itself because it has no feedback process with which information could be created (if it did it would then be phenomenally conscious and no longer a p-zombie). Computationally this describes the class of computers we call deterministic finite automata or DFAs. A, on the other hand, can create information via feedback; this gives it "memory" which we can model as a Turing machine with a finite tape (the tape must be finite since it is embodied in the real world rather than a theoretical construct). Other models are possible, but these should suffice for our purposes.

Since Z is a DFA, it must have at least one state for each world state x it encounters to ensure the correct mapping from x to y for each {x, y} in X, thus Z has at least O(|X|) states. This puts a lower bound on the size of Z, so to show that Z is exponentially larger than A in terms of |X| we need to show that A has an upper bound on its size of O(lg|X|) states and tape length. We will do this by constructing a phenomenally conscious version of Z by converting it from a DFA to a Turing machine. It's easier to understand how to convert a finite Turing machine into a DFA, though, so we'll do that first and then reverse the process to show that A is no larger than O(lg|X|).

Since A is a finite Turing machine it can be converted into a DFA by an algorithm analogous to loop unrolling that requires creating at least one state in the DFA for each possible configuration of the tape. Let us denote this DFA-ized version of A as A'. This means that, given that A has a tape of length n, we need A' to contain at least 2^n states. Supposing A consists of a state machine of fixed sized independent of |X| and always encodes information about each world state x on the tape, then A will need a tape of at least length lg|X| to encode every x, and A' will need to contain at least 2^(lg|X|)=|X| additional states to replace the tape. This gives us a lower bound on the size of Z since A' has O(|X|) states, so Z has O(|X|) states.

Reversing this algorithm we can construct a Turing machine Z' from Z by using the tape to encode the O(|X|) states of Z on at least O(lg|X|) of tape where Z' otherwise consists of a constant-sized finite state machine that emulates Z. Since Z' is a Turing machine it meets our requirements for phenomenal consciousness and so gives a lower bound on the size of A. A has a natural upper bound of O(|X|) since A can be simply constructed as Z with a tape, so A has between O(lg|X|) and O(|X|) states and tape length.

This unfortunately does not give us the desired result that A has an upper size bound of O(lg|X|) and only says that A is no larger than Z and Z is at most exponentially larger than A in terms of |X|. Nonetheless I believe it to be true that A also has at most O(lg|X|) states and tape length because this is what we empirically see when constructing, for example, less phenomenally conscious state machine solutions to programming problems rather than more phenomenally conscious versions using nested loops or recursion. I also suspect it's possible to prove an upper bound lower than O(|X|) closer to O(lg|X|) for A but have not found such a proof in the literature nor proved this myself. There may alternatively be a weaker relationship between the sizes of A and Z that has the same consequence that Z must be much larger than A but is quantitatively less than exponential.

(If you're interested in collaborating with me on this please reach out since this would be a nice result to have locked down for when I publish my work on formally stating the AI alignment problem! I believe it's possible, but I'm rusty enough on automata theory that the effort I would have to put in the prove this is high. If you already have automata theory fresh in your mind then this is likely more straightforward.)

Finally we need to connect this notion of number of states and tape length to physical size. This is straight forward, but to be explicit about it each state or tape position needs to be embodied in physical stuff. Even assuming a single transistor or a single bit of storage is all that is needed for each state or tape position, this means we can use our size calculations to estimate and compare the relative physical sizes of A and Z by setting up a correspondence between theoretical state and tape constructs and physical computer implementations. This let's us conclude at least a weak version of what we set out to prove: that a p-zombie Z of a phenomenally conscious agent A is no smaller than A in terms of |X| as measured in physical stuff, is probably much

larger than A, and is exponentially larger than A if we can prove that A has at most O(lg|X|) states and tape length.

Deleted

Deleted

Resolving human values, completely and adequately

In previous posts, I've been assuming that human values are complete and consistent, but, finally, we are ready to deal with actual human values/preferences/rewards - the whole contradictory, messy, incoherent mess of them.

