Map and Territory Cross-Posts

1. Map and Territory: a new rationalist group blog
2. In Defense of Kegan
3. A Foundation for The Multipart Psyche
4. Internalizing Existentialism
5. Fluid Decision Making
6. The Developmental Role of Play
7. Revealed and Stated Identity
8. Debate and Dialectic
9. Act into Fear and Abandon all Hope
10. Nothing is Forbidden, but Some Things are Good
11. Phenomenological Complexity Classes
12. What Value Hermeneutics?
13. What Value Epicycles?
14. Unstaging Developmental Psychology
15. The Personal Growth Cycle
16. Angst, Ennui, and Guilt in Effective Altruism
17. Developmental Psychology in The Age of Ems
18. What Value Subagents?
19. Inscrutable Ideas
20. Embracing Metamodernism
21. Is Feedback Suffering?
22. Cognitive Empathy and Emotional Labor
23. Regress Thyself to the Mean
24. Doxa, Episteme, and Gnosis
25. Avoiding AI Races Through Self-Regulation
26. Evaluating Existing Approaches to AGI Alignment
27. Suffering and Intractable Pain
28. Akrasia is confusion about what you want
29. Let Values Drift
30. Scope Insensitivity Judo
31. Normalization of Deviance
32. You are Dissociating (probably)
33. Forcing yourself to keep your identity small is self-harm

Map and Territory: a new rationalist group blog

If you want to engage with the rationalist community, LessWrong is mostly no longer the place to do it. Discussions aside, most of the activity has moved into the diaspora. There are a few big voices like Robin and Scott, but most of the online discussion happens on individual blogs, Tumblr, semi-private Facebook walls, and Reddit. And while these serve us well enough, I find that they leave me wanting for something like what LessWrong was: a vibrant group blog exploring our perspectives on cognition and building insights towards a deeper understanding of the world.

Maybe I'm yearning for a golden age of LessWrong that never was, but the fact remains that there is a gap in the rationalist community that LessWrong once filled. A space for multiple voices to come together in a dialectic that weaves together our individual threads of thought into a broader narrative. A home for discourse we are proud to call our own.

So with a lot of help from fellow rationalist bloggers, we've put together Map and Territory, a new group blog to bring our voices together. Each week you'll find new writing from the likes of Ben Hoffman, Mike Plotz, Malcolm Ocean, Duncan Sabien, Anders Huitfeldt, and myself working to build a more complete view of reality within the context of rationality.

And we're only just getting started, so if you're a rationalist blogger please consider joining us. We're doing this on Medium, so if you write something other folks in the rationalist community would like to read, we'd love to consider sharing it through Map and Territory (cross-positing encouraged). Reach out to me on Facebook or email and we'll get the process rolling.

https://medium.com/map-and-territory

In Defense of Kegan

This is a linkpost for https://mapandterritory.org/in-defense-of-kegan-2ed7ab51b4c8

NB: Originally published on Map and Territory on Medium. This is an old post originally published on 2016-09-10. It was never previously cross-posted or linked on LessWrong, so I'm adding it now for posterity. It's old enough that I can no longer confidently endorse it, and I won't bother trying to defend it if you find something wrong, but it might still be interesting.

I find Kegan's model of psychological development extremely useful. Some folks I know disagree on various grounds. These are some accumulated responses to critiques I've encountered.

Before we dive into these critiques, though, allow me to attempt a brief introduction to the theory (though this is a tough undertaking, as we'll discuss below). Robert Kegan, later along with Lisa Lahey, put forward a theory of developmental psychology rooted in complexity of meaning making. It is influenced and builds on the work of Piaget, but also Erikson and Kohlberg, who extended developmental psychology to consider the possibility of adult development.

Kegan's theory focuses on the maximally complex models people can make of the world in near construal mode. Development is along a continuous gradient but with clear "levels" where a particular kind of complexity is fully available to the thinker. These are classified from 1 to 5 and can be summarized in many ways, though fundamentally they correspond to when a person can form fully-articulable models of things, relationships between things, systems, relationships between systems, and systems of systems (holons), respectively.

This is an exceedingly dense introduction, though, and I know of no good short explanation. The best resources on the topic remain Kegan's seminal *The Evolving Self* and his later *In Over Our Heads* for a more approachable, example-laden presentation.

The first, and perhaps strongest, critique of Kegan is that it's very hard for anyone to explain it. Kegan begins *In Over Our Heads* with the story of getting a letter from a student assigned to read *The Evolving Self*. The student writes that *The Evolving Self* is full of interesting ideas but gets so frustrated trying to make sense of it that he wants to "punch [Kegan] in the teeth".

Partly this is because of Kegan has a strong literary and classics background, so *The Evolving Self* is full of very precise language with subtle meanings, many so subtle that Kegan takes multipage digressions to explain them. But *In Over Our Heads* and his later books written with Lahey and other coauthors use more familiar language and yet still leave people confused.

The theory seems to defy simple explanation. I've yet to find one written by Kegan, Lahey, or anyone else that was able to reliably convey in less than 40,000 words a reasonably coherent and complete view of it. As one person I know put it, the theory reads to him as analogous to someone saying there are invisible dragons, undetectable by readily available means, but which you will notice if you devote at least 20 hours to the study of invisible dragons.

Yet, those of us who have put in the time to “see the invisible dragons” tend to be pretty excited about the theory. Kegan gives us a way to understand and construct many aspects of human behavior and thought that time and again prove consistent and reflective of reality. So if it works so well, why is it so hard to explain?

There are a few ways things can be hard to explain. One is that they are unintuitive. Physics is like this: we perceive the world as if it operated the way Aristotle imaged it worked, but it turns out this approximation breaks down at extremes and the quest to find a complete theory forces us to consider ever more exotic phenomena.

Another way things may be hard to explain is that they’re complicated. Machines and living things are like this, with engineers and biologists mostly struggling to make clear what’s happening in systems where lots of details matter. A clock might fail if it’s missing a tooth on one gear or a frog might die if it’s missing a sequence in its DNA, and understanding why is a messy business of picking through tightly interwoven threads of causality.

But perhaps the most vexing way something can be hard to explain is when it’s complex. That is to say, even if it has few details and works in a straightforward manner, thinking through how it works is still hard. Game theory, economics, and most everything touched by mathematics is like this: just a few “simple” rules lead to bewildering complexity under combination.

So when trying to explain a theory like Kegan’s that has at its heart a developmental progression in human capacity to cope with complexity, it’s perhaps unsurprising that the complexity can collapse back in on itself and make the theory look like disjointed rubble. The theory, in fact, predicts this, because it’s one about the relationships between systems (i.e. the change in human meaning making over time), so by its own expectations it will prove difficult to gain an intuitive grasp on without the reader having themselves first already attained the capacity to naturally reason about relationships between systems in near construal mode (level 4 in Kegan’s model).

To most people this feels like the theory saying “you can’t understand it until you already understand it”, but there’s more going on here. It’s instead saying that Kegan’s developmental theory belongs to a class of things that cannot be fully understood without the ability to naturally, intuitively work with the relationships between systems. Without that ability it may be understood in other ways, in particular using far construal mode, but that is demanding on the level of learning algebra, calculus, or differential equations, which is to say something that even the brightest among us struggle with.

But if it’s really this hard, why do people feel they can reject Kegan when they can’t reject, say, abstract algebra in the same way? They may find they completely lack the capacity to understand what’s going on in abstract algebra in near mode, yet aside from a few mathematicians with technical objections, no one thinks abstract algebra fails to model well the parts of reality it is attempting to model whether they understand it or not. At worst it’s just some of that “math stuff” other people worry about but they don’t “get”.

The difference, as Robin Hanson has observed in the general case, is that Kegan is a theory about stuff we are intimately familiar with: people. We are happy to defer to experts and theories we don’t understand on topics we don’t feel we have much of a grasp on, like abstract algebra, but as things get progressively more “real” we feel

less inclined to trust complex theories experts put forward that we don't understand ourselves.

There's a sort of escalating scale of how many people don't trust experts that's a function of distance from lived experience, social agreements on who has expertise, and availability of evidence to check our understanding. Basically everyone trusts experts in math because it's far from lived experience and we agree that mathematicians are the math experts even though we can only easily validate the veracity of the simplest mathematical claims without training.

Most, but slightly fewer, folks trust experts in physics. People agree that physicists are the experts and have lots of evidence to prove their unintuitive theories are right (planes fly, electricity powers our devices, computers work with no moving parts). The only difficulty for physicists is that we all live physics, so there's a constant battle against violations of intuition they must overcome to convince us of their theories.

Less trusted still are doctors, economists, and philosophers. Somewhere between economists and philosophers we find anyone attempting to explain human behavior. We all have lots of experience with it, there's lots of evidence around to check against, so the only thing holding up the experts is that we agree their expertise exists because someone gave them an advanced degree in it.

So in general people feel free to reject arguments about human behavior that don't seem intuitive even if they are provided by experts. It's for the same meta-reason that no one listens to the economists and philosophers have been engaged in the same discussions for millennia: it feels easy to reject what doesn't feel true when it's something we have a lot of experience with and can easily gather data on.

Is this why folks who find it hard to understand Kegan often choose to reject it? I suspect probably yes, but then again I'm asking you to accept my argument about human behavior concerning belief strength in theories people don't fully comprehend and are not expert in, so I'll leave my response to this critique here before I ascend too far up a house of cards.

So suffice to say, Kegan is complex, complex enough that it's predictably hard to understand, and about a topic where we little trust experts.

Kegan is sometimes presented as "wrong" because it's not always accurate. That is to say, because it's a theory that presents a model of the world, it has edge cases at which it seems to break down. This is a standard objection to all models in all domains and is uninteresting, but since there is a high likelihood of confusion due to lack of trust in expertise here, it's worth covering.

A model is some explanation and prediction of how the world works. For example, atomic theory gives us a way of understanding matter as indivisible (atomic) particles. Like all theories, it's "wrong" in that reality is not actually made up of atoms—it's just reality. Instead atoms are a way of understanding reality that let us explain phenomena we see and predict future phenomena with some degree of accuracy. To the extent that atomic theory predicts what happens in reality, it is useful to the purpose of predicting future events. This doesn't make it "right", just predictive enough for our needs.

When atomic theory fails to make correct predictions it's not "wrong". Instead it's that the theory is not complete because it's a model and not reality and the only perfect

model of reality is reality itself, just as the only perfect map of the Earth is the Earth itself.

So Kegan's developmental theory is naturally not a perfect predictor of reality. We can only judge it by how accurate it is for the things we want to use it for. Whether or not it's accurate enough to be useful is what we'll explore in the remaining critiques.

The remaining two major objections are technical in that they assume an understanding of Kegan and find problems on internal grounds. Feel free to just skip to the end if these are not of interest to you.

The first problem is that Kegan differentiates expectations of what you can do in near and far mode. I'll note here, though, that Kegan does not explicitly reference construal level theory, dual process theory, or any other two part theory of mind. This is mostly an artifact of its time: Kegan wrote *The Evolving Self* before these theories were well developed, and instead spends a decent number of words to explain that he's focused on the capacity of intuitive, immediate, natural ratiocination.

Having lacked a referent to contrast near and far modes, some people naturally object to the theory on the grounds that mathematicians, for example, show incredible capacity to reason about holons in their 20s despite lacking the behavior patterns expected of someone with this capacity. The difference, of course, is that mathematicians do their work in far mode, and are in fact exceptionally talented at thinking in far mode, but because far mode is not used to engage in day-to-day activity because it's too complicated to fit in far mode, that far mode capacity for handling greater complexity does not extend to near mode.

It's unclear whether capacity to handle complexity in near mode extends to far mode. It seems likely but there's not much data on, for example, people becoming mathematicians in their 50s after struggling with math for the previous 5 decades.

The second objection is that Kegan is not directly testable because it's a theory about changes in the way of meaning making which is inherently unobservable since it exists only as a dialectic between perception and reality. While it may be true that you can't directly test if the model is how reality is structured, this is a problem for all theories of mind and it has the same solution as they do: you can test the predictions. We can check whether the expected behavior of people at particular Kegan levels correlates with their actual behavior.

There's unfortunately very little data on this. About the best we have comes from Lahey and her work in applying Kegan's model to education reform and management consulting, and most of the available data I'm aware of is collected post level assessment or informally, so it's suspect. I happily concede this is a major issue and would love to see more data collected but consider it unlikely because in the past 30 years the theory has gained little traction, largely it seems due to its complexity, so not enough people are working in the area to generate the needed data to sufficiently test the theory.

I've tried to address here the most common objections I've encountered to Kegan's theory. If there are additional objection categories I've left out that you notice, feel free to bring them up in the comments, and I'll see if they are tractable problems or tear the whole thing down.

If reading this has piqued your interest in Kegan's work, I highly recommend reading *In Over Our Heads* and *The Evolving Self* in that order. For applications of the theory

you can check out Kegan and Lahey’s later works and for a philosophical incorporation of Kegan I suggest reading David Chapman.

A Foundation for The Multipart Psyche

This is a linkpost for https://mapandterritory.org/a-foundation-for-the-multipart-psyche-79a66292a7a

NB: Originally published on Map and Territory on Medium. This is an old post originally published on 2016-09-14. It was never previously cross-posted or linked on LessWrong, so I'm adding it now for posterity. It's old enough that I can no longer confidently endorse it, and I won't bother trying to defend it if you find something wrong, but it might still be interesting.

In a recent post Scott Alexander gives a review of some recent results in neurobiology that suggest a powerful, unifying set of mechanisms for how information is integrated in the brain. I recommend you read his article and the original research if you can, but I'll summarize it briefly.

There are various chemicals regulating activity in the brain. There is now evidence that these chemicals act in coordinating an information pump in the brain. Change the chemicals and you change the parameters of the information pump. It seems specifically the information pump in play fits the Bayesian model in that certain chemicals regulate the presentation of prior evidence, others new evidence, and yet others confidence in those evidences.

What I find compelling is that the model described provides a plausible mechanism by which the theory of a 2-part psyche might work. There are several two-part theories of psyche, that is to say theories of how mental processes are organized. My preferred one is near/far construal theory, but there is also the S1/S2 distinction, the fast/slow distinction, in Chinese philosophy yin and yang, and even the hot/cold blood model from medieval European thought. Each of these acts as a way of classifying thoughts and behaviors along a spectrum between two extremes.

The interesting thing about the 2-part psyche theories, and why I prefer the near/far distinction, is that they all seem to operate along the same dimension. Near/far uses the metaphor of distance (because it happens we use similar reasoning patterns when working with things that are physically near versus far) to differentiate between things that are heavy on details and light on patterns versus those that are heavy on patterns and light on details. S1/S2 uses basically the same dimension, as does fast/slow, yin/yang, and hot/cold: stuff with lots of details is near, fast, yin, hot, and part of S1 while stuff with less details and stronger patterns are far, slow, yang, cold, and part of S2. This suggests that they are all pointing at the same sort of thing, though in slightly different ways.

And, as it happens, this is basically the same dimension along with chemicals in the brain seem to affect cognition, balancing between how much to weigh new evidence (details) against prior evidence (patterns). So it seems that we now have a plausible biological basis for the two-part psyche we've reasoned exists and find useful, whereas before it was just a pattern that worked without strong evidence of a mechanism.

The 3-Part Psyche

So that takes care of the 2-part psyche, but what about the arguably more popular 3-part psyche model. The 3-part model dates back at least to Aristotle in the West and the gunas in India, was revitalized by Freud, and has bloomed into various descendent theories in modern psychology such as Internal Family Systems. Each version has different boundaries and explanations, so for simplicity I'll use Freud's well-known terminology.

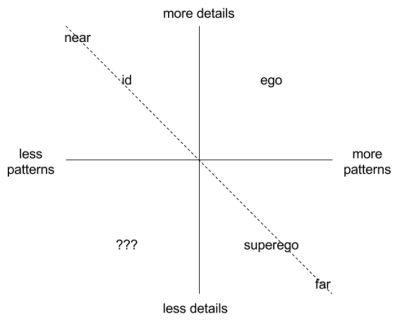
Briefly, these theories all see roughly the same three parts in the psyche: the id, the ego, and the superego. The id is the part that acts and responds "on instinct", the ego is the part that is "rational" and integrates the other two, and the superego is the part that operates on "moral" grounds. These parts are viewed as working in relation to one another, frequently in opposition, with what someone does and thinks arising from their interaction.

These theories connect with the 2-part psyche model in that near corresponds to aspects of the id and ego and far corresponds to aspects of ego and superego. When we see this kind of correspondence with superposition, it suggests both are mixing up the underlying reality in different ways. We can use this to try to pick apart what's really going on.

The commonalities of id and ego seem to be an inclusion of details, same as for near. What's different is that ego has a concept of integration with patterns whereas near and id do not.

The commonalities of ego and superego are just the opposite: inclusion of patterns. Same goes for far. The differences are that ego includes details while far and superego do not.

I propose from this that if we separate out details and patterns onto separate dimensions we can get a 2-dimensional model that captures both the 2-part and 3-part psyche models and even suggests a 4-part psyche model.



Now the 2-part model corresponds to the line between the more details, less patterns corner of the space and the less details, more patterns corner with near and far as their division down the middle, respectively. This is drawn as a dotted line in the above chart.

The 3-part models corresponds to 3 of the 4 quadrants formed around the middle of the 2-dimensional space: id is more details with less patterns, ego is more details with more patterns, and superego is more patterns with less details. This also leaves a suspiciously empty 4th quadrant to be part of the psyche with less details and less patterns.

And, to make things even better, this fits with the biological model Scott summarizes: there are chemicals regulating how much to favor details and how much to favor patterns. Normal thinking and behavior fall in the ego quadrant or at least near the center while mental disorders appear when the chemical regulation of detail and pattern strength are out of their typical balance.

So for all their faults in the past, maybe our theories of the psyche have been pointing us in the right direction all along, just in a confused way.

The 4th Quadrant

This still leaves us with the fourth quadrant that's gone unaddressed. Here I'll offer some brief speculation on what it might be before wrapping up.

Since this theory points to something that will feel qualitatively different to us from the inside the way id, ego, and superego do when both details and patterns are weak, we should go looking for things we consider mental states that don't fit well in the existing 2-part or 3-part models. One immediately comes to mind: dreams.

Dreams are, among other things, a time when you have low sensory information and seem to have trouble completing patterns. We talk about "dream logic" because in dreams you can jump between fitting patterns to limited data that often violate the causal narrative we expect to find in our thinking. And dreams seem to incorporate memories, often recent and important memories, in place of outside sensory data. This is by no means a slam dunk, but it does weakly fit the evidence.

Which is the point on which I wish to end: all of this is based on fairly weak evidence. Although the 2-part and 3-part psyche models are fairly robust, they have always had problems because they readily fall apart upon rigorous inspection and have not had a clear biological basis so are subject to introspection bias. Additionally, my new evidence from Scott is an interpretation of an interpretation of recent findings and stretches well beyond what we can safely conclude.

At the same time, these ideas are exciting and, I think, worth exploring because they give us a potential model for better understanding human thoughts and behavior. I fully expect this 2-dimensional, 4-part psyche of details and patterns to make wrong predictions, but I'm hopeful it makes more right predictions and fewer wrong predictions than either 2-part or 3-part psyche models do. I look forward to testing them and seeing what more we can learn about our messy selves.

Internalizing Existentialism

This is a linkpost for https://mapandterritory.org/internalizing-existentialism-72831ef04735

NB: Originally published on Map and Territory on Medium. This is an old post originally published on 2016-09-18. It was never previously cross-posted or linked on LessWrong, so I'm adding it now for posterity. It's old enough that I can no longer confidently endorse it, and I won't bother trying to defend it if you find something wrong, but it might still be interesting.

Over the last couple months, due to reading Daoist philosophical texts, I've come to deeply internalize something I've known for a long time: morality doesn't exist "out there" in reality and is instead a construct of our preferences and the dialectic between different people's preferences.

If you stumbled upon this and didn't realize morality wasn't essential, well, um, I'm not going to try to convince you of that. Probably a not terrible reading recommendation is the Less Wrong series on metaethics.

I started down the path to giving up an internal sense of essential morality when meditating on the Daoist position that there is fundamentally no differentiation. For example, chapter 41 of the Daodejing reads in part:

> Thus it is said:
> The path into the light seems dark,
> the path forward seems to go back,
> the direct path seems long,
> true power seems weak,
> true purity seems tarnished,
> true steadfastness seems changeable,
> true clarity seems obscure,
> the greatest are seems unsophisticated,
> the greatest love seems indifferent,
> the greatest wisdom seems childish.

And in chapter 20 we find:

> Stop thinking, and end your problems.
> What difference between yes and no?
> What difference between success and failure?
> Must you value what others value,
> avoid what others avoid?
> How ridiculous!

And in both the texts of Zhuangzi and Liezi we are given multiple stories where beauty and good acts do not lead to happiness and ugliness and wickedness do not hinder virtue. On the surface we are given contradictions, but by looking deeper the contradictions dissolve if we perceive that the dichotomy is false.

Even speaking of virtue is itself an interesting case. The word used in Chinese, 德 or de, means virtue with a moralistic component in normal use just as is found in English, but de also has a meaning of step and shares with virtue's Latin roots in meaning

strength or capacity. So even here we find, when it looks as though we are being given moral advice, it only seems that way if we take it to be that: take away the perception of morality and we are given possible steps along the path.

With this in mind, I set out to experiment with removing my use of moralistic language. We tend to say things are good or bad when really what we mean is that we like them or we don't. And if I want to find out if morality really does not exist as an essential property of the universe, it's worthwhile to try to take it out of my language and see if it comes up missing.

So I have tried to do this. I try to no longer say things are good or bad, and instead try to say I like or dislike things, or I want more or less of things. And aside from having a hard time breaking the habit of using common phrases that happen to contain "good" or "bad" like saying "this tastes good" to mean "I like how this tastes", it's proven very straight forward and thrown into contrast those times when I was projecting my own preferences onto the universe.

This projection happens through the turn of phrase. If I think what my friend is wearing is ugly and and I say to them "that looks bad", I'm implicitly suggesting their appearance goes against an external measure of style. But if I say "I don't like what you're wearing", I have to be the owner of the preference, and I know it's not living out in the universe apart from me. And if we look deeper, there's no sense in which something can "look good" if there is no observer to assess the quality, so it seems through language we casually mistake preferences for essences.

And so I have now more internalized the existential nature of morality I have long intellectually known.

Fluid Decision Making

This is a linkpost for https://mapandterritory.org/fluid-decision-making-d8e4e41a51c0

NB: Originally published on Map and Territory on Medium. This is an old post originally published on 2016-10-04. It was never previously cross-posted or linked on LessWrong, so I'm adding it now for posterity. It's old enough that I can no longer confidently endorse it, and I won't bother trying to defend it if you find something wrong, but it might still be interesting.