Define "completely resolving human values", as an AI extracting a consistent human reward function from the inconsistent data on the preference and values of a single human (leaving aside the easier case of resolving conflicts between different humans). This post will look at how such resolutions could be done - or at least propose an initial attempt, to be improved upon.

EDIT: There is a problem with rendering some of the LaTeX, which I don't understand. The draft rendered it fine, but not the published version. So I've replaced some LaTeX with unicode or image files; it generally works, but there are oversized images in section 3.

Adequate versus elegant

Part of the problem is resolving human values, is that people have been looking to do it too well and too elegantly. This results in either complete resolutions that ignore vast parts of the values (eg hedonic utilitarianism), or in thorough analyses of a tiny part of the problem (eg all the papers published on the trolley problem).

Incomplete resolutions are not sufficient to guide an AI, and elegant complete resolutions seem to be like most utopias: not any good for real humans.

Much better to aim for an *adequate* complete resolution. Adequacy means two things here:

- It doesn't lead to disastrous outcomes, according to the human's current values.
- If a human has a strong value or meta-value, that will strongly influence the ultimate human reward function, unless their other values point strongly in the opposite direction.

Aiming for adequacy is quite freeing, allowing you to go ahead and construct a resolution, which can then be tweaked and improved upon. It also opens up a whole new space of possible solutions. And, last but not least, any attempt to formalise and write a solution gives a much better understanding of the problem.

Basic framework, then modifications

This post is a first attempt at constructing such an adequate complete resolution. Some of the details will remain to be filled in, others will doubtlessly be changed; nevertheless, this first attempt should be instructive.

The resolution will be built in three steps:

- a) It will provide a basic framework for resolving low level values, or meta-values of the same "level".
- b) It will extend this framework to account for some types of meta-values applying to lower level values.
- c) It will then allow some meta-values to modify the whole framework.

Finally, the post will conclude with some types of meta-values that are hard to integrate into this framework.

1 Terminology and basic concepts

Let H be a human, whose "true" values we are trying elucidate. Let M be the possible environments (including its transition rules), with $\mu \in M$ the actual environment. And let H be the set of future histories that the human may encounter, from time t = 0 onward (the human's *past* history is seen as part of the environment).

Let $R = \{R : H \to R\}$ be a set of rewards. We'll assume that R is closed under many operations - affine transformations (including negation), adding two rewards together, multiplying them together, and so on. For simplicity, assume that R is a real vector space, generated by a finite number of basis rewards.

Then define V to be a set of potential values of H. This is defined to be all the value/preference/reward statements that H might agree to, more or less strongly.

1.1 The role of the AI

The AI's role is elucidate how much the human actually accepts statements in V (see for instance here and here). For any given $v \in V$, it will compute $w_H(v) \geq 0$, the weight of the value v. For mental calibration purposes, assume that $w_H(v)$ is in the 0 to 100 range, and that if the human has no current opinion on v, then $w_H(v)$ is zero (the converse is not true: $w_H(v)$ could be zero because the human has carefully analysed v but found it to be irrelevant or negative).

The AI will also compute $\theta(v) \in [-1,1]^{R \cup V} = \{f : R \cup V \to [-1,1]\}$, the *endorsement* of v. This measures the extent to which v 'approves' or 'disapproves' of a certain reward or value (there is a reward normalisation issue which I'll elide for now).

Object level values are those which are non-zero only on rewards; ie the $v \in V$ for which $\theta(v)(v') = 0$ for all $v' \in V$. To avoid the most obvious self-referential problem, any value's self-endorsement is assumed to be zero (so $\theta(v)(v) = 0$). As we will see below, positively endorsing a negative reward is not the same as negatively endorsing a positive reward: $\theta(v)(-R) = a$ does not mean the same thing as $\theta(v)(R) = -a$.

Then this post will attempt to define the resolution function Θ, which maps weights, endorsements, and the environment to a single reward function. So if $O = [0,100]^V \times [-1,1]^{R \cup V} \times M$ is the cross product of all possible weight functions, endorsement functions, and environments:

$$\Theta : O \to R .$$

In the following, we'll also have need for a more general θ, and for special distributions over H dependent on a given $v \in V$; but we'll define them as and when they are needed.

2 The basic framework

In this section, we'll introduce a basic framework for resolving rewards. This will involve making a certain number of arbitrary choices, choices that may then get modified in the next section.

This section will deal with the problems with human values being contradictory, underdefined, changeable, and manipulable. As a side effect, this will also deal with the fact that humans can make moral errors (and end up feeling their previous values were 'wrong'), and that they can derive insights from philosophical thought experiments.

As an example, we'll use a classic modern dilemma: whether to indulge in bacon or to keep slim.

So let there be two rewards, R_b the bacon reward, and R_s, the slimness reward. Assume that if H always indulges, $R_b = 1$ and $R_s = 0$, while if they never indulge, $R_b = 0$ and $R_s = 1$. There are various tradeoff and gains from trade for intermediate levels of indulgence, the details of which are not relevant here.

Then define $R = \{aR_b + bR_s | a, b \in R\}$.

2.1 Contradictory values

Define $v_1, v_2 \in V$ by $v_1 = \{\text{I like eating bacon}\}$, and $v_2 = \{\text{I want to keep slim}\}$. Given the right normative assumptions, the AI can easily establish that $w_H(v_1)$ and $w_H(v_2)$ are both greater than zero. For example, it can note that the human sometimes does indulge, or desires to do so; but also the human feels sad about gaining weight, shame about their lack of discipline, and sometimes engages in anti-bacon precommitment activities.

The natural thing here is to weight R_b by the weight of v_1 and the endorsement that v_1 gives to R_b (and similarly with v_2 and R_s). This means that

$$\Theta(w_H, \theta, \mu) = w_H(v_1)\theta(v_1)(R_b)R_b + w_H(v_2)\theta(v_2)(R_s)R_s .$$

Or, for the more general formula, with implicit uncurrying so as to write θ as a function of two variables:

$$\Theta(w_H, \theta, \mu) = \sum_{R \in R, v \in V} w_H(v)\theta(v, R)R.$$

For this post, I'll ignore the issue of whether that sum always converges (which it would almost certainly do, in practice).

2.2 Unendorsing rewards

I said there was a difference between a negative endorsement of R, and a positive endorsement of $-R$. A positive endorsement is just a value judgement that sees $-R$ as good, while the negative endorsement just doesn't want R to appear at all.

For example, consider v_3 = {I'm not worried about my weight}. Obviously this has a negative endorsement of R_s, but it doesn't have a positive endorsement of $-R_s$ - it explicitly doesn't have a desire to be fat, either. So the weight and endorsement of v_3 are fine when it comes to reducing the positive weight of R_s, but not when making a zero or negative weight more negative. To capture that, rewrite Θ as:

$$\Theta(w_H, \theta, \mu) = \sum_{R \in R} \max\left(0, \sum_{v \in V} w_H(v)\theta(v, R)\right) R.$$

Then the AI, to maximise H's rewards, simply needs to follow the policy that maximises that reward.

2.3 Underdefined rewards

Let's now look at the problem of underdefined values. To illustrate, add the option of liposuction to the main model. If H indulges in bacon, **and** undergoes liposuction, then both R_b and R_s can be set to 1.

But H might not want to undergo liposuction (assumed, in this model, to be costless). Let R_n be the reward for no liposuction, $R_n = 0$ if liposuction is avoided, and $R_n = -1$ if it happens, and let v_4 = {I want to avoid liposuction}. Extend R to $\{aR_b + bR_s + cR_n | a, b, c \in R\}$.