A lot of folks these days talk about “flow” to mean some kind of mystical state where they experience something like automatic decision making where they get out of their own way and just do. Less mysteriously, other folks use “flow” to describe periods of focused attention. More mysteriously, some folks talk about something similar but as the Daoist idea of action through non-action. Let’s see if we can make some sense of what’s going on here.

As far as I can tell we talk about flow because Daoist philosophy explains virtuous behavior (de) as being a mind like water. Chapter 78 of the Daodejing reads:

> Nothing in the world
> is as soft and yielding as water.
> Yet for dissolving the hard and inflexible,
> nothing can surpass it.
>
> The soft overcomes the hard;
> the gentle overcomes the rigid.
> Everyone knows this is true,
> but few can put it into practice.
>
> Therefore the Master remains
> serene in the midst of sorrow.
> Evil cannot enter his heart.
> Because he has given up helping,
> he is people’s greatest help.
>
> True words seem paradoxical.

And Chapter 15 reads:

> The ancient Masters were profound and subtle.
> Their wisdom was unfathomable.
> There is no way to describe it;
> all we can describe is their appearance.
>
> They were careful
> as someone crossing an iced-over stream.
> Alert as a warrior in enemy territory.
> Courteous as a guest.
> Fluid as melting ice.
> Shapable as a block of wood.
> Receptive as a valley.
> Clear as a glass of water.

> Do you have the patience to wait
> till your mud settles and the water is clear?
> Can you remain unmoving
> till the right action arises by itself?
>
> The Master doesn't seek fulfillment.
> Not seeking, not expecting,
> she is present, and can welcome all things.

Though this is not necessarily the direct lineage of the modern usage, this gives a sense of what metaphors of being liquid-like are trying to imply. In flow a person acts naturally, takes the most direct path, and problems yield before them. It's reported to feel from the inside like achieving without trying, and also like a full integration of knowledge into action that does what's intended. This integration is where I think we can get a grasp on what is happening in flow.

I previously explored a 2-dimensional theory of psyche along the dimensions of detail and pattern. In brief, we can model the brain (in part because it now appears it may physically work this way) as operating by giving high or low weight to details and high or low weight to patterns. When details are high and patterns are low, it's what we might term near construal mode, S1, or the id. When details are low and patterns are high, it's what we might term far construal mode, S2, or the superego. And when details and patterns are high, it's what we might term an integrated mode or the ego.

This integration of details and patterns allows a balanced approach to updating and acting on information. Too much focus on detail and there's an overreaction to specifics that ignores known patterns. Too much focus on pattern and there's a failure to account for specific circumstances. But it is also not enough to integrate details and patterns: they must be each weighted appropriately to result in confident action.

We know that simply integrating the two is not enough because we experience cognitive dissonance all the time. Admittedly there is cognitive dissonance among competing patterns and among contradictory details, but what I'm focused on here is the disagreement of patterns and details, like in this apocryphal story of a physics class:

> The period bell rings and the students shuffle into another day of high school physics. The teacher is standing over by the radiator balancing a 1-inch think metal sheet against it. She asks a student to come up and observe the metal sheet. He looks at it, and at the teacher's invitation, touches each side of the sheet. To his surprise, the side away from the radiator is hotter than the side toward the radiator! He returns to his seat and the teacher asks the students what's going on.
>
> "Air currents convecting heat to the far side", says one student.
>
> "Nope," says the teacher. "There's not enough air movement over here to cause that."
>
> "The metal sheet is made of two metals, a different one on each side, and the far one absorbs heat faster than the one near the radiator," says another.
>
> "Interesting idea," says the teacher, "but this metal sheet is definitely all made of the same alloy."

> “MAGNETS!!!” cries a third.
>
> The other students and teacher ignore this one.
>
> After the students have exhausted all their theories and the teacher has shot them all down, they give up. “Tell us, sensei, what is going on here?”
>
> “Simple,” said the teacher, resplendent in the glow of the afternoon sun shining through the window, “I turned the sheet around just before any of you came into the classroom.”

Of course, most real-world situations are not quite so intentionally dubious. Instead we have theories about how our cars work, why our friends say what they say, and what our cats are thinking and then they fail to predict or explain our car problems, our relationship troubles, or the inscrutable actions of our feline companions. We are constantly faced with inconsistencies between pattern and detail and are trying to fix them up. So just because the ego is in control, because we are integrating S1 and S2, near and far, yin and yang, we may still find we aren’t flowing.

So if an integrated thinking mode that combines details and patterns is to explain flow, we’re going to need more than integration alone. Going full ego is not enough. If there is anything here it is probably in how the details and patterns are integrated. This, fortunately and unfortunately, is somewhat well studied but poorly understood or applied, and goes by the name of rationality.

Rationality as a procedure pops up in multiple fields: game theory, economics, psychology, sociology, probability theory, and artificial intelligence to name a few. We can broadly think of it as the optimal way of integrating information that satisfies an arbitrary value function. It’s not exactly so-called “cold logic”, though that is a degenerate case, and fully encompasses anything that most directly approaches “winning” for whatever winning means to you. It looks like Bayes’s Theorem in its pure form, and learning to apply it to your thinking is one of the goals of a classical education.

By applying rationality, i.e. optimal information pumping, to the integration of details and patterns, we describe something that sounds a lot like flow. Details come in (evidence), they are weighted and balanced against patterns (priors), and combined to achieve an updated state that can be read to point to a clear next action. Even under uncertainty flow is possible, with the best possible option floating to the top. The next thought or action comes automatically because it is the best currently available way forward in the present integration.

But flow, to what extent we can find ourselves in it, is easily broken, and our model points to the ways in which it breaks. On one hand, details can be overvalued so that patterns are not sufficiently heeded. We get lost in the moment, forget our better judgement, and act without thinking. On the other, patterns can be overvalued so much so that the details don’t change our minds enough. We get in our own way, give in to fear, and overthink. Rationality, and thereby flow, is easily broken by being out of balance between details and patterns, and it’s only through skilled practice that we can keep our minds like water.

So it seems somewhat useful to think of flow as rational detail and pattern integration. But this explanation seems fail to square with the way most folks talk about what flow feels like from the inside. It’s described as being automatic, doing without trying, and acting through nonaction. People say it feels peaceful to be in flow, time seems to fly

by, and the running self-narrative diminishes or stops. This sounds a lot more like a cessation of activity than an optimization of it.

But what is optimization if not a cessation of chaos? Normally the mind feels full of competing systems trying to pull us in different directions. Some people report experiencing their own minds as a conversation between multiple agents, and not just metaphorically but as in they think to themselves as multiple characters in communication. It seems not a coincidence that we describe dynamic systems as quiet and stable when they are working as expected. So it's perhaps not surprising we should say flow feels like things in our minds are stopping because in a sense they are: they stop causing problems and work together.

So there we have a theory of flow that is based on physiological underpinnings that, while still not proven, provide a reasonable explanation of basic processes that we can use to construct simple, seemingly useful models of more complex mental processes. I'm interested in exploring possible weaknesses I've missed in this theory, so comment with your objections and I'll see if they can be addressed.

The Developmental Role of Play

This is a linkpost for https://mapandterritory.org/the-developmental-role-of-play-443549cfed3b

When I was a kid, maybe around 8 or 9, I had a cabinet full of Legos. To the casual observer it was a few messy shelves crammed with plastic blocks, but to me it was a portal into another world. Opening it up I'd see little buildings and people resting in suspended animation, and as I turned my attention to them they came alive. Through them I'd become heroes and villains fighting over secret bases, explorers searching the universe, and old friends hanging out and having fun. Hours would pass without notice, and more than once I had to be forced to bed or I would have played all night.

But somewhere between turning 11 and 12 the portal closed. I'd still open the cabinet and try to get my characters to come alive, but I had to work harder and harder to will them into existence until eventually I couldn't do it anymore. Whatever magic I had possessed was gone and all I had left was anthropomorphic plastic.

I was reminded of this when reading Mike Plotz's thoughts on Bateson and play, and it got me thinking about play from the perspective of Kegan and developmental psychology. It's straightforward to claim that my tweenage self developed from one psychological stage to another and in the process changed my play preferences, but that opens up deeper questions. Why did I lose a form of play? Why couldn't I do something I could previously do? Did I gain anything for my loss? And what does this suggest about the value of the developmental perspective?

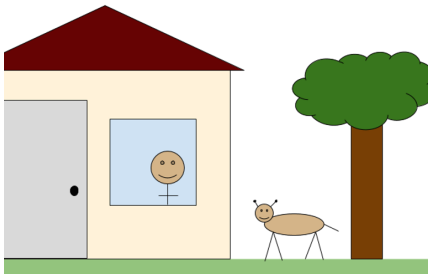
To begin, I'll let Mike present Bateson's view on play and it's role in communication (emphasis mine):

> Play is thus an exploratory behavior, a way of being that goes some way towards a kind of what-if reconnaissance of potentially dangerous territory without taking the extreme risks that the "real thing" would entail. Really fighting to establish dominance might lead to life-threatening injuries, while play-fighting is similar enough to the the real thing to determine who would likely dominate. Compare this function to that of dreams, which also serve as a way for the organism to simulate dangerous scenarios safely.
>
> […]
>
> Returning to the phenomenon of play as described in the previous section, it's now clear that play itself is demarcated in its own frame, which establishes a context in which certain behaviors are interpreted in a special way, e.g. as friendly rather than aggressive. **But frames (e.g. dreams, stories, movies, plays, video games) are often invisible while we're immersed in them—they seem to encompass the whole world.** And so there is a danger of forgetting the context and interpreting those behaviors as real aggression. Bateson sees the development of play as a necessary step on the road to a mature epistemology[.]
>
> […]
>
> The process of psychotherapy, according to Bateson, involves **taking stock of the patient's unexamined habits of thought and behavior, stepping outside of that comfortable frame, and establishing new habits of**

> **thought and behavior**—new rules—in a process analogous to natural play behavior[.]

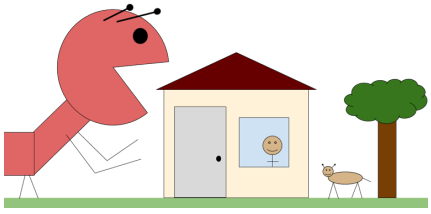
What Bateson is saying, in a developmental psychology context, is that play is a way to engage in development, and his talk of "frames" specifically suggests that play is made in terms of the subject in the subject-object distinction. If you're not familiar, it's related to but more specific than the general philosophical concept of subject and object. Something is said to be object for a person if it's the sort of thing they can model with sufficient complexity to understand it intuitively and have this proven by making accurate predictions about its future states. Things that are subject are not this: they are more complex than one can naturally integrate together the details and patterns of.

It helps to think of objects as things and subjects as contexts, lenses, or frames. Imagine yourself standing in a house looking out the window. Outside the window you see a lawn with a dog and a tree. You can learn a great deal about the dog and the tree, but only through the lens of looking out the window.



An idyllic scene

You can't put the dog and tree together in a bigger context that includes the house because you are inside it. If you want to do that, you'll have to find your way out of the house in order to be able to take in the larger scene.



Just your neighborhood giant ant coming to borrow some sugar

So long as you remain in the house, you are subject to it: you have no way to see the world that doesn't implicitly include the context of looking out from inside the house. Its window frames your experience of the world and the objects in it. This is the way in which the subject subjects you to its way of knowing the world.

Kegan, describing a developmental theory of psychology, looks at how people develop in terms of the categories of things they can hold as subject and object. People are seen to go through phases where their contexts are constructed in terms of objects (in the intuitive, physical sense), object relationships, systems, system relationships, and holons. So if play is, as Bateson argues, a way of moving to larger frames, we should find that play is essential to the developmental process Kegan describes.

Play Focuses on the Subject

As an example, let's consider what the subject or frame of my Lego play was and see if it was part of my developmental process. Assuming Kegan's model, we should be able to identify if my play was about objects, object relationships, systems, system relationships, or holons, and to some extent my play was with objects, but consider how Legos work. You take multiple blocks and put them together to form larger objects. If my play were framed by physical objects (Kegan stage 0), I wouldn't have been able to reliably act out scenarios through the Legos: I'd be focused on understanding how the objects worked in isolation.

Jumping ahead, if my play were framed by systems (Kegan stage 2), my play would have been less about individual characters and their actions and more about their organization. Rather than a good guy versus a bad guy, for example, my play might have been about a team of good guys fighting against the evil they found in the world. Systems seem more complex than what I was engaged with in my play, so this bounds us to consider object relationships as subject (Kegan stage 1).

If my play was framed by object relationships, I'd be primarily focused on understanding how the objects work together. At first this sounds more like a precursor to the play I describe: I was playing with constructed Lego objects and people and not with constructing them. I could build what I wanted so far as I had the parts to do it, and that certainly requires holding object relationships as object in order to be able to put individual objects together.

But consider the nature of my play with the pretend people. My play was specifically with the relationships between them. My favorite games were to pit good guys against bad and to have friends hanging out. My play was about exploring how the characters interacted, and in particular projecting my knowledge of the world onto my Lego people to create pretend scenarios so I could experiment with how the world works. I used my toys like visual aids to help me work out my thoughts.

And this points to why this play became impossible for me as I grew older: I was developing into the next phase, where object relationships become object and systems become subject so that I no longer needed to try to understand object relationships because I could already skillfully manipulate them in my mind. I lost a style of play not because I could no longer do it, but because it was no longer an interesting challenge. My mind could no longer be absorbed in studying object relationships because it knew too well how they worked.

To consider another example, as a teenager and young adult, I played a lot of Civilization. I got hooked playing a demo of Alpha Centauri, then moved on to Civ 3, 4, and 5 as they were released. Although I don't have exact stats, based on data I do have it's safe to say I spent well over 2000 hours in game. I liked playing it because I was able to explore my thoughts on systems and, later, how different systems interacted through the proxy of simplified, pretend nation states. It was fun for me in Kegan stage 2 and Kegan stage 3, but upon developing further it stopped being fun and just felt like an uninteresting way to spend time, the same way my Legos had stopped being fun by the time I was 12.

These days the only game I play is DOTA 2. Unlike my past games, DOTA is an online multiplayer team game where you are randomly matched with similarly skilled players to fight 5 on 5 for about an hour to achieve complex game objectives. Although some players are technically skilled enough to win without teamwork at lower levels, at

higher levels of play the only way to win is through coordinated action. And I find this interesting because it forces me to engage with that which remains subject to me—the integration of all details and patterns into a holon. This is part of my Kegan stage 4 play as I continue to develop towards stage 5.

So at least in my own life it seems play has been key to psychological development. And although I have checked against my life experience in the details to make sure the general case holds there, a quick googling shows that there is broad consensus that play is positively correlated with psychological development. There is a lot of room for disagreement on the mechanism of how play relates to development, but the general principle seems to hold strongly enough that it's well worth supposing the play/development correlation and investigating further to better understand causation.

Appendix: Technical note for those familiar with Kegan

If you are familiar with Kegan, my connection of play styles to stages may seem misaligned with the normal developmental timeline. I've left out the detail above of talking about leading and trailing edges, but the short of it is that a form a play seems to open up to you when you enter a stage and what you are subject to expands, and a play style becomes uninteresting when you leave a stage and your sense of object for a class is complete. So I don't leave stage 1 play until 12 not because of retarded development but because that's when my trailing edge reached stage 2. Similarly it seems it wasn't until I was about 30 that my trailing developmental edge fully entered stage 3 since only then did building individual systems become uninteresting.

Update 2017–05–11

My thinking on these topics has evolved. Although the surface and intermediary details remain correct, I think there are better ways of understanding the relationship between development and play. Although I've not revisited play explicitly, if you found this interesting you might want to follow up with these newer posts.

Unstaging Developmental Psychology

Phenomenological Complexity Classes

Revealed and Stated Identity

This is a linkpost for https://mapandterritory.org/revealed-and-stated-identity-64c6ec070f4f

In modern America, many people think of their identity as a thing they own. If I say I'm a Brony, have a non-binary gender, or identify as transracial, then that's it, and so much the worse for you if you perceive me as otherwise. This is an empowering idea that lets people escape the confines others may try to put on them and seems to make people happier than being forced into identities they don't want. But it comes at the cost of ignoring the reality that you are not in full control of your identity.

Much of what we might call identity is decided by others even if we don't like it. I don't get to decide if you think I'm tall or short: you're going to look at me and categorize my height. I might want you to think of me in a particular way, but you have to make your own determination about me, and all I can do is present information via various channels to try to influence your decision. Eminem gets at this tension when he raps

> And I am, whatever you say I am
> If I wasn't, then why would I say I am?
> In the papers, the news, everyday I am
> Radio won't even play my jam
> 'Cause I am, whatever you say I am
> If I wasn't, then why would I say I am?
> In the papers, the news, everyday I am
> I don't know it's just the way I am

So it seems we have a problem. People want to be able to choose how others think of them, yet wanting a particular identity is not enough to get it. Identity can become confusing and frustrating in cases where, say, I think of myself as kind but others don't see me that way. How do I deal with this conflict? Do I know the essential me better than others? And if so, why don't they see it? And if they see the "real" me, what does that mean about the value of my self-perceptions? Can I be something even if no one thinks I am, and if not does that mean I never really had the right to choose my identity?

Rather than trap ourselves in that quagmire, I think we can borrow a concept from economics to approach a better understanding of the conflicts in understanding identity.

Early in the 20th century, economists noticed that sometimes people said they wanted one thing, like cake, but then bought ice cream instead. Rather than accept that sometimes people do mixed up things, Paul Samuelson developed a theory of revealed preferences to help make sense of demand by looking at what people did rather than trying to model their utility functions (preferences) directly.

A *revealed preference* is observed when you do things, like choosing ice cream over cake when you're at the store, and then analyzing one or more data points to infer a preference. Since it's based on the ground truth of what you did there is relatively little room for dispute in what someone's revealed preference is. This gives us reliable information about your preferences at the cost of ignoring why you made particular choices.

A revealed preference has the natural dual of a *stated preference* that captures the why at the cost of certainty. If you say “I like cake more than ice cream,” then you have a stated preference for cake over ice cream, which is interesting because it tells us how you think about the choice between the two, but provides little to no way to check its accuracy.

Luckily, we don’t have to chose to look at only one of revealed or stated preferences. As duals they provide different insights into the same underlying preferences and give us a way to see into our complex and often opaque decision making process. There is no contradiction when stated and revealed preferences disagree: they are different frames for looking at the same reality. I can say I like cake more but consistently choose ice cream because maybe both my stated and revealed preferences miss the fact that the only cake flavor the store ever has is coconut, which I would rather not eat at all.

So if we can say we’d prefer to do one thing and then do another without contradiction, maybe we can apply the revealed/stated duality to identity to resolve the confusion of being both whatever you say you are and whatever others say you are.

Untangling identity

First, to be clear, when I’m talking about identity here I’m talking about the way in which a person is understood as part of the world. This includes how they relate to themselves and others, how others relate to them, and how they engage in systems. So identity isn’t just about hot-button identity politics issues around race and gender: it includes all the mundane ways in which we are distinguished from and included in the surrounding environment.

Stated identity is straightforward—it is that which we claim about ourselves. If you say to yourself or others that you are nice, then niceness is part of your stated identity regardless of your actions or other’s perceptions of you. In this way stated identity is often aspirational and corresponds to how we see ourselves in far mode and is something we feel in control of.

Revealed identity is more complex, though only as complex as we choose to make it. Like revealed preference it’s found by inferring a pattern on observed behavior, but now the data includes our and others’ perceptions. For example, suppose you witnessed one of your friends literally take candy from a baby. Would you describe your friend as nice or mean? Maybe you think this is wrong in all cases and so your friend is mean, but suppose you knew that this particular baby was allergic to candy: now it seems your friend did a nice thing to protect the baby. What your friend’s actions reveal about their identity is subject to interpretation by yourself and others, and that’s because identity is as much about the actions someone takes as how those actions are understood.

We can apply this stated/revealed identity duality to better understand what’s going on in cases where it previously seemed that identity was confused. Consider again Eminem rapping in The Way I Am.

https://www.youtube.com/watch?v=mQvteoFiMlg

It’s not so much that Eminem’s identity is unclear here as it is that his stated and revealed identities expose different aspects of his identity. People may alternatively

perceive him as a hero, a destabilizing cultural force, or a sellout depending on their experience of him, and this is often at odds with both who Eminem says he is and how he experiences his own identity. This of course does nothing to resolve the conflict per se, but it does make sense of it and allow a more nuanced, complete view that has room for observed disagreement without a need to force competing views out so that they fit a model.

We can similarly use the stated/revealed duality in our own lives to understand apparent identity conflicts. I've been described as eccentric, viz. weird, while I think of myself not as deviating from norms but as making choices with a greater willingness to consider larger solution spaces. Both can be true, though: I can appear weird while never intending to be so. It doesn't change the fact that I present as weird, but it also doesn't mean my weirdness is caused by a desire to be weird or my having some essential property of weirdness. I instead unask the question of "am I weird?" to gain a greater understanding of myself and the world.

And if all this sounds like I've just very cleverly found a way to say you're not wrong to call me weird but also I'm not weird, you're right. Reality as we perceive it is not a single thing, but many contexts for understanding the world. There's some real thing out there, but you can't get at it directly, and acknowledging that by giving these frames names, like stated and revealed, helps us better understand not just external reality but how our understanding of reality is constructed because it is all part of reality. Identity is just one more way in which we are forced to make sense of our complex, opaque universe.

Debate and Dialectic

This is a linkpost for https://mapandterritory.org/debate-and-dialectic-850b3585dad4

Election season is over in the US, but folks are still talking about how divided political conversation is. We hear that people are trapped in filter bubbles that limit their exposure to varying opinions and that America is more divided than ever thanks to modern media enabling our tribalism. And although I suspect we are primarily seeing a return to normal levels of tribalism from the unusual dream time of cosmopolitanism enabled by recent mass migration to cities that ended in the 1970s, there is a deeper source of divide found in the structure of our public conversations that lies beneath the stormy seas of politics.

Specifically, most public political discussion is debate. News programs do this explicitly both when they organize formal debates between candidates and everyday when they ask experts to argue policy. Politicians and experts also do this implicitly when they give solo speeches by speaking in the context of what others have said and arguing for their positions against others. And given how politicians and experts are often criticized for going "off point" when they fail to maximize their opportunities to argue for their positions, it seems there is little room for anything other than debate in political discourse.

A similar pattern exists outside politics. Professors debate contentious academic topics. Theologians contest the fine points of religion. Lawyers argue cases to convince judges and juries. People even debate against themselves by making pro/con lists. All of these are attempts to find the truth, to understand reality as it is, and all do so through adversarial talk.