Because H hasn't thought much about liposuction, they currently have $w_H(v_4) = 0$. But it's possible they may have firm views on it, after some reflection. If so, it would be good to use

those views now. When humans haven't thought about values, there are many ways they can develop them, depending on how the issue is presented to them and how it interacts with their categories, social circles, moral instincts, and world models.

For example, assume that the AI can figure out that, if H is given a description of liposuction that starts with "lazy people can cheat by...", then they will be against it: $w_H(v_4)$ will be greater than zero. However, if they are given a description that starts with "efficient people can optimise by...", then they will be in favour of it, and $w_H(v_4)$ will be zero.

If $w_H(v_4)(t, h)$ is the weight of v_4 at future time t, given the future history $h \in H$, define the discounted future weight as

$$w_H^{\infty}(v_4, h) = \frac{\int_0^{\infty} \gamma^t w_H(v_4)(t, h)\, dt}{\int_0^{\infty} \gamma^t dt}$$

for, say, $\gamma = 999/1000$ if t is denominated in days. If h is the history with the "lazy" description, this will be greater than zero. If it's the history with the "efficient" description, it will be close to zero.

We'd like to use the expected value of $w_H^{\infty}(v_4)$, but there are two problems with that. The first is that many possible futures might involve no reflection about v_4 on the part of H. We don't care about these futures. The other is that these futures depend on the actions of the AI, so that it can manipulate the human's future values.

So define H_{v_4}, a subset of the set of histories H. This subset is defined firstly so that the H will have relevant opinions about v_4: they won't be indifferent to it. Secondly, these are future on which the human is allowed to develop their values 'naturally', without undue rigging and influence on the part of the AI (see this for an example of such a distribution). Note that these need not be histories which will actually happen, just future histories which the AI can estimate. Let p_{v_4} be the probability distribution of future histories, restricted to H_{v_4} (this requires that the AI pick a sensible probability distribution over its own future policy, at least for the purpose of computing this probability distribution).

Note that the exact definition of H_{v_4} and p_{v_4} are vitally important and still need to be fully established. That is a critical problem I'll be returning to in the future.

Laying that aside for the moment, we can define the expected relevant weight:

$$\begin{aligned} W_H(v_4) &= E_{p_{v_4}} w_H^{\infty}(v_4, h) \\ &= \sum_{h \in H_{v_4}} p_{v_4}(h) \frac{\int_0^{\infty} \gamma^t w_H(v_4)(t,h) dt}{\int_0^{\infty} \gamma^t dt}. \end{aligned}$$

Then the formula for Θ becomes:

$$\Theta(w_H, \theta, \mu) = \sum_{R \in R} \max\left(0, \sum_{v \in V} W_H(v)\theta(v, R)\right) R,$$

using W_H instead of w_H.

2.4 Moral errors and moral learning

The above was designed to address underdefined values, but it actually does much more than that. It deals with changeable values, and addresses moral errors and moral learning.

An example of moral error is thinking that you want something, but, upon achieving it, you find that you don't. Let us examine v_2, the desire to be slim. People don't generally have a strong intrinsic desire for slimness just for the sake of it; instead, they might strive for this because they think it will make them healthier, happier, might increase their future status, might increase their self-discipline in general, or something similar.

So we could replace v_2 with v_2 = {I desire X}, where X is something that H believes will come out of slimming.

When computing $W_H(v_2)$ and $W_H(v_2)$, the AI will test how H reacts to achieving slimness, or achieving X, and ultimately compute a low $W_H(v_2)$ but a high $W_H(v_2)$. This would be even more the case if H_{v_2} is allowed to contain impossible future histories, such as hypotheticals where the human miraculously slims without achieving X, or vice-versa.