If this is what you're used to this probably seems normal, but there is a fundamental problem with seeking truth through debate. To understand why, consider what's happening in debate from a game theoretic perspective. Two parties, A and B, are presenting information to a third party, C, who will use the information to develop beliefs. The payout is something like this: C wins if C ends up with true beliefs; A wins if C adopts A's beliefs, and B wins if C adopts B's beliefs. The game is not zero-sum: if A's beliefs are true then C and A can win; likewise B and C can both win. But if we ignore C's payout, the game is zero-sum: A or B or both must lose and at most one can win. This informs the strategies that A and B take.

The best situation if you are A or B is to have true beliefs. Unfortunately this is hard. And although for simplicity here I talk about "true" beliefs, in fact we should be talking more about how closely a belief corresponds to reality and talk about winning more or less depending on the strength of the correspondence. So really what we should expect if we are A or B is to have somewhat true beliefs where we're not even sure how true our own belief's are. That's the kind of situation in which the debate game has to be played.

Even if A and B were aware of how true their own beliefs are, the structure of the game still encourages them to play to win with known false beliefs such that either A or B is the only winner. In such a scenario C is not necessarily even helped by A and B; C must try to win despite A and Bs' attempts to persuade them. This is the problem with debate: it can work, but its structure ensures that it will mostly diverge from the

ideal case and create a game where A and B will actively try to make C lose. This is, in short, a very ineffective way of trying to understand reality if your goal is true beliefs.

The game of debate forces its interlocutors into motivated reasoning, an attempt to find a plausible explanation for conclusions already determined rather than an attempt reason out a conclusion. Motivated reasoning skews our view of reality towards what we wish was true was rather than what is. Surely we can approach collaborative understanding of the mundane world around us without resorting to rationalization.

A debate is etymologically a fight. It's a kind of dialogue or "across speech" that puts the participants against one another. If I were to be so bold as to define a new term, debate is a kind of dislogue or "against speech". What we want instead is dialogue that brings participants together through logical reasoning. What we want is dialectic.

Dialectic is philosophical, rational speech aimed and addressing contradiction. Rather than trying to resolve disagreements, though, it accepts them as a source of developing deeper understanding. Through repeated acceptance of contradiction, or aufhaben, it grows comprehension towards a completeness that incorporates both thesis and antithesis.

That's all a bit abstract, though, so here's an example of a dialectic. While watching it, notice how negation advances understanding rather than contracting it.

At each turn of the narrative, a negation of a previous conclusion arises that leads us to broader understanding. This is nearly the opposite of what happens in debate, where negation is used to drive back competing explanations to create space that can be filled with the speaker's preferred reasoning and conclusions. If debate is a fight, then dialectic is a dance where contradictions partner in graceful harmony.

But, I anticipate you protesting, isn't this just debating politely? Perhaps the outcome is similar, but the key difference between debate and dialectic is that dialectic is not a game in the game theoretic sense. Even if there are multiple participants in a dialectic, there is no one who can win. The goal is truth, so well as it can be understood, without regard for who helps find it. Dialectic makes no space for individual credit, so it removes the payout matrix and replaces it with a single payoff shared by all.

This of course means that we must still debate.

For although dialectic offers a better way to reconcile differences in approaching the truth, nearly everyone is incentivized to (meta)defect and turn dialectic into debate. Unless you are like Hegel and ask to be judged by your own methods, you will have greater prestige if more people adopt beliefs like yours because you argued for them. So if there is nothing to guard against it, there is incentive to defect and treat dialectic as a game, and then to defect a little in dialectic to gain a little prestige but still mostly get at the truth. But defection begets defection, the balance is broken, and we soon find ourselves back at the equilibrium of debate.

But if debaters still care about the truth, perhaps because they are both participant in and judge of the debate, there may be ways to work within the game to produce more winning outcomes. One such approach, the "double crux", is how fellow Map and Territory contributor Duncan A Sabien tries to improve debate.

As for me, I ironically find myself soldiering on in a campaign to engage in dialectic despite the forces of reality opposing my choice. In doing so I reveal a preference for truth over prestige, perhaps making dialectic natural for me anyway.

Act into Fear and Abandon all Hope

This is a linkpost for https://mapandterritory.org/act-into-fear-and-abandon-all-hope-81bcc114c5fd#.2723rfhvv

The first time I received truly life-changing advice was when a friend pointed me in the direction of David Allen's Getting Things Done. I was in the middle of getting my graduate degree, struggling to keep up with the demands of school, teaching, friends, and romance, and although I wasn't drowning—that would come a couple years later—I was fighting to tread water. Learning about GTD was like being thrown a life preserver.

The core idea of GTD is simple: write down everything you want to do in a trusted system, keep the info in the system honest and up-to-date, then look at the system to figure out what to do. It doesn't matter if your system consists of scraps of paper, emails, a spreadsheet, or fancy to do software: as long as you trust it to include everything and be available, it acts as a sort of extended memory that you can offload your worries to. This let's you get things done because the system is more reliable than your memory alone and it frees your mind to focus on the here and now rather than everything else you could be doing.

I'm happy to say that GTD changed my life! I went from forgetting to do things to remembering everything and from spending hours worrying about all the things I wasn't doing to spending hours in flow. My grades improved, teaching was less stressful, I had more time for friends, and I managed to do one or two romantic things. It gave me so much more capacity for doing stuff that it even created time for me to spend telling other people about how great GTD was. Everywhere I looked I saw problems in people's lives that could be solved if they would just read Allen's short book. My life was made better, and I wanted share my new-found wisdom with anyone who would listen.

But I often found myself in the position of a zealot preaching salvation upon deaf ears. Most people weren't that interested in my advice to try GTD, and some people even found it offensive. "What do you mean I should use a system instead of my memory? That doesn't feel natural—it takes all the soul out of my actions!" I myself even eventually turned against GTD, finding a strong need for less structure. The great advice that had helped me so much just didn't seem to work for other people, including my future self.

I've seen this happen to lots of other folks, too, where they discover an idea that changes their life and then have difficulty sharing it with others. Instead of increased productivity, they might get advice that helps them lose weight or dress better, or they might pick up a whole bundle of ideas like a religion or philosophy that helps them experience greater satisfaction with their life. In all cases it feels from the inside like finding the secret to winning at life. But then something happens when you try to share the secret with others, and most people ultimately reject it. And after you see this pattern enough times you start to notice that advice is awfully hard to give.

From an economic perspective this must be so. For one, if advice were easy to give there would be no business in it, yet we find entire fields of professionals—therapists, doctors, pastors, gurus—who make a living in part or in whole because people pay them to be expert personal advice givers. Further, if it were easy to give advice we'd

expect there to be very little of it left to give that could help much: we'd expect everyone to have already taken up all the advice that generates most of the value leaving us with diminishing marginal returns on additional advice. But since we don't live in such a counterfactual universe where few give advice and most advice is heeded, giving advice must be hard.

Despite this, I'm going to try to share some advice anyway. I have no reason to suspect I'm more likely than most to have greater than average success at conveying advice. Nor am I even sure this is the advice you need, because as often as not someone needs exactly the opposite of the advice you're giving. But I'm pretty sure that this advice may be tremendously useful to some, and so on the occasion of a new year when people are traditionally more open to hearing advice and trying new things to achieve their goals, I'll take the risk that I might help someone.

It's in this spirit that I advise you, act into fear.

The idea is simple. If you are afraid of something, do it. Afraid of talking on the phone? Talk on the phone. Afraid of telling your crush of your unrequited affections? Tell them. Afraid of not standing up for yourself? Let someone dominate you. And if no fears immediately come to mind, go looking for them so you can act into them. In doing so you can free yourself from fear and gain the capacity to more often make well-considered decisions.

It's not immediately obvious that you'd want to overcome fear, though. It alerts us to dangerous situations and causes us to avoid such situations or engage in them with hyperarousal. Fear exists because it kept our ancestors alive, and given we see evidence of fear in almost all animals, it likely originated hundreds of millions of years ago, so we're probably already operating near the evolutionarily optimal amount of fear. Except that humans have undergone recent, rapid increase in brain size which, based on comparisons of humans to other animals, gave us greater ability to think about patterns abstracted away from details. This gives us an advantage in long-term planning and ratiocination in general, so it seems likely we could use that to achieve more preferred results than our ancient fear responses produce, especially since survival in the modern world depends less on making quick decisions and more on impulse control. But, that's only possible if we can think and act based less on fear and more on reasoning.

To lessen fear, we must understand it. In the broadest sense, fear is an emotional response to the possibility of violated preferences. At the extreme, fear neuroses and phobias tend to be focused on a strong preference to not experience particular things, like spiders, open spaces, or heights. More commonly, fears tend to be about those things we'd rather not face or will only face from a fighting stance, like confrontation, failure, and hardship. That we talk of "facing" fears means we often perceive them as external to us and that we can make ourselves safe from them if we stay away from, defeat, or otherwise eliminate them. But as anyone who has tried to contemplate their own death knows, fear can also come from within, and so fear must encompass many feelings born of not getting what one wants.

Defining fear so broadly, we must accept it to include and be tied to what we often consider separate emotions, such as worry (fear of future events), anger (a possible response to fear), and depression (systemic sadness from experiencing feared things). We can call all of this suffering in the Buddhist sense, where fear points to any violation of your expectations of how you would like the world to be and a failure to

accept the world as it is. This is the broad sense of fear I mean when I advise you to act into it.

And with fear so broadly understood, it seems changing your behavior to act into fear could include nearly any intervention that remotely touches on fear. If we consider religious and philosophical interventions—asking yourself what Jesus would do, following the example of Confucius, aspiring to live by the Sharia—this seems to be the case. They are largely about increased self-control to approach an ideal, and as such must ultimately address fears that hinder acting on ideals. But some interventions do more to directly address fear than others, and psychology has made a science of directly intervening to overcoming fear.

Psychotherapy, which seeks to correct maladaptive thinking, has long recommended facing fears as a cure. The most recent form of this, cognitive behavioral exposure therapy, has been successful applied in treating phobias, post-traumatic stress disorder, and some cases of depression and anxiety. This is a direct application of the advice to act into fear, and it has given some people their lives back. But this only addresses pathological fear. What about addressing "healthy" levels of fear?

Positive psychology, which focuses on maximizing preference satisfaction, has also found value in acting into fear. We find this throughout the self-help literature, quintessentially expressed in Susan Jeffer's *Feel the Fear and Do It Anyway*, a book laying out a way of doing exactly what it says in the title. Similar sentiments can be found in Habit 1 of Stephen Covey's *The 7 Habits of Highly Effective People*, Nathaniel Branden's *The Six Pillars of Self-Esteem*, and even in the foundational *How to Win Friends and Influence People*. This pattern continues with modern applications of psychological research operationalized for management training and organizational development, exemplified by Kegan and Lahey's immunity maps that lever on worries, which is to say fears.

And thus, as is so often the case in my thinking, it comes back to Kegan. Exposure therapy showed me that debilitating fears could be assuaged, positive psychology pointed me towards overcoming fear in broader contexts, but ultimately Kegan and developmental psychology showed me why we should ultimately expect this to work and be desirable.

The fundamental animating process of developmental psychology is resolution of confusion by creating more complex modes of understanding. To put it concretely, babies turn into children turn into adults by first not understanding the world and then exploring it until they understand it so well that it gives them new frames for thinking. When we're young this process mostly happens without conscious effort because the world demands we develop a more complete understanding of it in order to get the things we need to survive and thrive, but as we reach adulthood it's increasingly possible to get by without understanding more. By making clever choices in the company we keep and the cultures we engage, as adults we can insulate ourselves from the fullness of the world, and by doing so cut ourselves off from the need for further development.

Key to creating such a buffer against further psychological development is fear. Fear is like a fence, keeping a person "safe" by separating them from the things that would challenge their understanding of the world. Fear keeps out new info that would invalidate existing beliefs and creates a bubble inside which only confirming evidence can be found. Fear protects us from cognitive dissonance, but in so doing cuts off the vital confusion that helps us grow in our complexity to make sense of the world.

And making sense of confusion is not just the purview of psychology and science. Philosophy has long observed that fear inhibits growth by protecting oneself from the world. After all, developmental psychology came out of that very observation as expressed in Continental and Analytic philosophy. And as mentioned, in Buddhist thinking fear is key to understanding suffering and ultimately finding relief from it. So let's conclude by seeing what we can learn about fear from Daoist philosophy.

In Chapter 13 of the *Daodejing*, Laozi says this on fear:

> What does it mean that hope is as hollow as fear?
> Hope and fear are both phantoms
> that arise from thinking of the self.
> When we don't see the self as self,
> what do we have to fear?

What does it mean that fear is a phantom that arises from thinking of the self? I interpret this as saying that fear is experienced only by virtue of seeing the self as something apart from the holon of the world. The *Daodejing* then recommends that fear can be eliminated via extinction of the self as something apart from the world by understanding it as both integrated with the whole and having differentiating internal structure. Such a complete stance evaporates fear by giving it no means of replenishing itself. It does the same to hope.

As Laozi says, hope is as much a phantom as fear. Properly stated, hope is the dual of fear—the feeling felt when you have a preference for things to be a particular way. And being its dual, all that is true of fear is also true of hope, so the dual of the advice "act into fear" is "abandon all hope". This is the insight that Western thinking often misses, that psychological development requires not just succeeding in the face of failure but also failing in the face of success. To quote Chapter 13 of the *Daodejing* again:

> What does it mean that success is as dangerous as failure?
> Whether you go up the ladder or down it,
> you position is shaky.
> When you stand with your two feet on the ground,
> you will always keep your balance.

So in this new year and all days to come, I encourage you to keep your balance by acting into fear and abandoning all hope, for as Laozi concludes Chapter 13:

> See the world as your self.
> Have faith in the way things are.
> Love the world as your self;
> then you can care for all things.

Nothing is Forbidden, but Some Things are Good

This is a linkpost for https://mapandterritory.org/nothing-is-forbidden-but-some-things-are-good-b57f2aa84f1b#.t2m31g3qi

I recently wrote about my vegetarianism to casually document it for friends, and it proved to be more popular than anything else I've written lately. My guess is that it was popular because it was about something highly relatable that many people think about already—food choices—but maybe it was engaging because it was ultimately about morality. Since I don't have much more to say about food, I'll say some more things on morality to see if it proves similarly engaging to folks.

On my moral thinking I said this:

> As you may have noticed, I became a vegetarian via preference utilitarianism, but stay a vegetarian to signal virtue. That would be pretty confused moral reasoning, except properly I don't think morality is a category of thing that exists in the world and instead is an illusion created by seeing the world through a frame that does not include system relationships. But I do recognize preferences, my own preferences include a preference for the maximization of the preferences of others all else equal, and as a result I think in a way that generally aligns with the moral theory of preference utilitarianism, but if I had different preferences about the preferences of others I could just as easily be a deontologist or virtue theorist in terms of morality, so I see no problem in the contradictions that result from flatting my thinking into terms of morality.

Echoing this sentiment, a few days later this pithy line showed up in my Facebook feed:

The rules say we have to use consequentialism, but good people are deontologists, and virtue ethics is what actually works.

Depending on your thinking on morality it may sound to you like I'm being evasive and Eli is being too clever, but I think both of us are trying to convey that something more complex is going on that ends up doing weird things if you squeeze it into the abstraction of morality. Eli wrote extensively on this topic back in 2008, but you may not like his writing, be unwilling to sift through it all, or simply find it unconvincing, so I'll try to cover the core of the issue here in my own way.

Morality concerns systems for reasoning about what is good and bad. It has two primary components we can examine: the systems themselves and the value judgements of right and wrong these systems operate on. Although some moral theories tie these two aspects tightly together, as in natural law theory and contractualism, we still need a way to choose, and thus judge, a moral system to use, so some part of value judgement must reside outside the system if we want the system to be complete so we can use it to justify the use of the system. But by doing this we introduce a free variable that can contradict anything else in the system since it is not bound by its logic, thus forcing a choice between consistency and completeness, and since morality specifically concerns "systems" it must choose consistency over completeness in the same way mathematics must choose to be consistent instead of complete.

This is no more fatal to morality than it is to mathematics, but it does mean we're going to see some weird stuff where we have to leave part of reality out to get a consistent remainder. Just as there are numbers we can't count and questions we can't answer with mathematics, we find repugnant conclusions—that we should maximize the creation of minimally good lives—and trolley problems—that you may act to harm less but still cause harm—in morality. These scenarios show us that even if moral theories usually align with our value judgements they don't always do so, thus we may benefit from having a way of thinking about right and wrong that aims for completeness instead of consistency.

Rather than beginning in logic and reasoning, as morality does, let's begin in value judgement. Value judgements are preferences, which is to say liking more and being more likely to take some actions than others, all else equal. Preferences are complete, in that there is a relationship between any two things such that we can say one is preferred to the other, but we may end up with inconsistent preferences, such as preferring apples to bananas and bananas to cucumbers but preferring cucumbers to apples. This feature of preferences is annoying if you want to study rational actors and is often worked around as needed, but is great if you want to understand how decisions are made by real people.

Thinking about value judgements as preferences, you notice that value judgements in particular and preferences in general are not only about one's own behaviors, but include value judgements and preferences on the behavior of others. These exist even in the absence of a universalizing moral system, since even the most libertarian of people must have a preference for not having specific preferences about what others do. And since most people make some significant value judgements about the behavior of others, we quickly find ourselves in an interacting network of value judgements that collectively proscribe certain behaviors. That is, from value judgements alone we get a collection of heuristics and rules about what is good and bad, even if those judgements are not made systematically, and those heuristics and rules are shared among a community. Instead of a moral system we might call this a moral standard since it's inconsistent, hence not a system, but does provide standard answers to as many moral questions as possible.

One rule people generally include in moral standards is integrity, i.e. a kind of self-consistency in moral choices. And the best way to have integrity is to have a systematic way of reasoning about the value judgements of your actions, thus creating a desire for the existence of a moral system. Unfortunately, our moral standard can by design never be fully consistent, so we're stuck with the dual of our original problem. We can neither find a satisfying answer to all moral questions using a moral system nor can we achieve integrity while satisfying all our value judgements.

This is what I mean when I say our thinking on what is good and bad becomes confused when squeezed into moral theory. The moment you manage to cover all situations you introduce contradictions and when you eliminate all the contradictions you leave out satisfying all value judgements. Morally, we are forever between a rock and a hard place.

Of course, this all makes a lot more sense if we think of morality as a kind of illusion or mirage we create by trying to impose our understanding of good and bad onto the world. If morality exists not in the territory, but in our maps, and if we rely on our maps so much that our thoughts become a part of our perceptions, then morality will seem to exist “out there” when it’s mostly “in here”. We will mistake the subjective and intersubjective for the objective, and find we can never quite pin down reality under the thumb of morals.

So what are we to do if morality cannot tell us always how to act, we cannot satisfy all our preferences with integrity, and morality exists in the space between reality and our understanding of it? Should we give up, accept that nothing is true and everything is permitted? Probably not, since most people wouldn’t like to live in such a world, and we’ve shown a collective revealed preference for combining integrity and values as best we can throughout history. Instead, we must find ways of addressing moral nebulosity, and thankfully we know that we can, because we have in fact been doing it all along, only we didn’t realize it.

The De of My Decisions

I’ll leave it to you, dear reader, to figure out how you want to live in our messy, complex world. But, so that I don’t leave you with the null advice of “do what’s right for you”, I’ll at least give you a picture of how I’ve chosen to deal with the nebulosity of good and evil.

As preface, I’m something of a native virtue ethicist. Throughout my life I’ve had some vague sense of virtue and acted to fulfill it on the belief that it would lead me to living the sort of life I’d like to live. This is probably due to my upbringing, where my parents simply taught us to be good people in the absence of any strong religion or formal moral philosophy. If put in terms of Albion’s Seed, I grew up in the broader Quaker culture of America, even though none of my ancestors were specifically Quakers (that I know of).

This is probably why other moral stances didn’t appeal to me. I grew up in Florida, surrounded by a mix of Cavalier and imported Puritan culture. Most people I knew during my school years took a deontological view informed by their religion. To me this way of thinking felt confining and unhelpful, even as it seemed to work for them, and I rejected being yoked by specific moral rules.

As an adult, between university and a whole host of new friends I connected with first online and then in person, I met a lot of consequentialists. My guess at the cultural origin of this, if I had to give one, lies in a combination of Jewish religious scholarship and adopted Enlightenment rationalism. The consequentialist approach felt better than deontology, but it also seemed too complicated to work out robust solutions to moral quandaries. I was never sure I had reasoned far enough since more than once I had to change my mind after realizing I had failed to make a full account of a moral question, and while rationalism and consequentialism reasonably demand that you be able to change your mind, I wanted less fragile answers to moral questions where a wrong determination was less likely to result in accidentally causing major harm. Cf. the historical examples of scientific racism, The Great Leap Forward, and most of the history of psychopathology.

So I stuck to my virtue, ill defined as it was, because it was both flexible and robust enough to work for me. I still wished I had a firmer grasp on what exact behaviors

were virtuous, but having gotten by my whole life with only a rough idea of what it meant to be a good person I was at no risk of not being able to figure out what to do.

Then about a year ago, after a failed attempt to write a version of my key life advice into a book, I fell into nihilism. Not so much because I failed at the book project, but because in trying to write it I was forced to confront the fundamental lack of meaning that exists external to us in the universe. This was not really a crisis: I knew I was going to come out the other side eventually because developmental psychology gave me a roadmap for where I was going, but I wasn't sure how I was going to get there. As it happens, it was around this time I finally got around to reading the Daoist classics.

You might say Daoist philosophy has two core parts: dao, or way, and de, or virtue. Dao tends to get a lot of focus and for good reason (besides being literally the name of the philosophy): it's about a way of perceiving reality as a holon and mirrors much of my own thinking on understanding the world. In comparison de seems boring. It translates to virtue in English and also has a parallel etymology. There at first seems no new insights in de, and it can be read as a collection of stuff Daoists just so happen to think is good. But upon deeper exploration, de turns out to be as complex as dao, and extremely useful if you want to live in the world instead of just contemplate it.

Our first clue to this is that there is no list of virtues found in either Laozi or Zhaungzi. Instead we get a lot of examples and templates of virtuous behavior from which we are asked to discover for ourselves the organizing principles of wuwei, usually translated as non-action, and ziran, which translates as both naturalness and spontaneity. These may seem at odds, but wuwei should not be taken to literally mean doing nothing, but instead to mean doing nothing which is not ziran. And ziran is not about giving in to impulse, but instead about acting in accord with dao, which is to say acting in a way that goes with the world instead of against it. It's not so different, then, from trying to do the best thing, but from a perspective that more treats the world as so complex that discretion is the better part of valor and no action is often better than unwise action.