The use of W_H also picks up systematic, predictable moral change. For example, the human may be currently committed to a narrative that seems themselves as disciplined, stereotypical-rational being that will overcome their short term weaknesses. Their weight $w_H(v_2)$ is high. However, the AI knows that trying to slim will be unpleasant for H, and that they will soon give up as the pain mounts, and change their narrative to one where they accept and balance their own foibles. So the expected $W_H(v_2)$ is low, under most reasonable futures where humans cannot control their own value changes (this has obvious analogies with major life changes, such as loss of faith or changes in political outlooks).

Then there is the third case where strongly held values may end being incoherent (as I argued is the case of the 'purity' moral foundation). Suppose the human deeply believes that v_5 = {Humans have souls and pigs don't, so it's ok to eat pigs, but not ok to defile the human form with liposuction}. This value would thus endorse R_b and R_n. But it's also based on false facts.

There seems to be three standard ways to resolve this. Replacing "soul" with, say, "mind capable of complex thought and ability to suffer", they may shift v_5 to v_6 = {I should not eat pigs}. Or if they go for "humans have no souls, so 'defilement' makes no sense", they may

embrace v_7 = {All human enhancements are fine}. Or, as happens often in the real world when people can't justify their values, they may shift their justification but preserve the basic value: v_8 = {It is natural and traditional and therefore good to eat pig, and avoid liposuction}.

Now, I feel v_8 is probably incoherent as well, but there are no lack of coherent-but-arbitrary reasons to eat pigs and avoid liposuction, so some value set similar to that is plausible.

Then suitably defined H_v would allow the AI to figure out which way the human wants to update their values for v_5, v_6, v_7, and v_8, as the human moves away from the incorrect first values to one of the other alternatives.

2.5 Automated philosophy and CEV

The use of W_H also allows one to introduce philosophy to the mix. One simply needs to include in H_v the presentation of philosophical thought experiments to H, and H's reaction and updating. Similarly, one can do the initial steps of coherent extrapolated volition, by including futures where H changes themselves in the desired direction. This can be seen as automating *some* of philosophy (this approach has nothing to say about epistemology and ontology, for instance).

Indeed, you could define philosophers as people with particularly strong philosophical meta-values: that is, putting a high premium on philosophical consistency, simplicity, and logic.

The more weight is given to philosophy or to frameworks like CEV, the more elegant and coherent the resulting resolution is, but the higher the risk of it going disastrously wrong by losing key parts of human values - we risk running into the problems detailed here and here.

2.6 Meta-values

We'll conclude this section by looking at how one can apply the above framework to meta-values. There are values that have non-zero endorsements of other values, ie $\theta(v, w) \neq 0$.

The previous v_7 = {All human enhancements are fine} could be seen as a meta-value, one that unendorses the anti-liposuction value v_4, so $\theta(v_7, v_4) = -1$. Or we might have one that unendorses short-term values: v_9 = {Short-term values are less important}, with $\theta(v_9, v_1) = -1$.

The problem with comes when values start referring to values that start referring to themselves. This allows indirect self-reference, with all the trouble that that brings.

Now, there are various tools for dealing with self-reference or circular reasoning - Paul Christiano's probabilistic self-reference, and Scott Aaronson's Eigenmorality are obvious candidates.

But in the spirit of adequacy, I'll directly define a method that can take all these possibly self-referential values and resolve them. Those who are not interested in the maths here can skip to the next section; there is no real insight here.