These ideas of wuwei and ziran, being imbued with centuries of use by folks who lived the sorts of lives I would like to live, gave me a scaffolding to build my way out of nihilism. They helped me internalize what I already knew and have argued above—that morality is not part of the territory but a particular map for understanding individual and collective value judgements. They were concepts I needed to continue constructing a complete stance for myself. And they allowed me to understand that even if no behavior is forbidden by some hidden moral structure in the universe, there are still ways to figure out what is good.

So morality may be a mirage, but it's a useful mirage that helps us find life-giving meaning in what would otherwise be a desert of pure perception. I found de to be a helpful bridge towards holonic integration, but you might prefer Sharia law, act utilitarianism, or any number of moral or ethical ideas. Whatever your choice, in this way morality serves as an oasis that will sustain you on your journey to find meaning, especially when all meaning seems lost to the harsh winds of an uncaring world.

Phenomenological Complexity Classes

This is a linkpost for https://mapandterritory.org/phenomenological-complexity-classes-8b41836437b9#.253ac2tbh

This is an introduction to phenomenological complexity classes. The formulation is my own, but is heavily informed by existing results in developmental psychology and complexity theory and parallels phenomenological complexity theory enough that I had to reinvent many aspects of it from the fragments of phenomenology I knew before I started on this project. As such I am working in the field of phenomenological complexity theory ex post facto, so I'll first explain the history of the idea as I discovered it before formally introducing it and giving an accounting of known phenomenological complexity classes.

The Story of Phenomenological Complexity Classes

For the last few years I've been wrestling with an idea. An idea whose threads I found woven through the fields of developmental psychology, phenomenology, epistemology, organizational development, economics, history, information theory, theory of computation, literature, and self-help. I've struggled to understand it, and of late I've focused my philosophical work on sewing together its many patches of insight into whole cloth. My stitching is still a bit rough, but after long hours of needling, I'm satisfied that my intellectual tapestry can tell its story well enough to be understood.

The seed of the idea formed in me when I was 18. Throughout my junior and senior years of high school I was on something of a quest to find a "true" and "authentic" way to live my life. I tried on libertarianism, Objectivism, classical rationality, pop Stoicism, and more. I constructed for myself a complete worldview out of each, only to inevitably tear each one down when, after a few months, I ran into evidence and arguments that knocked over my reasoning as quickly as a breeze collapses a house of cards. It was during one of these demolition phases, before new building had begun, that I had an epiphany.

I was reading, of all things, Eliezer Yudkowsky's Creating Friendly AI. I seriously doubt that CFAI has any special power to do this, but in the midst of reading Chapter 2, my relationship to reality changed. I realized that all my creating and destroying of worldviews had been attempts, not to understand reality, but to try to make reality as I wanted it. When I looked at libertarianism and the rest it hadn't been that I thought they offered a way to live the kind of life I wanted to live, but instead that I thought that through them I could organize the universe to make it behave the way I wanted it to. But the universe keeps its course no matter how we think of it, and coming face-to-face with my mistake, for the first time I chose to engage with reality rather than hide in my perception of it.

My reason for telling this story isn't to share this particular insight, which I hope is so obvious as to seem banal. What's important is that this was the first time I experienced my mind shifting from one way of thinking to another. Something in the character of my thought changed that day, and though it took years to solidify and really stick, it set me down the path to discovering that understanding can change not just in amount but in kind.

But one event does not a pattern make, so I needed a second hard striking with the clue bat to see the outline of the big idea. I had to wait 11 years, and in that time I earned a master's degree, got married, had a nervous breakdown, took care of my sick wife, got a "real" job, watched my wife lose her job, went broke, defended a dissertation, and then failed out of a math Ph.D. anyway. It was during that ill-fated doctoral career that I learned a lot of decision and game theory, and the more I understood about Newcomb problems and Schelling points, the more decision theory seeped into my soul. Eventually I could see nearly everything through the lens of decision theory, down to the level of mundane activities like negotiating traffic on my daily commute.

So it's perhaps unsurprising that one day I was in the shower, mulling over my general unhappiness with life at the time, when it suddenly struck me that I could use my knowledge of decision theory to be happier. And what it specifically suggested was that my marriage, which was my primary source of unhappiness, was no longer worth the emotional cost I was paying for it, and I should choose different. So I cried, because in that moment I accepted into my heart that I was the agent of my decisions, and I would have to do things that my attachments and fears were screaming desperately for me not to but which knew I would have to do if I wanted to be happy.

So again my thinking had changed in kind. I went from seeing the world as something external that I could understand to something that I could actively engage with using my understanding. Again that may sound obvious and trivial, but in some real sense I wasn't doing that before, and getting there required a shift in paradigm.

This made me suspicious that something more was going on than simply changing during the course of my life. Sure, I was maturing, but the nature of that maturation was more complex than I had been lead to believe it would be. It was not simply that more experience gradually lead to more wisdom, but instead that there seemed to be definite milestones a person could pass that opened new ways of seeing that were previously closed to them. I could see in my past self and in those around me that some people had access to more complete world views than others, and those whose thoughts encompassed more were better able to live the lives they wanted to live. To put it another way, it was not just my thinking that changed, but my way of thinking—my meta-thoughts—that expanded to let me think things I could never have thought before.

I again had to wait to make more progress on seeing the idea. It took almost two years from the date of my shower epiphany for me to work up the courage to tell my wife I wanted a divorce, and that only happened because I found myself backed into a corner with little choice but to confess. But cowardice or no, I made my choice, and this, along with changing job prospects, lead me to move to the San Francisco Bay Area in Fall of 2013, where our story continues.

It was August, 2014. I was at a house party in Berkeley, where our idea of a party is one part carousing, one part philosophizing, and one part naked hot tubbing. Between drinks and whirl jets, I stayed up till sunrise with Malcolm Ocean and Ethan Dickinson as they told me about the work of Robert Kegan. I was riveted. Kegan, along with Lisa Lahey, had built a framework for approaching human psychology that encompassed and explained the psychological development I had noticed in myself first when reading CFAI and later when I decided to end my marriage. Their ideas gave me a structure in which to understand why I had experienced these changes, why they

were so profound for me, and also why I had such difficulty explaining what had changed in me to others.

This was the last piece of the puzzle I needed, but I had to spend a couple years working through the details so that I could deeply understand it. I read Kegan's *The Evolving Self* and *In Over our Heads*, and with each chapter I checked the ideas against my experiences and the reported experiences of others, referenced the cited books and papers, and worked out the implications of the ideas into new areas. I revisited my conceptions of human behavior to see if and how they fit with Kegan's theory, and I tested what I discovered on myself, my friends, and my coworkers to predict, explain, and change behavior. I came out of it a dedicated student of developmental psychology.

I wrote a blog for a while where I explored the lessons of mastering the earlier stages of development and how to transition between stages. I gave up this work when I tried to discuss the later transitions, though, realizing I needed more powerful tools to explain myself. Kegan and Lahey's work on immunity to change was influential, and from there I dove into self-help literature to see how others had approached similar issues in the past. I even tried to write a self-help book and find ways to operationalize my understanding of adult development so that it could help others in a very short lived career as aspiring life coach.

Through this work I distilled my advice for producing the same kind of psychological development in others that I had experienced in myself down to a single phrase: act into fear and abandon all hope. And in the process I had used Daoist philosophy as a scaffold to help me integrate my understanding of Kegan along the same lines as David Chapman has used Buddhism, and as a result generated a "complete" world view similar in tone to the philosophical views of Heidegger and Sartre. My ideas remained nebulous, though, until Robin Hanson asked for people to trade engagement in neglected ideas, and Robin made it clear no such trade would be possible so long as my neglected idea—what I now call phenomenological complexity classes—was both unnamed and explained to others only by reading and thinking about the same material as I had read and thought about.

To my surprise, when I looked for a concise description, I now had one! Thus, this.

Be forewarned, though: the following takes a semi-academic approach as necessitated by the subject matter and is decidedly not light reading. The links are often citations and sometimes point to huge concept masses you may need familiarity with to appreciate the examples. I've done my best to keep the philosophical arguments self-contained, but even there the background is considerably larger than I have space to cover. A full presentation of phenomenological complexity classes would likely require a book, so I offer you my thanks in advance if you're willing to struggle through to the end. I think it's worth it.

Introduction to Phenomenological Complexity Classes

Phenomenology is an approach to philosophy that primarily concerns itself with experience. It in particular assumes experience is intentional, which is a somewhat confusing way of saying there's something, a subject, that has experiences of things, called objects, and there's no way to talk about experience that is not in the context of a subject experiencing an object. The approaches to phenomenology are varied, but here we will take the existentialist view that existence is prior to experience and experience is prior to essence, by which we mean that subjects and objects really

exist, that by existing subjects are able to experience objects through causality, and that through experience ontology (the structure of the universe) can be known.

When a subject makes itself the object of experience, viz. the subject experiences itself, the experience is termed consciousness. This is an intentionally broad way of thinking about experience and consciousness so that we can talk about the "experiences" of rocks and the "consciousnesses" of trees. The terminology sounds strange because phenomenology was discovered by thinking about the human experience of self before being generalized, but it is perhaps no stranger than when physicists use "observe" to talk about causality transferring information, and in fact a phenomenological approach to physics would view "experience" and "observation" as one and the same. Luckily, engineers have given us a less anthropomorphic term for consciousness—feedback—and the recognition that consciousness/feedback is widespread lead in the mid-20th century to the establishment of cybernetics as an interdisciplinary field of study over control systems, computation, psychology, politics, economics, and more.

It's from cybernetics that we can begin to ask "if nearly everything is conscious, what makes a person different from a clock?" since existing phenomenology, even the phenomenology of complex systems, does little to address this question directly. Building on the work of earlier cyberneticists like Bateson, Wiener, and von Neumann, Douglas Hofstadter explored the boundary between the experiences of "mechanical" and "living" control systems first in *Gödel, Escher, and Bach* and later in *I am a Strange Loop*. Independently, Kegan wrote *The Evolving Self* to explain how phenomenological experience develops in humans from childhood to middle-age, and he overlaps with Hofstadter in examining the core qualities of what it feels like to be conscious. Unfortunately, even if we mash their works together, we don't get a fully integrated approach to the qualitative differences that appear to exist between "mechanical", "living", child, and adult because, while Hofstadter addresses cybernetic systems generally, he stops with humans as a category, and Kegan's work ignores cybernetic systems that are not humans, giving us no extension of his theory that might apply outside psychology. So we'll have to flesh out the details ourselves, and we can do that by looking at the underlying theme both believe to be the source of qualitative differences in experience: complexity.

Complexity is, roughly speaking, a measure of how difficult it is to do something. In everyday speech we say things like "the Israeli-Palestinian conflict is a complex topic" or "cars are more complex than skateboards" to mean "there are a lot of things to consider with regards to the Israeli-Palestinian conflict" and "it'll take a lot more words to describe how a car works than to describe how a skateboard works". These notions are formalized in computer science and information theory, where complexity refers both to Kolmogorov complexity, defined as the length of the shortest possible description of a string of symbols, and computational complexity, a related notion about the simplest program that can complete some specified task. Extending this concept to other cybernetic systems—phenomenologically conscious subjects—we can similarly talk about phenomenological complexity as the simplest way in which a subject experiences an object.

To begin examining the complexity of the phenomenological {subject, experience, object} tuple, we must have a grasp on what counts as a subject. As it happens, basically everything you can think of counts as a conscious subject in phenomenological terms, from quarks to humans. This is because, from the perspective of intentionality, the only things in physics that might precede experience are vectors in state space and only for so long as you don't have causality so that one

vector can transition to another. As soon as you introduce causality each vector can experience the vectors that came “before” it and they become subjects with their “past selves” as objects that are experienced through the information transferred to them via causality. And since even quarks and other still-developing theories of fundamental stuff are already operating above this level by talking about paths through state space—things—we’re forced to accept that everything in our current models of reality experiences feedback in the cybernetic sense and is hence phenomenologically conscious.

So setting aside possible non-conscious subjects like state space vectors with causality, all subjects are phenomenologically conscious because they feed information about themselves back into themselves. But some do this in more interesting ways than others. For example, rocks are pretty boring. To the extent that we employ the fiction that they are “things” they mostly just react—behavior being the way we can detect experience in objects—to being part of the unfolding state of the universe. But some of that state is internal to the rock and interacts with itself over time so that, if nothing else, the rock maintains a stochastically stable molecular state for millennia. That hardly sounds like the kind of dynamic entities we think of as being the proper subject of cybernetics, but what’s interesting is that from a phenomenological complexity standpoint rocks are just as complex as more exciting systems, like trees.

Trees clearly respond to feedback: they grow by taking in water and sunlight, they stop growing when they reach the limit of their capabilities, and they recover from damage on their own. But rocks can grow by magnetism, nuclear forces, and gravity fusing them with other rocks, they may stop growing when there is no more stuff near by to glom on, and they can “repair” themselves by “preferring” more stable configurations. Sure, the epiphenomena or emergent behaviors of rocks are in an absolute sense less complex than those of trees because they require less information to describe, but they seem to be doing the same kinds of things at different scales.

We might say that rocks and trees are subjects in phenomenological tuples that are members of the same **phenomenological complexity class**, mirroring the terminology used in computer science to talk about groups of problems that can be solved by similarly complex abstract machines. Because rocks and trees—and clocks and many other things besides—are capable of having the same general sorts of experiences, we can imagine an abstract phenomenological subject capable of doing what any subject in this class is capable of, namely having direct experiences of reality. And if behaviorism were a complete accounting of animal psychology, this class would also include cats, dogs, and most importantly humans. But behaviorism is limited specifically because it doesn’t take into account the ways in which animals are not like trees by ignoring the experience of experience.

The experience of experience, or meta-experience, is much closer to what we think of everyday when we use the word “consciousness” than the phenomenological definition I’ve been using. It’s a capacity to experience and react not just directly to things in the world, but to experience and react to a subject’s own experiences of the world. Meta-experience is lacking in things like trees, but found in abundance in humans where it even makes possible the notion of subject and object in phenomenology, because without meta-experience there is no way to know that experience itself exists and thus no way to see the members of the subject, experience, object tuple. Because meta-experience allows this greater complexity of experience, we can see that it places a bound on the phenomenological complexity of “flat” subjects that directly experience reality without meta-experience. Thus we can

think of rocks, trees, and the rest as forming our first phenomenological complexity class, the class of subjects with phenomenologically flat experiences.

We must then naturally ask, if meta-experience effectively bounds the phenomenological complexity of subjects with only flat experience, does something bound the phenomenological complexity of subjects with meta-experience? Unlike the jump from flat to meta-experience, there is no sense in which a subject can have meta-meta-experiences since those are still meta-experiences, so on the surface it seems this is the whole story. But just as deterministic Turing machines turn out to not to define the broadest computational complexity class, so too does meta-experience not define the extent of phenomenological complexity.

The additional complexity arises within meta-experience from the complexity of the meta-experience itself, in particular the complexity of the meta-experience of objects. Most of our knowledge about this comes from psychology because so far on Earth only neurons arranged into brains engage in meta-experience, though there is perhaps a case that some advanced artificial intelligence software achieves rudimentary meta-experience. And the complexity of how subjects meta-experience objects is exactly what Kegan explored in *The Evolving Self*.

But as already mentioned, we can't simply point at Kegan's work and say "quod erat demonstrandum" because it starts but doesn't complete a bridge from humans to generic phenomenological subjects in meta-experiences. The group of meta-experiencing subjects is believed by many experts to include at least mammals and probably most tetrapods, cephalopods, and theoretical future strong AI, and a theory of phenomenological complexity classes needs to be applicable to all subjects. So we need to generalize from Kegan's work on humans to include the full range of non-flat, meta-experiencing subjects.

In *The Evolving Self*, Kegan gives a taxonomy of **5 stages of human psychological development** that can be identified in multiple ways. There is also an often unlisted sixth stage prior to the first, sometimes referred to as stage 0, that corresponds to a human's early development from flat experiencing subject to meta-experiencing subject. Some ways Kegan and others have identified these stages, numbered 1 to 5, include

1. Perceptive experience; impulsive action; minimal self narrative
2. Concrete experience; deliberate action; other authored self narrative
3. Abstract experience; planned action; socially authored self narrative
4. Systematic experience; emergent action; self authored self narrative
5. Dialectical experience; fluid action; self transforming self narrative

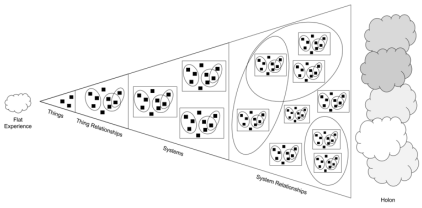
And with all due respect to Harvard University Press, Kegan himself gives a helpful chart summarizing the stages partway through *In Over Our Heads:*

DEVELOPMENTAL STAGE/ ORDER OF MIND (typical ages)	WHAT CAN BE SEEN AS *OBJECT* (the content of one's knowing)	WHAT ONE IS *SUBJECT* TO (the structure of one's knowing)	UNDERLYING STRUCTURE OF MEANING-MAKING
1st Order: Impulsive Mind (~2-6 years old)	One's reflexes	One's impulses, perceptions	Single point
2nd Order: Instrumental Mind (~6 years old through adolescence)	One's impulses, perceptions	One's needs, interests, desires	Categories
3rd Order: Socialized Mind (post adolescence)	One's needs, interests, desires	Interpersonal relationships, mutuality	Across categories
4th Order: Self-Authoring Mind (variable if achieved)	Interpersonal relationships, mutuality	Self-authorship, identity, ideology	Systemic
5th Order: Self-Transforming Mind (typically >~40, if achieved)	Self-authorship, identity, ideology	The dialectic between ideologies	System of systems

The last column is especially relevant because it describes the structure of meaning-making, which is another way of talking about the phenomenological complexity of meta-experience. But Kegan is imprecise about what things like “categories” are and what it could mean to structure meaning making “across categories” because he has a background more founded in classics and continental philosophy than mathematics and analytical philosophy. So to give an accounting of phenomenological complexity classes more suited to my context within the rationalist community and executable philosophy, we need a more detailed approach to the phenomenological complexity at play in each stage in order to generalize them into phenomenological complexity classes.

Phenomenological Complexity Classes of Meta-experiencing Subjects

Taking Kegan’s stages as a rough guide, we can begin to explore the heights of phenomenological complexity within meta-experiencing subjects. Corresponding to each stage, we’ll look at the complexity classes of the phenomenological tuples of meta-experience, in particular the complexity of the object as meta-experienced by the subject to understand the capabilities and limitations of each complexity class. Because I find this to be a tricky subject to think about without a visual aid, we’ll work from left to right through the diagram below.



We’ve already covered flat experience, so moving directly into the meta-experience triangle depicted above, the first level of phenomenological meta-experience to consider corresponds to Kegan’s first stage and is the meta-experience of objects as things. This is in some sense the complexity you get for free from meta-experience because it’s naturally produced by meta-experience through experience of the {subject, experience, object} tuple—or more precisely the {subject, meta-experience, thing} tuple—where “thing” can be understood as the minimally complex way a subject can meta-experience an object. Things in this technical sense are irreducible, isolated, and filled with perceived essence because all ontology must be located

within them by subjects with experiences in this complexity class since there is no where else for ontology to reside but as an inseparable part of the object itself.

Thing-level meta-experience forms a phenomenological complexity class by being the simplest expression of meta-experience while being bounded by the kinds of meta-experiences it doesn't include. In particular, T—our shorthand for the set of thing-level meta-experience phenomenological tuples—can only use flat experience and meta-experience of things to experience the world, which leaves out any way to account for the experiences of other objects. As a simple example, a baby with thing-level meta-experience could experience a person and a ball but not experience a person looking at a ball because the person and ball are isolated and lack a means of experiencing each other from the baby's perspective. To handle more complexity we must replace the category of things with a more complex structure.

In particular, a way is needed to include the phenomenological tuples of other subjects as objects of meta-experience. This is a very fancy way of saying we need things to have relationships, so we need to expand T to TR by allowing the inclusion of {subject, meta-experience, things-in-relationship} tuples. This is depicted in the meta-experience triangle as vertexes encircled by hyperedges, implicitly forming a hypergraph, in the "Thing Relationships" trapezoid.

Missing from **thing-relationship-level meta-experience**, though, is a way of experiencing the hypergraph itself rather than the vertexes and hyperedges within it. This matters because it misses the proverbial forest for the trees and gives the subject no way to meta-experience the whole beyond lumping it together again into an indivisible thing. This re-thing-ing of thing relationships is like the way cartoon trees merge into a unified forest as they recede into the background.

In order to maintain things and thing relationships simultaneously while zooming out we need a new level of complexity made possible by systems. A system is just a collection of things in relationships, but notably has a boundary at which we can talk about the system similarly to how we talk about things but without forgetting that the system is made up of things. Returning to the hypergraph, we can't talk about the graphically invariant properties of a hypergraph, like connectivity, by only referencing the vertexes and hyperedges. We need to have a way of talking about what vertexes and hyperedges are in a hypergraph for the notion to make any sense. Thus a system is like a thing in that it is isolated but different in that it is reducible, creating opportunities for more complex ontology where essence may reside in the system rather than in the things in it.

This need is what gives rise to **system-level meta-experience** that expands the complexity class TR to S by including phenomenological tuples of the form {subject, meta-experience, things-in-relationships-in-system} or more simply {subject, meta-

experience, system}. System-level meta-experience is what powers Kegan's third stage and and offers complexity rich enough to capture many of the experiences of adults. In fact, from a cultural-historical perspective, most people who lived prior to the 16th century likely had experiences exclusively within S because the cultural expression of S is often referred to as the traditional, communal, or pre-modern worldview. This is especially interesting because the pre-modern worldview shows us the limits of system-level meta-experience because it is missing a key capacity that enables modern, systematic worldviews: relationships between systems.

Pre-modern thinking has no space for Other, and when a concept of Other is first acknowledged by a culture it is often as isolated systems with little to no interaction. As a culture develops towards modernism the people within it learn to reify systems into things so that systems can be put in thing relationships, but this gives up reducibility and remains fully expressible via the phenomenological tuples of S. Modernism proper arises when systems offering complete worldviews, usually defined by some "-ism"—scientism, communism, capitalism—compete to embody ontology. Fully engaging this competition requires systems not be reified into things to relate to each other but remain fully deconstructable within their relationships so that they can be compared in whole and in part. If no one system is found to be better than all others in all ways for all people, cultural post-modernism emerges to continue the search for truth by analyzing the relationships between systems.