Let $n_V = ||V||$, and let σ be an ordering (or a permutation) of V, ie a bijective map from $\{1, 2, \dots, n_V\}$ to V. Then recursively define W_H^σ by $W_H^\sigma(\sigma(1)) = W_H(\sigma(1))$, and

$$W_{\mathbf{H}}^{\sigma}(\sigma(i)) = \max\left(0, W_{\mathbf{H}}(\sigma(i)) + \sum_{j=1}^{i-1} W_{\mathbf{H}}^{\sigma}(\sigma(j))\theta(\sigma(j), \sigma(i))\right).$$

Thus each W_H^σ is the sum of the actual weight W_H, plus the σ-adjusted endorsements of the values preceding it (in the σ ordering), with the zero lower bound. By averaging across the set P(V) of all permutations of V, we can then define:

$$\widehat{W}_{\mathbf{H}}(v) = \sum_{\sigma \in P(\mathcal{V})} \frac{1}{n_{\mathcal{V}}!} W_{\mathbf{H}}^{\sigma}(v).$$

Then, finally, for resolving the reward, we can use these weights in the standard reward function:

$$\Theta(w_{\mathbf{H}}, \theta, \mu) = \sum_{R \in \mathcal{R}} \max\left(0, \sum_{v \in V} \widehat{W}_{\mathbf{H}}(v)\theta(v, R)\right) R.$$

3. The "wrong" Θ: meta-values for the resolution process

The Θ of the previous section is sufficient to resolve the values of an H which has no strong feelings on how those values should be resolved.

But many H may find it inadequate, filled with arbitrary choices, doing too much by hand/fiat, or doing to little. So the next step is to let H's values affect how the Θ itself works.

Define Θ_0 as the framework constructed in the previous section. And let Ω be the set of all such possible resolution frameworks. We now extend θ so that $\theta(v)$ can endorse or unendorse not only elements of V and R, but also of Ω.

Then we can define

$$\Theta^\theta = \sum_{\Theta^i \in \Omega, v \in \mathcal{V}} \widehat{W}_{\mathbf{H}}(v)\theta(v, \Theta^i)\Theta^i,$$

and define Θ itself as

$$\Theta = \Theta_0 + \Theta^\theta.$$

These formulas make sense, since the various elements of Ω takes values in R, which can be summed. Also, because we can multiply a reward by a positive scalar, there is no need for renormalising or weighting in these summing formulas.

Now, this is not a complete transformation of Θ according to H's values - for example, there is no place for these values to change the computation of Θ^θ, which is computed according to the $\hat{W}_H$ previously defined for Θ_0. (Note: Those are where the LaTeX errors used to be, and now there are oversized image files which I can't reduce, sorry!)

But I won't worry about that for the moment, though I'll undoubtedly return to it later. First of all, I very much doubt that many humans have strong intuitions about the correct method for resolving contradictions among the different ways of designing a resolution system for mapping most values and meta-values to a reward. And if someone does have such a meta-value, I'd wager it'll be mostly to benefit a specific object level value or reward, so it's more instructive to look at the object level.

But the real reason I won't dig too much into those issues for the moment, is that the next section demonstrates that there are problems with fully self-referential ways of resolving values. I'd like to understand and solve those before getting too meta on the resolution process.

4 Problems with self-referential Θ

Here I'll look at some of the problems that can occur with fully self-referential Θ and/or *v*. The presentation will be more informal, since I haven't defined the language or the formalism to allow such formulation yet.

4.1 All-or-nothing values, and personal identity

Some values put a high premium on simplicity, or on defining the whole of the relevant part of Θ. For example, the paper "An impossibility theorem in population axiology..." argues that total utilitarianism is the only the theory that avoids a series of counter-intuitive problems.

Now, I've disagreed that these problems are actually problems. But some people's intuitions strongly disagree with me, and feel that total utilitarianism is justified by these arguments. Indeed, I get the impression that, for some people, even a small derogation to total utilitarianism is bad: they strongly prefer 100% total utilitarianism to 99.99% total utilitarianism + 0.01% something else.

This could be encoded as a value v_{10} = {I value having a simple populations ethics}. This would provide a bonus based on the overall simplicity of the image of Θ. To do this, we have introduced personal identity (an issue which I've argued is unresolved in terms of reward functions), as well as about the image of Θ.