Addressing culture like this is essential to understanding **system-relationship-level meta-experience** because culture is central to how humans experience system relationships. Outside of culture humans only seem to engage with system relationships in academic fields and confine that engagement to narrow slices of their total experiences. It's only through culture that humans regularly need to meta-experience systems in relationship to one another and only when the culture is cosmopolitan enough to point to gaps in understanding that must be filled by system relationships. Thus culture and its development provides one of the few familiar examples of where the complexity of S is insufficient and must be extended to SR to include phenomenological tuples of the form {subject, meta-experience, things-in-relationships-in-systems-in-relationship} or {subject, meta-experience, systems-in-relationship}.

Given that system relationships are sufficient to experience the complexity of post-modernism, SR seems like it should be sufficient for all phenomenological tuples in which people might find themselves subjects. Yet within post-modernism there are hints that SR does not capture all possible phenomenological complexity since most people find post-modernism unsatisfying and it often causes them to adopt nihilistic world views. This suggests that using system relationships as ultimate ontological vehicles fails to completely resolve questions about the true nature of the universe, so there is likely more complexity needed to fully experience reality and construct a satisfying ontology of it.

We can overcome this failure by the same mechanism we used to go from thing relationships to systems by introducing an additional abstraction for holding systems together in relationships. We call this a holon, which is simply Greek for "whole" and follows Arthur Koestler's use to mean something that is simultaneously a whole and a part. Unlike systems, holons can contain systems in relationships and things in relationships, so offer a way of talking about groupings of systems and their relationships without reifying systems into things. Thus holons are capable of holding repeatedly dividable systems of systems within a single context, enabling what is sometimes called a nebulous or fluid perspective. Here we refer to this as **Holonic**

meta-experience, and it extends SR to H by including {subject, meta-experience, holon} phenomenological tuples.

The realm of holons in poorly explored. Kegan and Lahey's research suggests less than 1% of adults reach stage 5, their equivalent to H, and those that do rarely do so before age 40, so the pool of potential investigators is small. Further, holons offer a way of forming what feels like, for humans, to be a complete world view and thus often gets mixed up in spiritual and religious language that is typically unwelcome in the very modern project of academic science and philosophy. That said, they are exciting for those same reasons and why much of my current thinking focuses on what I've termed holonic integration.

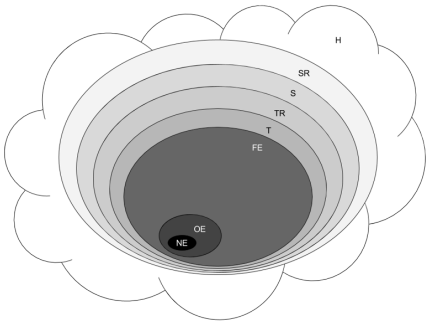
Holons may even capture all the complexity possible within meta-experience. To conclude with a little speculation, I suspect holons are likely enough to permit the inclusion of all possible phenomenological-tuples within our Tegmark Level I universe, aka our light cone or Hubble volume, by admitting only a single holon that covers all things, systems, and their relationships. Additional complexity may only be necessary to address theoretical cross-universe experiences and possibly their counterfactual instantiations in our universe. This is an active area of research for me, and one I hope to be able to address in the future.

Conclusions

Existentialist realist phenomenology permits an explanation of qualitative differences in experiences between subjects via the complexity of phenomenological tuples that subjects can be members of. The resultant complexity classes of phenomenological tuples start from non-experience (state space vectors without causality), outward experience (state space vectors with causality), and flat experience, then expand to include meta-experience and the increasing complexity of subjects' meta-experiences of objects. Each class is a proper subset of the next, and they are in order

- **NE** *(No Experience)*: {}
- **OE** *(Outward Experience)*: NE + {subject, flat experience, object-that-is-not-subject}
- **FE** *(Flat Experience)*: OE + {subject, flat experience, object}
- **T** *(Thing-level meta-experience)*: FE + {subject, meta-experience, thing}
- **TR** *(Thing-Relationship-level meta-experience)*: T + {subject, meta-experience, things-in-relationship}
- **S** *(System-level meta-experience)*: TR + {subject, meta-experience, things-in-relationships-in-system}
- **SR** *(System-Relationship meta-experience)*: S + {subject, meta-experience, things-in-relationships-in-systems-in-relationship}
- **H** *(Holonic meta-experience)*: SR + {subject, meta-experience, holon}

Below, we graphically show the proper subset relationships and acknowledge that it is unclear if H is equal to the set of all possible meta-experiences, even if only within our Tegmark Level I universe.



This formulation of phenomenological complexity classes flirts at times with mathematical formalism, but the theory itself does not have a rigorous mathematical foundation. I am working to correct this, but perhaps unsurprisingly this is hard because it requires building a mathematical foundation of phenomenology. I don't know when or if I will finish this project, but I am encouraged that others are also interested in and even tackling the problem, and yet others may be approaching the same problem from different perspectives.

There are also many applications of phenomenological complexity theory which I expect to explore given time: phenomenological complexity of ems, artificial intelligences, and animals; ways of increasing the phenomenological complexity of experiences subjects find themselves in; and comparative advantages of different phenomenological complexity classes within a subject's teleology. I've already been doing this to some extent on this blog and in personal conversations without a unifying theory to power my thinking; now I have a hard core around which to build my ideas rather than letting them sit on a vast, interconnected web that few were willing to explore.

Thanks

Many people helped me to put this together, not only by supplying intellectual work to build upon, but by offering constructive feedback and additional phenomenological data as I worked out these ideas. The people who helped me, in no particular order, whether they knew it or not, not already named, include Olivia Shaefer, Jacob Czynski, Paul Christiano, Ben Hoffman, Lauren Horne, Tilia Bell, Eric Bruylant, Luke Muehlhauser, Anna Salamon, Duncan A Sabien, Mike Plotz, Sara Lynn Michener, Ryan Singer, David Malki !, ThunderPuff, Scott Alexander, Scott Aaronson, Kevin Simler, Andromeda Cohen, and my cat, Sammie. I apologize for anyone I forgot to mention; I exhausted the list of folks who were salient with no more than 5 minutes of thought.

2017–6–5 Update

I think it's worth saying here that what I am pointing at with the meta-experience classes is a type-based approach to ontological complexity. Michael Commons has done the same, but in an eliminative approach rather than a phenomenological approach. I've now written a little about his ideas.

Unstaging Developmental Psychology

What Value Hermeneutics?

This is a linkpost for https://mapandterritory.org/what-value-hermeneutics-7a05c80d63b3#.qrkwwg5t0

I often take for granted that there is something to be said about how we know things. For example:

- In high school I took a class titled Theory of Knowledge that was a combination intro to philosophy course and applied epistemology workshop.
- My friends are concerned with questions about how to ensure artificial intelligence learns human values and so want to understand how we learn values ourselves.
- I am perhaps now among the ranks of existentialist philosophers of phenomenology.

So it's probably unsurprising that I have spent a lot of time thinking about and trying to improve my ability to know.

One methodology I've found especially helpful has been what I, for a long time, thought of as literary criticism but for interpreting what people said as evidence about what they knew about reality. I first started doing this when reading self-help books. Many books in that genre contain plainly incorrect reasoning based on outdated psychology that has either been disproved or replaced by better models (cf. Jeffers, Branden, Carnegie, and even Covey). Despite this, self-help still helps people. To pick on Jeffers, she goes in hard for daily self-affirmation, but even ignoring concerns with this line of research raised by the replication crisis, evidence suggests it's unlikely to help much toward her instrumental goal of habit formation. Yet she makes this error in the service of giving half of the best advice I know: feel the fear and do it anyway. The thesis that she is wrong because her methods are flawed contradicts the antithesis that she is right because her advice helps people, so the synthesis must lie in some perspective that permits her both to be wrong about the how and right about the what simultaneously.

My approach was to read her and other self-help more from the perspective of the author and the expected world-view of their readers than from my own. This lead me to realize that, lacking better information about how the human mind works but wanting to give reasons for the useful patterns they had found, self-help authors often engage in rationalization to fit current science to their conclusions. This doesn't make their conclusions wrong, but it does hide their true reasoning which is often based more on capta than data and thus phenomenological rather than strictly scientific reasoning. But given that they and we live in an age of scientism, we demand scientific reasons of our thinkers, even if they are poorly founded and later turn out to be wrong, or else reject their conclusions for lack of evidence. Thus the contradiction is sublimated by understanding the fuller context of the writing.

Once I had a way to make sense of modern self-help that is both wrong and right, I turned those same skills to more ancient texts to see if I might find some valuable results despite questionable reasoning. The situation turned out to be better than I expected, for although their metaphysics was tainted with the supernatural, ancient writers more often respected experience enough to rely on it in their reasoning rather than reject it as unscientific. Thus I found compelling wisdom in the Daodejing when I

needed direction out of the nihilism I found myself in after seeing fully through the inadequacy of the modern worldview. And as I worked through the nebulous and terse Laozi, I was unknowingly honing against the original purpose of my as-yet unnamed skill, because what it turns out I was doing was hermeneutics.

I first heard about hermeneutics in a Ted Chiang story that Mike Plotz pointed me towards and then ran into it again when doing research for "Phenomenological Complexity Classes". I probably didn't learn about hermeneutics earlier because it's mainly associated with interpretation of sacred texts and until the 20th century was practically synonymous with Biblical literary analysis. But Heidegger, hoeing the same rows as I am now nearly a century earlier, devised a notion of philosophical hermeneutics as phenomenologically informed epistemology that adapted hermeneutics for the secular world.

Let me explain.

Epistemology precedes metaphysics. Every major philosophical tradition in the world has come to realize that you can't talk about metaphysics directly. Everything we know about the world comes from our perceptions of it, and so instead we can study epistemology, through epistemology construct ontology, and through ontology may try to infer metaphysics, viz. the true nature of reality. That epistemology comes before metaphysics is a descriptive stance—it is trying to *explain* how we actually know anything about the world rather than tell us how we *should* know the world. This descriptiveness holds even if realism is true in its strongest form because it would still be the case that we know about the real world only via observation of it. Thus we can never truly divorce the discussion of what we know from how we know it.

It then follows that if we are going to know anything we must take account of how we know it. Classical epistemology and the epistemology of analytical philosophy are primarily concerned with reasoning, i.e. based on some known truths or facts, what can be logically concluded. This is valuable as far as it goes, but ignores the question of how we get any facts to start with. Indian philosophy has explored ways of knowing in-depth for over two millennia, and Indian ideas entered the Western tradition via 19th century German philosophers. This gave rise to phenomenology, which seeks to study how we know anything, and it finds that the starting point is observation. But if everything we know comes from observation, it must be that all knowledge is subjective in that it is mediated by a subject observing an object. And if all knowledge is subjective, maybe we should give up the project of epistemology all together in favor of solipsism.

But our experiences themselves point towards the existence of an external reality unaffected by our beliefs. Thus, strange as it is to say, solipsism seems to go against our subjective experiences. And so far as we are willing to believe that things exist prior to experiencing them, the existentialist viewpoint, then we can say that knowledge is intersubjective—it exists between multiple subjects experiencing the same object. And to go one step further, to know about subjects observing objects, a subject must experience that observation—an experience of experience—even to observe their own observations. Thus it seems that all knowledge, direct and meta, is obtained via experience; epistemology must have a phenomenological basis; and therefore epistemology is ultimately founded on phenomenological methods even if we are most familiar with science and other epistemological methods that aim to be robust to differences in experience.

Thus all knowing is fundamentally rooted in interpretation of experience or sense-making. This creates a interpretive circle where where full understanding makes holonic demands on the thinker to understand reality as a whole and as many parts in an integrated manner. And since interpretation of experience is what hermeneutics is all about, it seems an appropriate approach for dealing with not just texts but all experiences, both our own and those of others.

This is old hat, though. By the 1960s most Western thinkers, scientists included, were aware of these arguments in favor of hermeneutics. So why did I have to reinvent the concept on my own before I was able to look in the right places to learn about it? My best guess is that starting with Russell, Popper, and Gardner and continuing with Sagan, Dawkins, and Tyson, intellectual discourse has moved heavily towards favoring material realism as the only acceptable philosophical approach for right-minded, modern folks. This seemed especially necessary in the face of the anti-scientific counter-counterculture, and my own education certainly reinforces this idea: I was indoctrinated into the belief that only science and logic are able to determine what is true and to believe anything else would have been to betray my subculture of rational skepticism. That we live in a world where climate science is politics and human biodiversity is racism only reinforces the notion that anti-scientific thinking must be purged, collateral damage be damned.

Given this context, it might be that philosophical hermeneutics has been ignored because it does not neatly support the pro-science narrative. Interpretation of observation is the sort of opening that can be widened to let in any belief you like, so if you favor science it seems better to disallow any interpretation than give the enemy space. But even if we set this cultural context aside, it remains difficult to favor more hermeneutics due to the lack of rigor found in most social sciences and the replication crises sweeping as far afield as physics. That we as a society have difficulty doing trustworthy science, even when strengthened by the existence of opposition, leaves little hope that we can reliably apply philosophical hermeneutics and other phenomenological methods and maintain our epistemic virtue. Yet these structural problems do nothing to diminish the timeless philosophical value to be found in interpreting experience.

I see no easy resolution to this conflict of interest between the completeness of epistemological methods granted by using philosophical hermeneutics and the consistency of excluding broader notions of interpretation to protect us from gullibility. We live in a time when we trust our ability to know so little that we would rather give up useful techniques than be mislead, and while this is perhaps “healthier” than the pre-Enlightenment stance of preferring beliefs that were more convenient than true, it leaves room for the rationalist project and other efforts to curate a garden where stronger epistemology can grow. Perhaps from that garden can be spread the seeds of a more complete approach to knowledge.

What Value Epicycles?

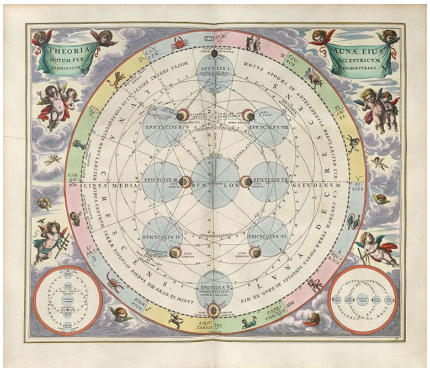
This is a linkpost for https://mapandterritory.org/what-value-epicycles-f8358678a23a

A couple months ago Ben Hoffman wrote an article laying out much of his worldview. I responded to him in the comments that it allowed me to see what I had always found "off" about his writing:

> To my reading, you seem to prefer in sense making explanations that are interesting all else equal, and to my mind this matches a pattern I and many other have been guilty of where we end up preferring what is interesting to what is parsimonious and thus less likely to be as broadly useful in explaining and predicting the world.

After some back in forth, it turns out what I was really trying to say is that Ben seems to prefer adding epicycles to make models more complete while I prefer to avoid them.

Epicycles come from Ptolemaic astronomy which puts the Earth at the center of the universe with everything, including the Sun, orbiting around it. To make this geocentric model fit with observations of retrograde motion, though, required the introduction of epicycles, imagined spheres in orbit around Earth on which the planets rotated. It's now part of the mythology of science that over time extra epicycles had to be added to correct additional observational anomalies until it became so complex that epicycles had to be thrown out in favor of the simpler heliocentric model. And although it seems this story is more misunderstanding than truth, "epicycle" has become the metonym for adding parts to a theory to make it work.



Epicycles of the moon

Epicycles have a bad rap. After all, they proved to be part of an incorrect model and are now associated with anti-scientific adherence to maintaining tradition because they were needed to support the cosmology backed by the Catholic church. But I mean to use "epicycle" here in a neutral rather than pejorative way to simply mean making a theory more complete by adding complexity to it. It's not necessarily bad to "add epicycles" to a model, and in fact doing so has its uses.

Epicycles let you immediately make an existing theory more complete. For example, if you have a unified theory of psyche, adopting something like dual process theory instantly gives you additional explanatory power by allowing the model to decompose

concepts that may have been confused in a unified psyche. In turn, 3-part and 4-part psyche models add more epicycles in the form of more parts or subagents to give yet more complete theories that can explain the existence of contradictory thoughts. That these models lead away from more limited models that neuroscience has given us strong evidence for is beside the point because these models help people explain, predict, and change thinking and behavior.

But asking models to be useful to a purpose doesn't necessarily imply adding epicycles. If our purpose is correspondence with reality then the principle of parsimony may serve better. The better known half of Occam's Razor, the principle of parsimony is to make things as simple as possible but no simpler. I learned this miserly modeling skill from Donald Simmons in his seminal work on evolutionary psychology, *The Evolution of Human Sexuality*, where he warns against the temptation to construct convenient etiological myths that often implicitly and unnecessarily inject complexity into evolutionary processes.

To wit, it would be presumptuous to say that giraffe necks got longer to reach high leaves, since to do so is to suppose evolution is goal-directed. Instead what we can justifiably say is that giraffes with longer necks had a reproductive advantage that led them to leave more descendants than shorter-necked giraffes and thus longer necks became more common. The latter explanation might sound overly cautious to the point of being convoluted, but it differs importantly from the former by not implying additional complexity where none is needed, specifically by not making evolution into a teleological process. The simple evolutionary explanation is enough, though we may wish to further study why long necks were reproductively advantageous and may conclude it's because they allowed giraffes to reach higher leaves.

The advantage of parsimonious explanations lies with probability: the fewer the propositions you multiply together the higher their combined likelihood can be, so a theory with less complexity is more likely to be true, all else equal. This is, incidentally, also the case against epicycles: adding them necessarily decreases the likelihood that a theory is true. Yet epicycles are often useful enough for us to posit their existence, so truth may not always be our highest value. It's in this light that we must consider if simpler explanations are always better.

To the extent you want to be right—to have an accurate map of the territory—you by necessity have an equal want to not be wrong. When you learn something new, as much as you want to change your thinking to integrate the new information, you have an opposite want to find the new information already accounted for so that no update is necessary. You don't want to have metaphorical comets shattering your epicycles all the time: you mostly want them to pass cleanly between the orbits because there is space in your model for them. But if you're lost at sea and need to navigate by the stars, and positing the existence of epicycles is the only thing that allows you to find your way home, then suppose the epicycles exist with all your heart and worry about comets later when you're out of existential danger.

And this perhaps explains the evolution of my thinking. Over time I solve more of the problems that used to vex me, and I suffer less; and as I find greater contentment I can more discard epicycles in favor of parsimony. That I still have a (relatively) complex theory of mind perhaps suggests I am still out at sea, trying to find my way home.

Update 2017–05–15

Val says something similar but with different emphasis.

Unstaging Developmental Psychology

This is a linkpost for https://mapandterritory.org/unstaging-developmental-psychology-fcdc5f9ba894

It's no secret that I think there is some there there when it comes to developmental psychology. However, I must admit theories of psychological development have some weaknesses. One issue we run into right away is that they sound elitist, and for many that's enough to reject them in the same way some people reject evolutionary psychology and human biodiversity as racist. Sure, developmental psychology can be abused to support status grabs and enforce social hierarchies, and I have no doubt some people intentionally try to do this, but this is not itself evidence against the ideas, just cautions against how we use the insights we draw from them.

The other major problem is that developmental psychology isn't a category of strictly scientific theories. One of Kegan's key insights was that developmental psychology is based on phenomenological methods and capta rather than scientific methods and data. This presents a problem when you try to evaluate developmental psychology through the lens of science as Sarah Constantine recently noted:

> Overall, the experimental evidence that distinct, cumulative stages of human development exist is rather weak. The strongest evidence is for Kohlberg's stages, and these (like all the other stages considered) are limited by the fact that they are measures of how people *talk* about moral decision-making, rather than what they decide in practice.
>
> Higher stages correlate with positive results in many cases: people at higher Kohlberg stages are less likely to be criminals or delinquents, positive psychological strengths like self-esteem correlate with the Eriksonian ego strengths, and leadership development measures correlate with Kegan stage. This is evidence that developmental stages do often correspond to real psychological strengths or skills with external validity. We just don't generally have strong reason to believe that they progress in a developmental fashion.

And given that phenomenological methods are not well developed because there are few folks trying to rigorously apply them relative to the way folks apply scientific methods, it's reasonable to reject the notion that we should trust results based on phenomenology. Developmental psychology might be pointing at real phenomena, but phenomenological theories of developmental psychology are going to fail to meet standards of scientific proof. For developmental psychology to become a more widely accepted theory, it needs to be reformulated in scientific terms.

Michael Commons has done exactly this with the model of hierarchical complexity, or MHC.

MHC takes developmental psychology and reverses its normal etiology. Rather than positing stages that give rise to behaviors, MHC considers individual behaviors and then gives a system of classifying their complexity. The classification is hierarchical, so behaviors in higher complexity classes are necessarily constructed from combinations of less complex behaviors, and the classifications match the shape of developmental psychology as discovered by Piaget and Erikson, but the hierarchy is derived independently via a mathematical abstraction.

MHC's formulation allows it to accomplish several things at once:

- Developmental stages can correspond to statistical averages of the complexity of behavior instead of making demands that we interpret thought processes.
- MHC can be tested via scientific methods because it's based on classifying observed behavior rather than classifying self-reported thoughts.
- Multiple theories can attempt to explain the evidence without MHC itself making strong theoretical claims, and MHC remains useful when not used to support theories of developmental psychology.
- It extends to animals we can't collect phenomenological data on because they can't communicate it to us, even if it seems likely they have complex thoughts.

This is much of what I was attempting to do with classes T through H in phenomenological complexity classes, but MHC has the advantages of being more thought out over more years and having more conventional theoretical foundations. Not that MHC necessarily displaces phenomenological complexity classes, Kegan's constructive developmental framework, or others, but it does reduce how much these theories must claim by phenomenological methods alone. MHC seems a powerful bridge between scientific and phenomenological methods of understanding human thoughts and behaviors.

Thanks to Map and Territory reader Dennis Pachernegg for pointing me in the direction of Commons's work

The Personal Growth Cycle

This is a linkpost for https://mapandterritory.org/the-personal-growth-cycle-34ec1c218615

We encounter psychology in our lives mainly through dysfunction. That is, people generally go to see psychologists and psychiatrists because they or others are unhappy with how they are thinking or behaving. This attention to error and correction makes most of applied psychology a kind of negative feedback process aiming to normalize thoughts and behaviors within desired tolerances. People find this psychopathological approach useful as far as it goes, but it leaves out ways of using psychology in positive feedback loops to help people achieve more than normal. Positive psychology focuses on this.

Some folks, however, take a negative view of positive psychology. From what I can surmise, they hold this view because of positive psychology's association with self-help, life coaching, and leadership training—domains where rationalization can often pass for explanation and hucksters can operate as long as they are sufficiently charismatic. And although plenty of folks are working on evidence-based self-help and other scientific personal development methodologies, the guilt-by-association seems sufficient for some people to dismiss the field. Yet I and others have found tremendous value in our lives by looking at such topics as esteem, psychological development, and meaning through the lens of positive psychology. If all that holds some people back from believing in the usefulness of positive psychology is that its sometimes use in cons, then we can advise that they should, as always, practice epistemic hygiene and not confuse correlation for causation.

But even acknowledging the theoretical validity of positive psychology, some folks still object that its techniques don't seem to work, at least not for themselves and not for their friends. To this point, it seems likely that positive psychology only works for the sufficiently privileged because they uniquely have the luxury of being able to spend time and energy on self actualization. Put another way, if you're too busy surviving, you don't have time for positive psychology and even if you did it's not likely to help because it's not your primary constraint. This seems to neatly explain much of why people, after controlling for scams and bullshit, still have varying success with positive psychology.

But I think there is something left unexplained. Among those for whom positive psychology works, some specific techniques will work and others will not. We could just chalk this up to the typical mind fallacy (a special case of the mind projection fallacy), but in doing so I think we ignore the chance to see deep patterns in exchange for the surface-level results. Sure, not every technique may work for every person, but for all persons some techniques should work. Can we find a pattern that would both explain the results of positive psychology and account for cases where specific techniques don't work?

A General Mechanism of Personal Development

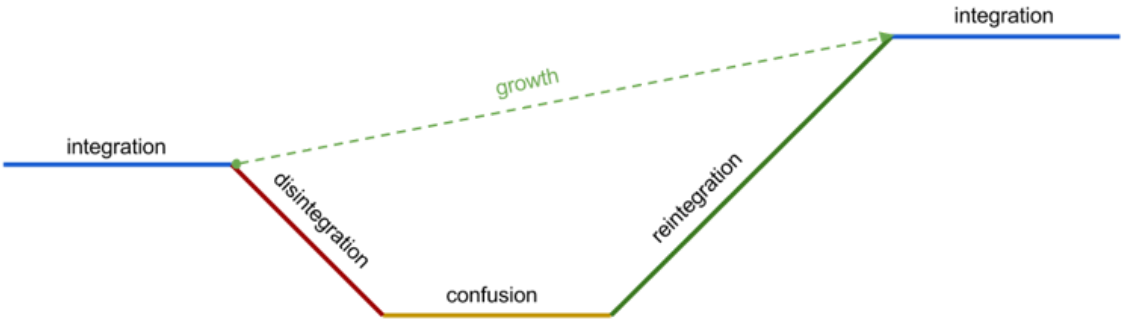
I think we can find such a pattern, and that pattern covers more than just positive psychology—it covers all of personal development and catches positive psychology in its net as a necessary consequence. That's because personal development, as we will construe it here, includes anything that takes a person from less to more of who they want to be. Sometimes this involves psychotherapy and psychiatry to address psychopathologies, but it also involves positive psychology after a person is already descriptively, statistically normal (as opposed to the prescriptive use of "normal" that people mean when they talk about social norms and idealized behaviors).

Because I know post-formal developmental psychology best, let's start looking there for a pattern of personal development. Post-formal developmental models posit psychological development after brains reach physical maturity, which Piaget identified as the formal stage. Erikson, Kegan, and others have investigated post-formal development, and Commons has given an account of developmental psychology that unifies pre- and post-formal stages in a functional, eliminative model. Commons shows that behaviors that characterize these stages must be mastered and applied to perform behaviors in subsequent stages. This means that each stage must, in some way, support development into the next stage.

Commons theory has a strong mathematical bent, but it essentially says that stages are differentiated and hierarchical because each stage builds on the behaviors of the previous one by combining them in nontrivial ways to enable behaviors that were impossible within the previous stage. This matches the traditional, phenomenological view in developmental psychology that stages correspond to increasing ontological complexity without the need to lean on intention. Commons further shows that development requires a period of confusion where existing behaviors may be combined in unproductive ways in an effort to search the space for useful combinations. This is followed by a period of integration where new behavior combinations are practiced, culminating in solidification of new behavior patterns at the next stage.

Noteworthy is that during development from one stage to another a person is likely to make "mistakes" they wouldn't have made were they not trying new behavioral combinations. This relates to Bateson's notion of play and the need for safety explicated by Kegan's conception of holding environment. Safety is a matter of providing physical and psychological space to explore thoughts and behaviors free from fear. Physical security, secure attachment, high esteem, and competence all play roles in creating such a safe environment because, without these many facets of safety, people take on defensive stances to protect themselves from perceived harm. Such stances may be prudent, but they inhibit exploration, learning, and experimentation, and so can retard development. Thus the privilege seemingly needed to successfully apply positive psychology is a manifestation of access to safety, a state that many people unfortunately cannot find themselves in.

So to summarize what is happening, development starts from a place of integration, followed by disintegration into confusion, which through active efforts at reintegration in a safe space results in development. If a safe space for reintegration is not available, development may not proceed. Metaphorically this is similar to muscle growth by weight lifting, where the muscle is initially whole, is torn by the stress of lifting weight, and then rebuilds the damaged, weakened muscle stronger than before when supplied with sufficient nutrients and rest. We can think of this process graphically as a line that dips and then rebounds higher than it started.



This seems certainly a model of personal development assuming growth of the sort examined by developmental psychology, especially since it heavily mirrors Kegan and Lahey's four-column model of psychological development from *Immunity to Change,* is similar to the patterns observed by fellow Medium writers (cf. Jonathan Cottrell's "Equation

for Personal Growth" and Brad Stulberg's "The Growth Equation"), and even looks an awful lot like the dual of one iteration of Boyd's OODA loop. But what of personal development more generally. Does this pattern hold or break?

The General Mechanism of Change

It seems trivial to show that this disintegration-reintegration model does not generically hold in personal development. Consider how we usually construe development of self-esteem. People talk about upward and downward spirals of self-esteem, pointing to a simple positive feedback loop without the one-step-backwards-two-steps-forwards pattern of developmental psychology. There's no obvious disintegration, confusion, or reintegration here, so I guess that's it.

Only, how does the positive feedback loop of self-esteem—or any other feedback loop for that matter—ratchet forward? Feedback requires passing information about a system back into itself and using that information to affect future states of the system. We generically refer to this as "updating", and we know the general pattern of updating well. It goes by many names—entropy, intention, causation, and simply change— but its most rarefied form is perhaps Bayes' Theorem and it is The Pattern behind everything that moves.

To explain, imagine you are in a a fixed, initial state called the prior. New information appears that muddies the waters, so the speak, because the new information does not exactly match the prior. Bayes' Theorem tells us the optimal way to resolve the confusion, and we use this to integrate the new information with the prior to achieve a new fixed state called the posterior. As you may notice, this is a familiar story: we go from integration (prior) to disintegration (transmission of new information) to confusion (knowledge of new information) to reintegration (application of Bayes' Theorem) to a new, more complete state of integration (posterior). In this light it seems the pattern of personal development given above is nothing more than another instantiation of The Pattern, so even if my pattern does not quite describe how all personal development happens, it is only because it is an overly specific version of the universal pattern of change.

This is perhaps disappointing to some, because it seems that the pattern I have seen in personal development is just the same pattern we see everywhere. But this is also exciting because it informs us more of the deep connectedness of the universe. I don't mean that as some kind of hippy-dippy appeal to monism—although I personally interpret The Pattern as identifying with the Dao—but rather as part of the great project of integrating our understanding towards a maximally parsimonious approach. I am encouraged to find that, in a topic as complex as human thought, the Dao reverberates deeply through it.

Angst, Ennui, and Guilt in Effective Altruism

This is a linkpost for https://mapandterritory.org/angst-ennui-and-guilt-in-effective-altruism-73cf62935dfd

There is danger in learning about existential risk. I don't mean that as some kind of bad joke, either: some people experience personal, psychological harm after learning that the world might end in preventable ways. They may become depressed, anxious, or nihilistic over not just their own death but the potential death of everything that means anything to them, and they may wallow in that despair. But, if they can wade their way through such feelings—and many people do!—they might seek to take action against existential risks. One such group of action-takers is effective altruists.

There are many existential risks, and so there are many ways effective altruists try to address them. Some tackle fairly prosaic risks and so can use prosaic methods, like applying activism to address risks from climate change and *not* raising money for emergency response. Others look at more exotic risks from nanotechnology and superintelligence and so may need more exotic mitigation strategies, like performing original research that would otherwise not happen or teaching people to be more rational. And it's among those who look at exotic risks and find only need of exotic interventions that a particular sort of angst, ennui, and guilt can arise.

Now I don't much think of myself as an effective altruist, but I quack like one, so you might as well count me among their ranks, and many of my friends identify as effective altruists, so I speak from a place of gnosis when I say effective altruists suffer in these specific ways. And it seems to happen because people follow reasoning along these lines:

- We face existential risk X.
- The best option to mitigate X is action Y.
- I should do Y.

Sometimes this works out great and we get fabulous folks I'm glad to know working at places like CSER, FHI, FRI, MIRI, and others doing vital work on addressing existential risks. Other times it doesn't, not for lack of will or clarity of thought, but for lack of fit to perform the action Y. But, if you've reasoned thusly, you're now stuck in a spot where you know the world is at grave risk, you know something specific to do about it, and yet you don't do it. If you feel frustrated or disappointed or to have caused harmed, then you may feel the existential angst, ennui, or guilt, respectively, from not addressing X.

I see lots of evidence of this in rationalist and effective altruist culture. Perhaps no groups have ever worried more over akrasia and ways to combat it or the potential impact their possible work could have. Nate Soares wrote a whole series of posts on guilt because people kept coming to him saying "help, I feel guilty for not doing more," and there's lots of self questioning about how effective or altruistic the whole effective altruism movement is anyway. Even at EA and rationalist parties I hear many conversations about what people think they should be doing with their own lives and how they can make themselves do those things. A cynic might say the community is

actively encouraging feelings of angst, ennui, and guilt from its members. Yet somehow I feel none of these emotions. How do I do it?

Well, if I'm honest, I dramatically oversolve the problem with meaningness, but that doesn't mean I don't do things that specifically address the angst I used to feel over not doing AI safety research. After all, I think risks from superintelligence generate the greatest potential negative outcomes for the world, I think the best thing I could do about that is to dedicate my energies to researching and otherwise working to reduce those risks and weaken the badness of the outcomes, yet I am not working on AI safety research. Instead I do some other stuff that I like and give a little money to fund AI safety. How do I live with myself?

I think the answer is that I fully accept the concept of comparative advantage into my heart. Sure, I could make less money doing work I find less interesting and that I'm less good at to directly advance AI safety, but I instead make more money doing things I find more interesting and that I'm better at and give the excess to fund others working on AI safety. As a result, everyone gets more of what they want: I get more life satisfaction, someone else gets more life satisfaction because I help pay them to do the AI safety research they love, and we both still see an increase in the total amount of effort put into AI safety. And if it were just me making this tradeoff it might be a bad deal because I could subvert my comparative advantage and also work on AI safety and we'd make a little more progress on it, but there are others doing the same and together we are able to mine the vast wealth of our dreamtime to get everyone more of everything, including AI safety research, than we could have otherwise had.

But if economics has taught me anything it's that economic reasoning is unsatisfying to most people, so allow me to reframe my position in more human terms. Ability is unequally distributed: it's sad and true. Even with lots of hard work we cannot all do all the things we might like to do as well as we'd like to do them, even if we are all free to do the things we want to do as much as we want. And although you can feel angst that the world is unfair in this way, ennui for that which you can never have, and guilt over not fighting back harder against reality, there is great dignity in accepting the world as it is and finding contentment with it. So maybe you won't save the world and personally prevent existential disaster. That's fine. You can still do what you can towards preventing X, even if all you can do is be a silent supporter not actively opposing Y or other interventions to address X.

And, despite my own advice, I have some hope that the future may be different. Right now there does not seem to be a Gordon-shaped hole in AI safety research, but I keep an eye out to see if one appears. Maybe one day the cosmos will give me the opportunity to defend it against risk from superintelligence or another existential threat, but if it doesn't have to, so much the better—I can be "merely" happy with getting to live in the wonderful world others create for me. I hope such a "terrible" fate befalls us all.

Developmental Psychology in The Age of Ems

This is a linkpost for https://mapandterritory.org/developmental-psychology-in-the-age-of-ems-6f159f701354

Back in December Robin Hanson asked people if they would trade engagement with him on neglected topics. I couldn't trade with him because I had to make my ideas clearer first, and I still have a ways to go before I can reasonably ask many others to engage much with my ideas. But he's written an entire book on the implications of brain emulations, or ems, so it seems reasonable that I may try to engage his neglected topic of ems through my neglected topic of phenomenological complexity, and in doing so may make clearer my own ideas while more widely spreading his. If I'm lucky, I may even get some engagement back. Let's see what happens.

Robin focuses on a possible future where brain emulation—the ability to run a human brain on artificially constructed hardware like computers—develops before other similarly disruptive technologies like superintelligence. Based on the application of the strongest results in physics, engineering, economics, psychology, and sociology, he works out the implications of such an em scenario. You can read his book, *The Age of Em*, for all the details, but he gives a short overview in this TED talk.

To summarize, we can imagine em cities coexisting with human cities, only em cities are much denser than human cities in order to balance communication, cooling, and energy needs. Rather than experiencing life directly in these cities, ems will live in a virtual reality running orders of magnitude faster than human subjective time so that years in em subjective time might pass in weeks or days. Although there may be millions or billions of unique ems, most ems will be copies of the few hundred most successful ems, and further still most of those copies will be spurs that are deleted after completing a specific task. Ems will live near subsistence levels and need to constantly work to support the resources they run on or retire to slower speeds.

Importantly, since ems are emulations of human brains with little modification, ems can be understood psychologically as humans living in a very different environment, comparable in difference from our modern world to the difference between humans living in forager and farmer environments. Thus much of the analysis of em behavior comes down to taking humans who evolved for forager life, spent several millennia adapting to farmer life, then spent a few hundred years adapting to industrial life (a process which is still ongoing), and asking them to adapt to em life. Most of that change is likely to be cultural, but since the em age will last for subjective millennia for the fastest running ems, some amount of "biological" adaptation can be considered, similar in level to the kinds of adaptations populations living in farmer societies for millennia have picked up.

One of the biggest changes Robin expects from ems is a return to more farmer-like values, such as more respecting social hierarchy and more carefully planning for the future. Modern humans have moved away from farmer values during the industrial era dreamtime for more forager-like values, I suspect largely due to changing amounts of relative economic excess per capita. As a result ems are likely to be more calculating, fearful, judgmental, nepotistic, self-controlled, conservative, loyal, honest, self-

sacrificing, polite, hard-working, and religious than industrial era humans with far reaching effects given their technological advantages.

We see this in Robin's expectations for child-rearing, education, training, and work habits especially. Young ems and humans who may be uploaded are likely to be educated in more varied ways to increase the probability of big winners, with the best performers moving on to join the workforce. Early adulthood is likely to be filled with much high variance apprentice and journeyman work to prepare ems/humans for later work in a variety of roles, and the most productive adult ems (native and uploaded) will later take on master-level work around a subjective age of 40, assuming industrial era individual productivity trends hold. Without biological deterioration, adult ems will likely work for subjective decades to centuries before becoming insufficiently flexible in their thinking to remain productive, at which point they will likely be forced to retire to slower speeds to make way for younger adult ems to replace them, including possibly by younger versions of themselves.

This presents an environment in which we can analyze the phenomenological complexity of ems through the lens of developmental psychology, a topic Robin has thus far ignored presumably because there is not yet strong consensus on how developmental psychology works or if it's even a thing at all. I, however, think there are some robust results we can draw from the field, and I'll explore their implications in order of my confidence that the underlying theory is correct.

Caveats

But before we continue I want to note a couple things about what Robin has done and what I'm planning to do in response to his work. Robin is not a normal futurist. Most futurists present futures that are meant to reaffirm values they hold dear. For them futurism is less about predicting the future and more about expressing what they would like the future to be like. That's fine, but Robin is explicitly trying to give an accurate assessment of how he thinks an em scenario would play out without expressing an opinion on what he would like the future to be like. In fact, the things you may find repugnant about an em scenario are possibly similarly dissatisfying to Robin, but that doesn't change the nature of what we can reasonably predict. If we find the theorized em era in contradiction of our values, then so much the worse for our values.

The other thing to note is that what I'm attempting here is similar in tone to what Robin has done, so I'm going to describe possibilities rather than give you a specific narrative of the future. I don't know what the future will look like, so at best I can point in the direction of outcomes that are likely given what we can extrapolate about the em scenario now. Thus this is less futurism and more predicting the future, and so my claims are all tentative and couched in possibility and do not necessarily promote my values. Please keep that in mind when considering these ideas.

Growing up Em

The most robust result in developmental psychology is that people mature over time, where I take "maturation" to be a proxy for phenomenological complexity. This is said without explicit reference to notions of developmental stages, just a correlation of complexity with time, or if we're willing to take a small gamble to get a more precise statement, with variety of life experience. How does this suggest em society will differ from industrial society?

Since most ems will be subjectively middle-aged, the population of ems will skew much older than populations in most industrial societies. This implies that the average em will be more mature and phenomenologically complex than the average industrial human, so as a baseline we can look at how human preferences change with age to predict what most ems will prefer. As humans enter middle-age they remain concerned with personal values of pleasure, sex, power, prestige, and success that they hold when they are younger, but grow more concerned with abstract and social values like beauty, knowledge, health, stability, affection, belonging, support, obedience, religiosity, and tradition. This differs from a weaker abstract and social values focus when people are younger and a weaker personal values focus when people are older. From this we can expect em populations to share more balanced, conservative, and moderate views than industrial era human populations. This agrees with Robin's idea that ems will hold more farmer-like values.

Middle-aged maturity can be viewed not just as change with time but as a growth in understanding due to compounding complexity of thought. Until the full onset of senescence, older humans become increasingly likely to exhibit more complex behaviors as measured by Michael Commons' model of hierarchical complexity. I believe this is equivalent to saying they more think with greater phenomenological complexity as measured by Kegan, Kohlberg, and myself along the post-formal stages of thought from the traditional/pre-conventional/subject-level stage to the modern/conventional/subject-relationship stage to the post-modern/post-conventional/holonic stage. Further, although most humans in industrial societies still struggle to move beyond the subject-level stage, the people most successful in complex organizations are generally middle-aged and show signs of more often thinking at the subject-relationship and holonic stages. This suggests more ems will express thoughts and behaviors requiring subject-relationship and holonic complexity than humans do.

This may not be uniformly true of ems, though, because there are sometimes competitive advantages to less complex thought. An em who must do many repetitive, boring tasks will likely be happier if they less often experience complex thoughts because the level of task difficulty necessary for them to experience flow state will be lower. Additionally, greater phenomenological complexity appears to be mostly helpful in management and executive tasks and give little marginal advantage for performing tasks specified by others, so there may not be much selection pressure to encourage greater complexity among ems who function primarily as individual contributors within organizations. However it may still prove useful for individual contributors in em organizations if they more tightly coordinate than humans organizations do because phenomenological complexity enables more cognitive empathy and better communications skills.

So on the whole we should expect a smaller proportion of the em population to think primarily at the subject-level stage than among the human population and a larger proportion of ems to think primarily at the subject-relationship and holonic stages. In addition to further supporting the conclusion that ems will hold more balanced, moderate, and conservative views than humans do, this also suggests they will be better at self-control, more pro-social, and better at coordination than humans are on average. This slightly disagrees with the ems-are-farmer-like notion because it suggests ems will exceed industrial era humans in terms of social technology who themselves exceed farmer era humans in social technology, and unsurprisingly Robin draws a similar conclusion in *The Age of Em.*

Effects on Em Culture

After so much agreeing with Robin, you might wonder if there are any places where my expectations of higher average phenomenological complexity among ems predicts anything different. In fact, there are, and they come when we look closer at culture and social structures. For brevity, I will look at those topics on which I disagree with Robin. Note, though, that this section is far more speculative and based largely on my own observations and interpretations of capta rather than data. I think it's well reasoned for what it is, but new information could also easily surprise me and cause me to change my mind. I look forward to your comments.

Because most ems will hold post-conventional moral views, I expect em law to less depend on efficient enforcement of written rules and more depend on ambiguous social enforcement of conventions based on the judgements of respected ems. What written rules there are will be fewer and more strongly enforced when the consensus is that they are useful, but most of what is today "law" will instead become the unwritten rules of polite society. Ems will be free to defect when they so choose, with the main consequence being (temporary?) ostracism or exile, and most ems will choose not to defect because already living at subsistence levels they will not be able to afford to maintain themselves for long in such a situation. However this mostly applies to the "aristocracy" of the few hundred em copy clans that make up most of the population, and more efficient law enforcement may prove necessary for em clans with few or only one copy since among these ems society will likely be even more impersonal than it is in the industrial era.

For similar reasons I suspect em religions will be less farmer-like and continue the industrial era trend towards being less moralizing, more accepting, more encouraging positive behaviors, and more tolerating negative behaviors. Em religions will be more like Unitarian Universalism and modern Chan/Zen Buddhism and less like Abrahamic and Vedic religions because they will need religions that are more amenable to taking a complete stance. Thus em religions will mostly focus on ritual, social cohesion through positive feedback, and advice and less on metaphysics, social cohesion through negative feedback, and ethics.

I can be less certain of this, but I also predict ems will have much weaker notions of personal identity than even Robin suspects. Phenomenological complexity beyond the subject-level, corresponding to post-formal stages of developmental psychology, creates a more integrated understanding of an individual's place in the universe because the defining feature of the nominal holons of the holonic stage is that they have fluid boundaries that only come into focus by subjects taking particular perspectives. With many physical markers of individual identity made optional for ems, they will likely develop eusocial societies with ems playing roles similar to those of individual ants in a nest or bees in a hive. By late in the em era it seems possible that em society will look little like human society and much like an insect colony, at least among the few hundred copy clans that compose most of the population.

Conclusions

On the whole my analysis of the em scenario via phenomenological complexity agrees with Robin, and in those places where it disagrees the evidence for those disagreements is largely capta rather than data and so amenable only to phenomenological rather than scientific methods and thus outside the scope of what *The Age of Em* attempts. Taken together this seems to support the strength of Robin's

predictions and partially verify phenomenological complexity theory by seeing it generate results that agree with other theories so long as we restrict ourselves to evidence both may consider. Phenomenological complexity theory has a possible advantage of incorporating evidence unavailable to strictly scientific theories, but this relies on a willingness to employ phenomenological methods and evidence. Additionally, the em scenario is a rather strange theory to verify it against, but that’s what I get for engaging in neglected topics.

What Value Subagents?