Population ethics feels like an abstract high-level concept, but here is a much more down-to-earth version. When the AI looks forwards, it extrapolates the weight of certain values based on the expected weight in the future. What if the AI extrapolates that $w_H^\infty(v_i)$ will be either 0 or 10 in the future, with equal probability? It then reasonably sets $W_H(v_i)$ to 5.

But the human will live in one of those futures. The AI will be maximising their 'true goals' which include $W_H(v_i) = 5$, while H is forced into extreme values of w_H (0 or 10) which do not correspond to the value the AI is currently maximising. So v_{11} = {I want to agree with the reward that Θ computes} is a reasonable meta-value, that would reward closeness between expected future values and actual future values.

In that case, one thing the AI would be motivated to do, is to manipulate H so that they have the 'right' weights in the future. But this might not always be possible. And H might see that as a dubious thing to do.

Note here that this is not a problem of desiring personal moral growth in the future. Assuming that can be defined, the AI can then grant it. The problem would be wanting personal moral growth *and* wanting the AI to follow the values that emerge from this growth.

4.2 You're not the boss of me!

For self-reference, we don't need Gödel or Russell. There is a much simpler, more natural self-reference paradox lurking here, one that is very common in humans: the urge to not be told what to do.

If the AI computes $\Theta(w_H, \theta, \mu) = R$, there are many humans who would, on principle, declare

and decide that their reward was something other than R. This could be a value $v_{12} = \{\Theta$ incorrectly computes my values$\}$. I'm not sure how to resolve this problem, or even if it's much of a problem (if the human will disagree equally no matter what, then we may as well ignore that disagreement; and if they disagree to different degrees in different circumstances, this gives something to minimise and trade-off against other values). But I'd like to understand and formalise this better.

5 Conclusion: much more work

I hope this post demonstrates what I am hoping to achieve, and how we might start going about it. Combining this resolution project, with the means of extracting human values would then allow the Inverse Reinforcement Learning project to succeed in full generality: we could then have the AI deduce human values from observation, and then follow them. This seems like a potential recipe for a Friendly-ish AI.

On Defense Mechanisms

There is much talk about cognitive heuristics and biases in the Rationality community, whereas psychodynamic/psychoanalytic perspectives of psychology tend to be dismissed as "bad science". Now, I will agree that some Freudian ideas like penis envy and Oedipus complex are silly -- but that doesn't necessarily mean that nothing useful can be salvaged from the wreck. In particular, today I want to highlight the concept of *psychological defense mechanisms* developed by Sigmund Freud and his daughter Anna, and suggest how they might be of interest to rationalists.[1]

According to Freud's theory, our unconscious minds are host to the continual battles between the "id" (our primal, pleasure-driven aspect), the "ego" (our decision-making aspect, mindful of social reality) and the "superego" (our moral compass). These internal conflicts, when prolonged, result in anxiety and other unpleasant emotions like guilt. In order to get rid of this anxiety, we rely on "defense mechanisms", which are feats of mental gymnastics that work through self-deception. And since self-deception can be an obstacle to epistemic and instrumental rationality, it should be clear why we should at least be aware of defense mechanisms.

Here are some of the most common defense mechanisms, taken from Weiten, Dunn & Hammer (2015).[2]