This is a linkpost for https://mapandterritory.org/what-value-subagents-868b3b3fc076

Among rationalist thinkers the subagent theory of psyche is popular: CFAR teaches a version of internal family systems (IFS), Max Harms writes AI fiction from the point of view of a subagent, Brienne describes herself as metaphorically made up of multiple people, Alicorn seems to have been the first rationalist to explicitly suggest the idea, the notion goes back to our intellectual forebearers, and nearly everyone seems to use a subagent interpretation of dual process theory at the least. Yet although I've previously explored multi-part theories of psyche, I don't think of these parts as subagents. Instead, as I've hinted at, I think of the psyche primarily as a unified optimization process with inconsistent preferences that does not decompose into subagents. I didn't always think this way, though, and have at times thought of myself and others as made up of different configurations of subagents, so I want to explore some of how I came to give up subagent theories and see what value they may still hold.

Stepping back to start, what do rationalists mean by subagents of the psyche? "Agent" has roughly the economic sense of a process that makes decisions, but we should understand "decision" to include any kind of behavior, including thoughts, so in the context of theories of psyche "agent" is a rough synonym for "person" or "mind". A subagent, then, is an agent operating inside of another agent, forming a system of agents that externally present as a single agent. Explaining how subagents work in IFS should make this clearer.

IFS is a psychoanalytic model that, to my reading, originates from thinking of the Freudian id, ego, and superego like three homunculi competing to control the mind, but replaces the id, ego, and superego with groups of subagents serving exile, firefighter, and manager roles, respectively. These roles differer in both subtle and obvious ways from their inspirational forms, but the basic idea is that exiled subagents are those that would create undesired thoughts and behaviors, managers are subagents that work to keep undesired thoughts and behaviors from occurring, and firefighters are subagents that work to mitigate suffering when exiled subagents gain enough strength to influence a person's behavior. The psyche can then be thought of like a walled city where managers act as sentries to keep the exiles out and, when exiles inevitably make it inside, the firefighters serve as the garrison that chases the exiles out.

Rationalist thinkers have mostly dropped the Freudian shape of IFS, but they keep the idea of the psyche as a family of subagents. Rather than specifically exiled, firefighting, and managing subagents, the subagents are more often associated with observed aspects of an individual person's psyche and given attribute names like Lazy, Ambition, Face, and Pride. This internal family is frequently, when first identified, dysfunctional and fighting, corresponding to feelings of cognitive dissonance and conflicting desires. Through the application of various trainable techniques, though, the psychic family can be made functional and happy. Many people I know have expressed finding great value in this model, and they are living more satisfying lives through understanding themselves in this way.

I've, however, never much felt like I was made up of many subagents, and when I was introduced to the idea it seemed to me insufficiently parsimonious because it seems

to suppose the existence of neurological subprocess different from those that appear to exist. Instead for most of my life I've felt like my "I"—my psyche—was a homunculus trying to control a vast, semi-autonomous machine. This is to say my experience matches the description of dual process theory, sometimes referred to in this way as the conscious and subconscious, System 1 and System 2, or as the rider and elephant. The rider is an inner process that is the true self (Dasein if I'm generous; the authentic self if I'm not) and the elephant is an outer process that interacts with the world. The elephant is able to act on its own and has its own ideas about what to do, though, so the rider must make an active effort to direct the elephant. Dual process theory has some biological and evolutionary support, but even in the absence of that it's correlated with construal level theory, so this seems a potentially useful model.

But thinking of yourself as a rider and elephant can be highly dissociative. I often felt like I wasn't in control of myself and that I depleted my ego to do what I wanted. I thought of myself as suffering from akrasia because I could exhaust a lot of willpower trying to direct the elephant to little effect. And I lived like this from circa age 10 until I was 30. But as I transitioned from thinking in terms of systems to system relationships, the separation I felt between rider and elephant dissolved. Rather than being made up of two parts, it felt more and more like I was made up of just one. I could still see the shapes of the rider and the elephant, but now I could understand their interactions. This eventually led me to a holonic interpretation of the psyche where there are no distinct parts, but, as we'll see, there are many details.

That's because the holon of my psyche can't have no stuff in it: I know too much about how the brain works to really think of it as a black box. But I can understand it as a box full of gears, where the gears are preferences that interlock, sometimes harmonizing and sometimes grinding, and their combined motions produce my actual behavior. Being a holon, it's hardly clear exactly how the gears interact, but we have enough clues about the arrangement that it's possible to use it to make some predictions, not so strong that they risk cracking the epicycles and not so weak that we can claim anything. Unfortunately this theory makes heavy demands for ontological complexity when applied, thus it's likely not that useful to most folks seeking personal development.

Nonetheless, it's given me insight into the value others find in many-subagent theories of psyche when viewed through the lens of ego states or "masks". In *Impro*, Keith Johnstone describes the theatrical tradition of mask work where performers don masks and "become" the character a mask embodies, sometimes entering a trance where they forget themselves and effectively are the mask's character. When the psyche is viewed as a box of preferences this has a straightforward interpretation: the mask is a filter for expressing a subset of preferences. And if that's the case, then we can see psychic subagents as masks that reify subsets of preferences into agents.

In this light subagent theories of psyche make a lot of sense. Rather than being made up of many subagents, each subagent is a mask that allows a person to make salient their preferences that are normally lost in the cacophony of competing desires. By switching between masks you can give voice to the parts of yourself that are often ignored and even let the subagents talk to each other. But because these masks apply an ontology on top of a simpler system that can explain itself without them, subagents truly are like epicycles: they produce correct predictions, they are easier to work with than the full calculus of preference mechanics, and do this despite not describing reality parsimoniously.

So although I never did and no longer need to think of myself as made up of many subagents, I can understand them as an extremely useful form of play. My desire is then that they act as a bridge for people to develop more complete understanding of themselves.

Inscrutable Ideas

This is a linkpost for https://mapandterritory.org/inscrutable-ideas-c3211ee599af

David Chapman has issued something of a challenge to those of us thinking in the space of what he calls the meta-rational, many people call the post-modern, and I call the holonic. He thinks we can and should be less opaque, more comprehensible, and less inscrutable (specifically less inscrutable to rationalism and rationalists).

Ignorant, irrelevant, and inscrutable
*I have changed my mind. It should go without saying that rationality is better than irrationality. But now I realize…*meaningness.com

I've thought about this issue a lot. My previous blogging project hit a dead end when I reached the point of needing to explain holonic thinking. Around this time I contracted obscurantism and spent several months only sharing my philosophical writing with a few people on Facebook in barely decipherable stream-of-consciousness posts. But during this time I also worked on developing a pedagogy, manifested in a self-help book, that would allow people to follow in my footsteps even if I couldn't explain my ideas. That project produced three things: an unpublished book draft, one mantra of advice, and a realization that the way can only be walked, not followed. So when I returned to blogging here on Medium my goal was not to be deliberately obscure, but also not to be reliably understood. I had come to terms with the idea that my thoughts might never be fully explicable, but I could at least still write for those without too much dust in their eyes.

The trouble is that holonic thought is necessarily inscrutable without the use of holons, and history shows this makes it very difficult to teach or explain holonic thinking to others. For example, the first wave of post-modernists like Foucault, Derrida, and Lyotard applied Heidegger's phenomenological epistemology to develop complex, multi-faceted understandings of history, literature, and academic culture. Unfortunately they did this in an environment of high modernism where classical rationalism was taken for granted, so they failed to notice they were building off the strengths of modernism even as they derided its weaknesses. Consequently they focused so much on teaching the subjectivity of experience that they forgot to impress that it was subjective experience of an external reality and left their students with an intellectual tradition now widely regarded as useless for anything other than status signaling.

In comparison Buddhist traditions have, to the extent that bodhi is synonymous with meta-rationality and holonic thinking, done a better job of teaching post-modernism than the post-modernists did. Indic philosophical traditions hit upon post-modern ideas at least as early as the Axial Age and they became central to Buddhist philosophy around 200 CE. I'd argue the sutras of Buddhism do no better than the texts of the post-modernists at teaching holon-level thinking, but over the centuries Buddhist schools also developed tantric instruction that created environments in which practitioners were able to play and later work with holons. It appears to me these "esoteric" techniques tapped in to the same developmental psychology

operating in personal growth and in doing so created lineages that provide paths to holon-level thinking that people traverse to this day.

I suspect the key differentiator of the experiential learning of personal growth and tantra from the textual learning of academia and sutra is the focus on gnosis over episteme, and this suggests why the meta-rational is inscrutable from rationalism, but I'll do one better and prove it. To do that it will suffice to show that there exists at least one meta-rational idea that cannot be made scrutable to rationalism. I choose meta-rational epistemology.

Rational/modern/system-relationship epistemology aims to be consistent and complete, meaning it produces a complete ontology. To the extent that there is any disagreement over system-relationship epistemology it is disagreement over how to compute correct ontology. Meta-rational/post-modern/holonic epistemology denies the possibility of a complete ontology via consistent epistemology because epistemology necessarily influences ontology. That is to say, even if some epistemology is consistent, it cannot be complete because it cannot prove its own consistency, thus no consistent epistemology can produce complete ontology. Instead we might have a complete but inconsistent meta-epistemology that chooses between consistent epistemologies in different situations based on telos, like a desire for correspondence to reality or telling an interesting story, but telos asks us to make an axiological evaluation, not an epistemic one, and thus we are forced to admit that even our consistent and complete meta-epistemology needs a free variable, hence cannot actually be complete.

In this way holonic epistemology is necessarily inscrutable to system-relationship epistemology because it explicitly demands the latter do something it explicitly cannot. To be fair, system-relationship epistemology does the same thing to pre-rational/traditional/system epistemology by demanding consistency that the latter cannot tolerate because it would violate its internal completeness, but I think this is infrequently acknowledged because if you grew up in the shadow of the modern world you probably didn't notice when modernity demanded this of you. And unless you learned to ignore the problem, the modern world constantly gives you opportunities to experience the system-relationship level of complexity and obtain gnosis of it. But obtaining gnosis of the post-modern and holonic seems to require that a great tragedy befall you or that you have enough dedication to tolerate the pain of finding it, so beyond better building the episteme of holons for those few with more than doxa of them, I'm doubtful being less inscrutable will accomplish much of what Chapman seems to hope it will.

Embracing Metamodernism

This is a linkpost for https://mapandterritory.org/embracing-metamodernism-5d4ebe8a8ddf

I've never much felt like I was part of a cultural movement. I'm too much a "digital native" to be fully part of Gen X. I'm insufficiently idealistic to be a Millennial. I'm part of the transhumanist, the rationalist, and the effective altruist subcultures, but in a weak way that more resembles atomization than membership. And my philosophy is one of irreducible complexity. So I was surprised to discover I'm a metamodernist.

In a The Huffington Post piece from January, Seth Abramson describes metamodernism this way:

> [M]etamodernism believes in reconstructing things that have been deconstructed with a view toward reestablishing hope and optimism in the midst of a period (the postmodern period) marked by irony, cynicism, and despair.
>
> Generally speaking, metamodernism reconstructs things by joining their opposing elements in an entirely new configuration rather than seeing those elements as being in competition with one another. If postmodernism favored deconstructing wholes and then putting the resulting parts in zero-sum conflict with one another—a process generally referred to as "dialectics"—metamodernism focuses instead on dialogue, collaboration, simultaneity, and "generative paradox" (this last being the idea that combining things which seem impossible to combine is an act of meaningful creation, not anarchic destruction). Metamodernists will often say that they "oscillate" between extremes, which really just means that they move *so quickly* between two extremes that the way they act incorporates both these two extremes *and* everything between them. The result is something totally new.

Abramson goes on to examine how metamodernism manifests in music, art, film, literature, and memes and finds examples in the Childish Gambino, Shia LaBeouf, My Dinner with Andre, David Foster Wallace, and The Bee Movie. Elsewhere Abramson has compared metamodernism to other living cultural philosophies and sees similar relationships to those Chapman sees between different modes of meaning. Of greatest interest to me, though, is that metamodernism gives wider context to the philosophical work of myself and other post-rationalists.

Or maybe we should be the meta-rationalists to allude to metamodernism since there is much in metamodernism that jives with the meta-rational, especially the notion of reconstructing the deconstructed. Perhaps why we've so far struggled to make ourselves comprehensible is that our closest intellectual cousins, the post-modernists, were largely content to deconstruct things without reconstructing them. Comparing ourselves to them we are forced to explain how postmodernism failed despite having a good start, but if we compare ourselves to the metamodernists then the story is simpler because we also look to move beyond deconstruction to reconstruction, and not in spite of deconstruction but in the spirit of it. In this way our divergence from the rationalists seems to go beyond epistemological differences to a belief that the world can be deconstructed while maintaining its shape rather than being reducible.

[tennis is basically baseball]

Sorry if that seems like a lot of inside baseball, but how meta-rationality is understood by others has been a salient topic for me of late. Rest assured that my drafts folder is filled with posts on the relationship between feedback and suffering, the important of regression to the mean, and my ever elusive mathematical foundation for phenomenology. I expect to tackle all those topics in the next few months.

Is Feedback Suffering?

This is a linkpost for https://mapandterritory.org/is-feedback-suffering-cf18006deca8

NB: Some of the terminology and concepts I used here are incompatible with my more recent work because this was written before I formalized my philosophy, but is dangerously close enough to them that it may cause confusion. Caveat lector.

In the future, the world may be filled with creatures other than humans and animals. There may be ems, superintelligences, and other sorts of processes we'd recognize as alive, and it seems likely that there will be many orders of magnitude more of these living things in the future than at present. As a result there may be many orders of magnitude more future suffering. Depending on your moral stance, even if you're an anti-realist like me and simply express a preference for the satisfaction of preferences of others, this means that the lives of these future creatures are of great concern.

Brian Tomasik and others with the Foundational Research Institute have been exploring the interaction between suffering-focused ethics and the far future for a few years now. Their work has included measuring suffering and happiness, formalizing ethical calculations, and understanding the benefits of compromise with other value systems, but of particular interest to me are Brian's looks into issues around suffering of minimally conscious processes because his thoughts on consciousness resemble mine. Thus when I read FRI researcher Lukas Gloor's recent piece on tranquilism I was primed to find myself wondering, rather alarmingly, is feedback suffering?

To understand why, recall that feedback is another name for the process by which a subject is phenomenologically conscious and experiences itself as object. Further, most feedback is either positive or negative, meaning it causes action towards or away from a state, respectively. As humans we experience positive and negative feedback as a desire that the world be in a state other than the one it is currently in, and tranquilism suggests that suffering arises from desire since not experiencing desire yields a state we call contentment. But if we can avoid suffering by not experiencing desire, which is to say positive and negative feedback, then it looks a lot like feedback is inherently painful and thus the source of suffering. That's distressing because it implies most existence is pain and almost all physical processes suffer constantly, so it seems to me worthwhile to make a formal study of suffering to see if we might poke some holes in this line of reasoning.

I'll do this from an existentialist phenomenological stance, which is to say that stuff exists, stuff is only known through experience, and experience is an inseparable, directed, intentional relationship between stuffs where we term the experiencing stuff "subject" and the experienced stuff "object". Consequently in order to be precise and to avoid equivocation I'm going to have to write the word "experience" a lot, so my apologies for the word soup that sometimes follows.

Phenomenology of Suffering

Whether or not feedback is suffering is ultimately a question of axiology, the study of value. That is, when we ask if feedback is suffering we are asking a question about how to measure the value of feedback. Since I'm an existentialist and a phenomenologist, I'm additionally willing to say that we must calculate value for ourselves rather than find it because there is nowhere for it to come from other than

our experiences, and since value is a measure we must then choose what to measure against. This choice forces an inescapable telos upon value, and in this case that telos appears rooted in emotion since we think of suffering as an emotional state with negative valence. Further, since I feel positive emotional valence (happiness) when other people feel positive emotional valence and negative emotional valence (sadness) when other people feel negative emotional valence—viz. I am empathetic and care about other people—and most other people feel the same way, our telos is additionally ethical because we want to know how our actions will affect the happiness and sadness of others. Thus, as you probably already surmised, we're going to be valuing if feedback is ethically desirable or not.

Ethical systems, being axiological systems for the purpose of judging intent, are calculi for how to combine experience via relations that yield value. We can additionally place optimization constraints on these systems, like maximizing the value of preference-satisfying experiences or minimizing the value of preference-violating experiences, to produce specific axiologies, like hedonic axiology and negative hedonic axiology, respectively. This gives us a language in which to talk precisely about axiological and ethical systems, and in this language tranquilist axiology is the axiological system where experiences are combined to maximize the value of experiences of contentment.

To be intellectually honest to myself about what tranquilism means, I need to phenomenological deconstruct contentment. Descriptions of contentment make it sound something like an experience of anhedonia where the subject experiences indifference towards suffering and pleasure, but this seems unsatisfactory because these descriptions also assign contentment a positive rather than neutral valence. Lukas similarly agrees that contentment is a happy experience when he writes that contentment is "untroubled by any cravings for more pleasure", "experienced as completely problem-free", and that "conscious states completely free of cravings should thus elicit very positive associations". So it seems when one is content one is happy, indifferent to more happiness, and does not suffer, but this appears contradictory because to be content one would have to be happy about preference violation, a definitionally sad experience because negative valence emotions are the feedback mechanism we use for experiences of preferences.

This is only a problem, though, if we limit our axiology to valuing experiences and experiences of experiences. If we include experiences of experiences of experiences in our ontology rather than compressing them into experiences of experiences, thus making a systems-level demand for ontological complexity, we can understand contentment as an experience of happiness towards all experiences of experiences and use it to direct tranquilist axiological reasoning. In this way contentment wraps happy and sad experiences in an experience of happiness, making it of a different type and avoiding apparent contradiction by adding happiness to rather than changing the original experience from pleasure or suffering.

Having deconstructed contentment and understood tranquilism precisely, it appears my original concerns about feedback being suffering were confused because if suffering, a negative valence experience, is something we can be content with, then in this context suffering must be an experience of experience and thus feedback cannot necessarily be suffering because feedback can exist as a direct experience not just a meta-experience.

I seem not alone in needing to clear this confusion because our everyday use of the word "suffering" points to at least two different categories the same way

“consciousness” does. Just as we can separate naive notions of consciousness into phenomenological consciousness (self-experience) and phenomenological sentience (experience of self-experience), we can separate suffering into phenomenological desire (intention to make the world otherwise) and phenomenological suffering (experience of negative valence over desire). Cleaved in this way we see that feedback is desire but not necessarily suffering, but suffering is often confounded with desire in sentient subjects because meta-experiences can be both desire and suffering. This unfortunately leaves open the possibility that the feedback of sentience is often suffering even if it doesn’t have to be.

Panpsychism and Suffering

If something must be sentient to suffer, it may seem we need only worry much about suffering among animals and animal-like things such as ems and AIs, and even their suffering is only ethically relevant as far as your empathy extends. But what if more things are sentient than we think?

Phenomenological consciousness implies a weak form of panpsychism—the idea that everything is at least a little bit conscious. But most people don’t care about the phenomenological consciousness of rocks and chairs, so there may be panpsychism but it’s a very boring kind of panpsychism that isn’t affecting anyone’s ethics. A strong panpsychism based on sentience would be another story since many people, including many effective altruists, say sentience is the criterion for ethical relevance.

Many simple control systems, such as thermostats, may have phenomenological sentience because their operation can be construed in terms of meta-experience of self. The leading mathematical theory of sentience, Integrated Information Theory, would seem to agree, likely scoring such systems as minimally “conscious” (IIT uses “consciousness” where I use “sentience”). Although it seems unlikely that something as simple as a thermostat experiences suffering in the sense of a negative valence experience of desire, perhaps there is some way in which meta-experiences we’d recognize as suffering can be defined that do not depend on negative valence emotions. This is important because it would allow us to identify suffering or suffering-like experiences in subjects that do not have something resembling the evolved emotional systems of animals.

I don’t (yet) have any further thoughts on such suffering-like experiences, but having a phenomenological deconstruction of suffering it may prove possible to make progress on understand suffering in terms of non-emotional experiences.

TL;DR

Okay, that wasn’t too long, but it was pretty dense, so in case you were wonder what the heck I was saying, here’s the short version:

- Feedback is desire but not necessarily suffering
- Suffering is a kind of feedback with negative emotional valence and animals seem to experience desire as suffering.
- Contentment wraps up suffering in happiness by adding ontological complexity to create distance from the experience of negative emotional valence.
- Many things may have some sentience and so may be able to suffer or experience something like suffering.

- I’m going to think more about if there is some way to talk about suffering prior to emotion so we can considering the suffering of non-emotive sentient beings, especially for those who do not have the ontological complexity to learn contentment.

Cognitive Empathy and Emotional Labor

This is a linkpost for https://mapandterritory.org/cognitive-empathy-and-emotional-labor-cb256c38597d

The concept of emotional labor has been popularized in the last couple years as a way of talking about the work people do to manage other people's emotions. Most of that discussion has been around how women are often expected to perform emotional labor without compensation both professionally and personally. Women report being asked to perform social glue functions in the workplace without it being part of their job, part of how they are evaluated, or part of how they are paid, and they are culturally expected to perform most of the emotional labor in personal relationships. And perhaps most frustratingly, while men are lauded when they perform emotional labor and mostly given a pass when they don't, the situation is reversed for women who mostly only see punishment for not doing enough.

But ultimately emotional labor is for everyone, and although there is a sex differential in its performance, there is little new I can say on that aspect of the topic. What I can say is something about how emotional labor is related to developmental psychology and cognitive empathy. Specifically, how skill at emotional labor depends on the development of cognitive empathy and lack of cognitive empathy is a limiting factor in being able to perform emotional labor.

I described emotional labor as "managing other people's emotions", but to be more precise emotional labor is acting to influence the emotions of others. To do this one must have some knowledge of the emotions of others and how they can be affected. This knowledge typically comes from either affective empathy or cognitive empathy. Affective empathy is feeling another person's feelings, like being sad because your friend is sad or being scared because a character in a horror movie is scared. Affective empathy's source is probably mirror neurons, and a lack of affective empathy is associated with sociopathy. For this reason affective empathy is sometimes also called "primitive" empathy because it seems to naturally develop on its own and is rarely missing in a person.

Cognitive empathy, on the other hand, is the skill of thinking about others ontologically and is anything but "primitive". In order to be able to think of ways to make your friends happy or worry about what others will think of you, you must model other people and predict their responses. Those models can be simple, like how children employ thing and thing-relationship levels of phenomenological ontological complexity, but such simple models often fail if not backed up by affective empathy. As people age they build up enough cognitive empathy to effectively participate in society without necessarily feeling everyone's feelings, and they develop system level and higher ontological complexity that enables cognitive empathy techniques like seeing other people as made up of parts, distinguishing others' revealed and stated identities, and understanding others' needs and wants. And if a person continues down this path they may develop a generalized sense of cognitive empathy that can tackle broad axiological questions about how to treat themselves and others.