1. **Repression**: keeping distressing thoughts and feelings below your conscious awareness. *Examples*: Amber forgot the name of someone whom she really hates. Billy doesn't remember the time he nearly got killed while serving in Afghanistan.
2. **Projection**: attributing your own thoughts and feelings to someone else. *Examples:* Claire feels sexual tension with a colleague, which she attributes to the colleague's motive to seduce her. Dennis doesn't like his boss, but tells others that he actually likes the boss -- it's just that the boss doesn't like Dennis.
3. **Displacement**: re-directing emotions from their original source to a substitute target. *Examples:* Emily had a rough day at work, so she unleashes her anger onto her husband and cat. Frank has been disciplined by his parents, and takes his anger out on his little sister.
4. **Reaction formation**: behaving in a way that is the opposite of what your true feelings would imply. *Examples:* Gemma unconsciously resents her child, so she spoils the child with gifts. Harry speaks out against homosexuality, but has latent homosexual impulses of his own.
5. **Regression**: reverting to immature behavioral patterns. *Examples:* Irma has trouble finding a new job, but boasts about her massive talent in an exaggerated way. Jason throws a temper tantrum when he doesn't get his way.
6. **Rationalization**: creating false yet plausible excuses to justify socially unacceptable behavior. *Examples:* Kate binge-watches Netflix instead of studying because "studying more wouldn't help her anyway". Leon cheats someone in a business transaction because "everybody does it".
7. **Intellectualization**: looking at difficulties in a detached and abstract way to suppress your emotional reactions. *Examples:* Molly has been diagnosed with a terminal illness, so she tries to learn as much as possible about the disease's details and treatment. Nathan is deep in debt due to overspending, so he creates a complex spreadsheet of how long it would take to repay with different interest rates and payment options.

8. **Identification**: boosting your self-esteem by forming alliances (real or imaginary) with some person or group. *Examples:* Ophelia identifies with a number of famous rock stars, movie stars and athletes. Peter is an insecure college student who joined a fraternity to bolster his self-worth.
9. **Denial**: refusing to acknowledge the painful realities in your life. *Examples:* Quinn is failing a class required for graduation, yet allows her family to plan a trip to her graduation. Ray abuses alcohol but refuses to admit he has a problem.
10. **Fantasy**: fulfilling your wishes and impulses in your imagination. *Examples:* Samantha is unpopular but imagines that she has a large network of outgoing and popular friends. Tim is being bothered by a bully, but instead of taking action to stop it he daydreams about killing the bully.
11. **Undoing**: trying to counteract feelings of guilt through acts of atonement. *Examples:* Ursula compliments her mother's appearance each time after she insults her mother. Vernor dislikes his professor, so he gives the professor an apple as a gift.
12. **Overcompensation**: compensating for deficiencies (real or imagined) by focusing on or exaggerating positive characteristics. *Examples:* Wilma is a transfer student who hasn't made new friends, so she focuses on doing well in class. Xavier strives for status, power and wealth as ways to cover up his feelings of inferiority.

This list is not exhaustive; other defense mechanisms include isolation, introjection, reversal, splitting, acting out, passive-aggression, and sublimation, for example.[3]

What unites these various defense mechanisms is that they are common even in psychologically healthy people (akin to how all humans are vulnerable to cognitive bias), but can become problematic when we rely on them excessively. Moreover, defense mechanisms play a prominent role in *defensive coping*, which is a common albeit usually counterproductive and maladaptive response to stress. As coping strategies, they shield us from uncomfortable emotions like anxiety, guilt, anger, and dejection -- but this comes at the price of distorting our perceptions of reality, often in self-serving ways. Furthermore, they rarely provide actual solutions to our problems, and ironically may increase anxiety.

A superior alternative is to use *constructive coping* techniques. These are action-oriented, realistic, and require self-control. For example, when you're experiencing stress or burnout, you could try to regulate your emotions by exercising, meditating, writing about your experiences in a personal journal, or forgiving others. You could try to solve the problem that caused the stress by brainstorming solutions, seeking social support, or improving your time management. Finally, you could try to change your appraisal (evaluation) of the situation by reinterpreting events in a positive way, using humor, or avoiding negative self-talk like catastrophizing.

Question for discussion: How would *you* suggest we use the idea of defense mechanisms in theory or practice?[4]

Notes

[1] I was surprised that this topic hasn't been discussed explicitly on LW before.

[2] "Psychology Applied to Modern Life: Adjustment in the 21st Century" -- I have a summary with notes of this book on my blog, here.

[3] See the Wikipedia article.

[4] Scott Alexander seems skeptical but admits it can sometimes be useful.