Yet affective empathy and cognitive empathy rarely exist in isolation. In the context of emotional labor, people often first feel—use affective empathy to notice—that an

opportunity exists to affect someone else's emotions, and then use cognitive empathy to figure out what to do. And when cognitive empathy fails us we may fall back on affectively informed actions. This will work most of the time, but pesky philosopher that I am, I want to know what happens in the edge cases, like when you can do something to hurt someone else's feelings without them finding out.

Consider the case of the broken vase. I'm having a fancy dinner party and you lend me your vase to use as a centerpiece. On the way home I stumble and drop the vase, shattering it into a million pieces. Luckily this happens right in front of a store where I can purchase an exact replica, so I immediately replace it. The dinner party goes well, and I "return" the vase with you unable to tell I've replaced it. I have two options:

1. Say nothing about the break.
2. Tell you that I broke the vase and replaced it.

If I have no affective or cognitive empathy it seems likely I will do (1) since it is naively the option that produces the better payout: you'll be mildly happy in (1), whereas there's some chance you'll be angry in (2). If I have no cognitive empathy but plenty of affective empathy, it seems likely I will do (2) because I will feel the bad feelings you would feel if you knew I broke your vase and won't think about the fact that you don't know that I broke it. If I have no affective empathy but plenty of cognitive empathy, though, it now becomes a bit more complex to figure out how I will act. Maybe I will want to spare your feelings and do (1), but maybe I will reason that you would want (2) because it conveys information about me you want to know, and I do it out of a reasoned expectation that acting in this way will more create the world I want to live in. And the situation remains substantially the same if I have plenty of affective empathy to go along with my cognitive empathy, however my feelings will likely affect my axiological calculations in deciding which action to prefer.

In this scenario cognitive empathy enabled emotional labor. Without it I was left either playing a simple game or acting on my feelings with no consideration for you, and so my emotional "labor" was reduced to calculating a payout matrix or dealing with my own conflicting emotions. Cognitive empathy made possible real emotional labor, though, because it provided an ontology to reckon with. True, you might object, it still produced one of the outcomes that could be achieved without cognitive empathy, but emotional labor matters at the margins when we consider many cases where acting without cognitive empathy would produce inconsistent answers.

This is important because without doing enough emotional labor to come to a wise course of action, a desire to help someone borne out of earnest empathy for them may end up unintentionally hurting them. If I failed to understand you sufficiently well in the vase case when I was using cognitive empathy to perform emotional labor, I might have chosen to do (1) when actually (2) is what you would have preferred or vice versa. When we help we risk hurting if we do so unwisely, so helping depends on having the skill to accurately predict how our actions will affect others. This is why people say emotional labor is draining: not only is it mentally challenging but the stress of failure can weight so heavy on us that we find it hard to act.

So what can you do if you want to perform the emotional labor that will allow you to help others as they want to be helped? My own solution is to target virtue when my calculations are insufficiently calibrated, but otherwise you might take the same advice I'll always give: do the emotional labor you fear doing and give up on hoping for the emotional labor you want done for you, because even if you hurt people in the

short run this will enable the necessary personal growth towards increased complexity that will let you help them in the future.

Regress Thyself to the Mean

This is a linkpost for https://mapandterritory.org/regress-thyself-to-the-mean-932d5fd9789d

I don't usually write about well understood things: others are better explainers than I am, and I have more fun working at the edge. But a few weeks ago I was commenting on a Facebook post and the exchange went something like this:

> **Them**: I think [subculture] is fading out because it's being softened by the surrounding culture of the city.
>
> **Me**: Actually, that seems insufficiently parsimonious to me. I think a simpler explanation is that people in [subculture] are aging.
>
> **Someone else**: You don't even need to suppose it's aging related. It's probably just regression to the mean.
>
> Me:
>
> 
>

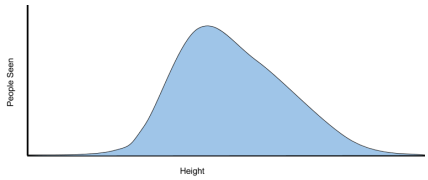
So to atone for my sin of neglecting the most likely reason anything ever happens, I present to you my penance—an introduction to regression to the mean.

Regressive Statistics

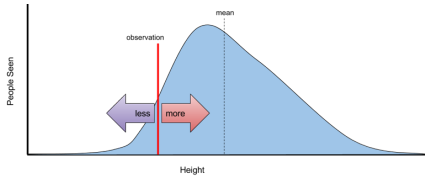
Suppose you are sitting on a park bench, watching people walk by. You notice the height of each person that passes you and develop a sense for how the height of people at the park is distributed. After some time a friend joins you on the bench and, hearing about what you've been doing, ask you if the next person to walk by is likely to be taller or shorter than the last person who did.

Regression to the mean tells you how to answer this. Whatever the distribution of the heights of the park goers, each person who walks by, lacking any other information to inform your decision, is likely to be in the direction of the mean within the height distribution relative to the previous person. So if a short person walks by, you tell your friend the next person is likely to be taller and vice versa, and most of the time you will be right!

There are several ways to understand why this works, but I think the most straightforward is to think in terms of sampling from a distribution. You know the heights of the people you've seen in the park, so if you charted a histogram of their height measurements you'd get a picture that looked approximately like this:



Then if you observe someone of any particular height, you can divide the histogram into two sections: those shorter than the observed height and those taller. Because of the way the data is distributed, though, the side with the mean will always have more of the probability density than the other side, so it is always more likely than not that the next sample will be in the direction where the mean is located. Thus the tendency to regress towards the mean.



Note though that regression to the mean does not imply that the next person to walk by will likely be closer to the mean, just more likely to be in the direction of it relative to the last sample. The distinction is important because for observations close to the mean the opposite—that the next person is farther away from the mean than the previous person—is more likely to be true because of variance. For example, if a person is only 1 inch taller than average but the variance in height is 6 inches, the next person is more likely to be more than 1 inch shorter/taller than average than the last person and thus farther away from the mean. After all, regression to the mean applies to variance too.

But you probably don't care much about how tall people are, so I hope you are asking yourself "what else can I use this for?". As it turns out, just about everything.

Put a Number on It

I might not go as far as Jacob Falkovich in my desire to put numbers to things, but a lot of opportunities open up if you think of the world as measurable, even when it seems totally impossible to take measurements. Being able to use regression to the mean as an explanation is one such opportunity.

Take the conversation that spawned this post as an example: someone said that a subculture had grown "less distinct" over time, but for a subculture to be less or more distinct means we must be able to measure it in terms of distinctness. It doesn't really matter for our purposes how this measurement is done, just that we can take a measure. Once we have our measure we can take sample measurements of different subcultures at different times and come up with a distribution of subcultures' distinctnesses. Then it's straight forward to say that more distinct subcultures are likely to become less distinct over time and vice versa because subsequent measures are likely to regress to the mean. If it looks like a subculture is less distinct now than it used to be, the obvious line of reasoning should be that its distinctness is regressing to the mean.

We can similarly apply regression to the mean to help understand almost anything we can measure—test scores, fashion, salaries, productivity, prestige, cuteness, etc.—

but only if we apply it correctly. Consider students in a class where their grade is determined by a midterm and final exam. Does regression to the mean imply that students who do poorly on the midterm will do better on the final exam and vice-versa? No. The trouble is that the students' scores are not identically and independently distributed random variables: how they perform on the midterm impacts how they will perform on the final. Some possibilities:

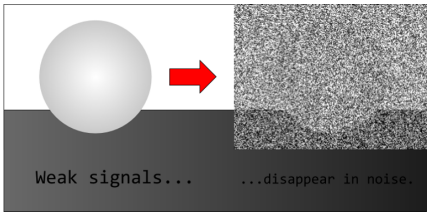
- A student who gets a bad score on the midterm may study harder to get a better score on the final so they can pass the class.
- A student who gets a good score on the midterm may slack off because they can get a lower score on the final and still pass the class.
- A student who does well on the midterm may also do well on the final because they study hard all semester.
- A student who fails the midterm may give up and drop the class or stop studying and intentionally fail.

In the language of control systems we might say the students are negative feedback loops working to keep their inputs on target, and in the language of information theory we would say that our samples (test scores) from each random variable (student) possess mutual information. We won't see regression to the mean in this case, though, even if all the students followed a move-in-the-direction-of-the-mean pattern, and any such resemblance to regression to the mean would be coincidental, because regression to the mean is what happens when nothing tries to affect the value measured. If work is being done then regression to the mean is not happening.

Entropic Regression

This sounds an awful lot like regression to the mean should never happen since the universe is full of things doing work, everything is tangled up with everything else, and mutual information always exists except when it doesn't matter. We could suppose to the contrary that every instance of what we think is regression to the mean is actually negative feedback, but this hardly seems parsimonious since nothing happening is generally more likely than something happening, and any feedback whatsoever demands at least phenomenological consciousness. So how can it be that we see regression to the mean so often if in reality the conditions of its operation are never met?

The answer is entropy. Regression to the mean isn't a process so much as a manifestation of the general tendency of the universe to move towards low energy states as seen through the lens of statistics. Thus it's not that the existence of mutual information prevents regression to the mean from happening, but that mutual information is the work being done to counteract regression to the mean, and if the amount mutual information is too small—if not enough entropy is displaced to notice a local rise in complexity—then it fades into the background noise and goes unnoticed because it's weaker than the static that is regression to the mean.



This is the “real” reason regression to the mean explains so much: it’s what happens by default if nobody does anything. Regression to the mean, entropy, static, noise, and kipple are all the same “thing” in that they are all the no-thing of emptiness that fills the void between intentional stuff. They are the yin to purposeful work’s yang; the resting state everything returns to absent effort. Thus regression to the mean is one of the many ways we see the nothing out of which the workings of extropy or Dao are born.

Doxa, Episteme, and Gnosis

This is a linkpost for https://mapandterritory.org/doxa-episteme-and-gnosis-ea35e4408edd

Ancient Greek famously made a distinction between 3 kinds of knowledge: doxa, episteme, and gnosis.

Doxa is basically what in English we might call hearsay. It's the stuff you know because someone told you about it. If you know the Earth is round because you read it in a book, that's doxa.

Episteme is what we most often mean by "knowledge" in English. It's the stuff you know because you thought about it and reasoned it out. If you know the Earth is round because you measured shadows at different locations and did the math to prove that the only logical conclusion is that the Earth is round, that's episteme.

Gnosis has no good equivalent in English, but the closest we come is when people talk about personal experience because gnosis is the stuff you know because you experienced it. If you know the Earth is round because you traveled all the way around it or observed it from space, that's gnosis.

Often we elide these distinctions. Doxa of episteme is frequently thought of as episteme because if you read enough about how others gained episteme you may feel as though you have episteme yourself. This would be like hearing lots of people tell you how they worked out that the Earth is round and thinking that this gives you episteme rather than doxa. The mistake is understandable: as long as you only hear others talk about their episteme it's easy to pattern match and think you have it too, but as soon as you try to explain your supposed episteme to someone else you will quickly discover if you only have doxa instead. The effect is so strong that experts in fields often express that they never really knew their subject until they had to teach it.

In the same way episteme is often mistaken for gnosis. At least since the time of Ptolemy people have had episteme of the spherical nature of the Earth, and since the 1970s most people have seen pictures showing that the Earth is round, but astronauts continue to experience gnosis of Earth's roundness the first time they fly in space. It seems no matter how much epistemic reckoning we do or how accurate and precise our epistemic predictions are, we are still sometimes surprised to experience what we previously only believed.

But none of this is to say that gnosis is better than episteme or that episteme is better than doxa because each has value in different ways. Doxa is the only kind of

knowledge that can be reliably and quickly shared, so we use it extensively in lieu of episteme or gnosis because both impose large costs on the knower to figure things out for themselves or cultivate experiences. Episteme is the only kind of knowledge that we can prove correct, so we often seek to replace doxa and gnosis with it when we want to be sure of ourselves. And gnosis is the only kind of knowledge available to non-sentient processes, so unless we wish to spend our days in disembodied deliberation we must at least develop gnosis of doxastic and epistemic knowledge to give the larger parts of our brains information to work with. So all three kinds of knowledge must be used together in our pursuit of understanding.

Unfortunately we tend to forget this and often privilege one kind of knowledge or discount another, resulting in several possible ways of failing to embrace the full richness of knowledge available to us. When a person downplays doxa relative to episteme and gnosis we might say they are too skeptical if we are generous and a sophist if we are not. When a person ignores episteme in favor of doxa and gnosis we might say they are unscientific or irrational. And when a person emphasizes doxa and episteme to the exclusion of gnosis we might call them an empiricist or, to be provocative, a rationalist.

In each case the cure is different. The sophist must learn compassion and empathy to be able to trust that producers of words, including their own past selves, may imbue them with some correlation to reality. The irrationalist must learn logic and parsimony so that they may accurately predict the world they will find themselves in. And the rationalist must learn humility enough to trust experience when facing what they don't (or can't) know through doxa and episteme. None is better than the other; all three are missing out on the fullness of what we can know.

It's coincidental that ancient Greek chose to break knowledge into three kinds rather than two or four or five, but because it did we can think of doxa, episteme, and gnosis like the three legs of a stool. Each leg is necessary for the stool to stand, and if any one of them is too short or too long the stool will wobble. Pull one out and the stool will fall over. Only when all three are combined in equal measure do we get a study foundation to sit and think on.

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.

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.

Suffering and Intractable Pain

This is a linkpost for https://mapandterritory.org/suffering-and-intractable-pain-e7115a5acae3

I've been writing a lot about AI alignment lately, so let's take a break from that to talk about a lighter subject: suffering.

Suppose you contract the rare and as yet untreatable disease boneitis. The doctors estimate you have one year to live. In response you might spend the last year of your life living it up, pour all your efforts into finding a cure, or become so depressed that you end your own life before boneitis can. Whatever course of action you might take, we can measure it along a dimension ordering your potential responses from least to most trying to affect change in the world, ranging from totally accepting and abiding the disease to totally rejecting it and working to prevent it from killing you. Let's call this the accept-reject measure. Where do you think, given things I have written, I would personally advise you to fall along this dimension?

I ask because several people have expressed to me a belief that I would take an accept-only strategy in this and similar scenarios, and this is a dangerous misunderstanding of my thinking to the extent that others may model their actions on my writing. Certainly I think most people would benefit by their own standards to abide more and try to change the world less, but how I see this working is perhaps even more important than that I see it, because naively applied a move in this direction encourages quietism, deathism, and general tolerance of suffering. And the key to how I see more abiding helping is through more abiding only intractable pain so we can transform it from suffering into a neutral or even positive experience.

The Phenomenological Origin of Suffering

We start by asking, "what is suffering?". Or, more tractably, "why do I suffer?". We approach this phenomenologically via a reduction of suffering, and to do that we need first to identify the intentional relationship under consideration, but we run into a problem right away because in English we say "I suffer", making the object of experience unclear. Thus before our reduction even begins we find ourselves challenged to understand what it looks like to suffer.

To get our ground, let's explore the etymology of "suffer". It has its roots in Proto-Indo-European "bher", the action of carrying or bearing as in "I carry/bear a heavy load". Latin added the "sub-" prefix (which mutated to "suf-") to give the idea that the bearing was done beneath something and this came to have a metaphorical sense in which you might "carry on under oppression" or some such. This gets us to the archaic sense in which people said that they suffered fools gladly and witches to live meaning that they tolerated something, and this eventually evolved into the modern form where it means more generally to experience something bad such as suffering a pain. Thus although "to suffer" has an intransitive form, it is clearly rooted as a transitive verb where the subject suffers something.

What might be the object of suffering, then, when we say more generally "I suffer"? Since the sense in which we mean "to suffer" here is to experience something bad, I will take that as our starting point and consider the reduction of "I suffer" to be equivalent to the reduction of {I, experience, something bad}. But what is this

something bad? There is a sense in which “bad” could mean “morally wrong” if you believe in the existence of moral facts, but there is a broader sense in which “bad” simply means something you don’t want or that you desire not to see in the world, so to ask what is bad seems to ask what it is to prefer or value against something.

Whence does againstness rise? To go against, to resist, or move away from something is for a subject to observe the world and act to change it. As with any kind of reaction, againstness implies a direction, derivative, or rate of change, so the subject must not only observe the world but repeatedly observe it and assess what action to take to move against relative to where the world is now. This means including part of the world in itself, either ontically or ontologically, and so sets up a loop that makes the subject cybernetic. Thus to talk about againstness we are necessarily dealing with feedback.

In humans and other animals feedback (known as homeostasis in the case of negative feedback) is implemented by various systems including hormones, physical pressure, and neuronal activity. Some of these systems act without the subject consciously experiencing the feedback process at work, and the subject is only aware of anything happening, if at all, when they experience the effects of the feedback process on the affected system, such as when physiological responses cause the heart to beat faster and the subject may feel the change in heart rate but not the experiences which caused the heart the beat faster. Of course the heart can also beat faster in response to observed experiences like stress, so this highlights that not all feedback is equally experienced by the conscious mind, and invites us to ask how suffering relates to consciousness.

To speak of the object of experience as “bad”, as in the case of suffering, is to imply an ontological understanding of the object and hence a sign of phenomenal consciousness. Thus {I, experience, something bad} must contain a nested experience to create the something bad within the ontological, so we might expand it to {I, experience, {I, experience as bad, something}} to see a deeper aspect of suffering, viz. the experience of negative valence experience. To distinguish it from suffering we will call the nested negative experience within suffering “pain” following the Buddhist distinction between suffering and pain, but using a broad sense in which pain includes any experience felt as having negative valence. And pain makes clear that we have lost something of the nature of suffering with this reduction.

Specifically the epoche of {I, experience, something bad} to {I, experience, {I, experience as bad, something}} renders suffering indistinguishable from the conscious experience of pain, yet as just mentioned Buddhism claims a distinction between suffering and pain and Buddhist practitioners report capta of neutral and positive experiences of pain they would not consider suffering. Thus despite what suffering initially seemed to be, it must have a second defining characteristic—the experience of pain as pain—or more specifically the phenomenon {I, experience as bad, {I, experience as bad, something}}. In this way we are led to an epistrophe showing suffering to not merely be the experience of pain, but to be the pain of experiencing pain, and in this sense “to suffer” literally is to bear pain under pain!

Separating Suffering from Intractable Pain

That suffering is a phenomenally conscious experience carrying negative valence suggests suffering plays a role in feedback systems within the cybernetic systems of animals that create consciousness, thus we may ask if it is wise to eliminate it. I think

it can be, but only when done carefully with a mind to reduce suffering by increasing ontological complexity to address sources of intractable pain.

By intractable pain I simply mean those sources of pain which cannot be addressed by changing things in the world other than your perceptions. For example, if there were a pin sticking into your skin causing pain, you might be able to eliminate the pain by removing the pin, so this would probably not be intractable pain. On the other hand, if you have a disease like fibromyalgia you might feel pain because your nerves falsely report it in the absence of an appropriate stimulus, and this would probably be intractable pain so long as no treatment existed to ease it. Of course fibromyalgia pain may not always be intractable since new drugs may be invented to treat previously intractable pain, so intractability reflects a property of pain in the moment rather than being essential or eternal, and hinges on whether or not there is some change that could be made to make the pain go away other than changing your perception of the pain.

For tractable sources of pain the correct course of action is probably to respond to the pain and change the world to avoid it—if your hand hurts because it's in a fire, pull it out; if you feel depressed because you are hungry, eat; and if you are anxious because you don't know something, learn it. In these cases pain is acting as a signal to get you to do something that you will probably like or, at the least, may provide evolutionary fitness, and so to the extent you want to be happy, continue living, and reproduce, acting in response to pain is adaptive. It's only for intractable sources of pain where the experience of suffering is not adaptive because it cannot lead you to make any change that will better satisfy your desires, and so it is only in this case that learning to experience pain without suffering is much useful.

There are two catches here, though. The first is that to learn to abide pain with tranquility rather than suffering you must practice abiding pain, and the easiest pain we have access to practice with is tractable pain since it can generally be created as easily as it can be stopped. Practicing with tractable pain to learn to abide intractable pain is likely part of the etiology for ascetic practices surrounding hunger, sitting, and celibacy. The second catch is that we may often be mistaken about what pain is tractable. That is, although we can often address surface-level sources of pain like hunger, they are usually manifestations of deeper sources of pain which may be intractable, like a desire for the world to always be better no matter how happy we are with it now. Thus much tractable pain has an intractable part that is left unaddressed when the tractable part is eased.

These two catches I think explain much of the nuance of my stance on suffering. On the one hand I view suffering as a point of practice from which growth is possible and suffering as something worth seeking out and working with. On the other I also view some pain as inescapable because pain is rooted in the operation of the feedback processes that make us both cybernetic and phenomenally conscious. Combined this means I value suffering for its instrumental value in serving the purpose of overcoming suffering, especially that suffering which is based on intractable pain, and in this way my thinking parallels Stoic and Buddhist views on suffering but with perhaps greater attention to the phenomenological distinction between pain as a cybernetic phenomenon and suffering as a quale.

I’m tempted to leave off there, but I think there is one more aspect of suffering and learning to overcome it through abiding pain that is part of my view that needs emphasis, and it ties deeply into my own terminal values. For although I practice Zen and predict that in the future there will be less “I” inhabiting my body-mind and more no-self and non-dual experience, I find that my I craves completeness or perfection. Now, in a proper sense I am already perfect and I just don’t realize it, but my inability to experience total gnosis of this (yet) motivates my desire to complete myself through realizing my potential and increasing my psychological complexity. Thus I value suffering and overcoming suffering through tranquility and contentment because they are the mechanism by which I grow towards completeness, perfection, and realization of Buddha-nature.

And on this point I do perhaps radically differ from many of my peers. I view our great project as not necessarily to end suffering, but to create a world where the only suffering is that which is necessary to the work of spiritual/psychological realization and then only for those who would seek it out rather than have it thrust upon them by an uncaring world. Until then we are stuck in a world where everyone must suffer whether they want to or not, and the only path the serenity seems through learning to abide pain.

Akrasia is confusion about what you want

> *It's 2 pm. You've had a report to write since yesterday morning but you just don't feel like doing it. You've tried to start on it several times, but each time your mind simply refuses to engage with the task. You stare at the screen for a while, hands perched over keys, but nothing comes. Eventually you do something else for a while hoping inspiration will strike while you're away, but instead you spend hours on other tasks while the report languishes. Eventually it's 6 pm, the report was due at 5, so you work late and force yourself to get it done, only finally making progress because you feel the threat of consequences for not delivering. You push through and finish around 11, crash, and then wake up the next day to find it a struggle to do even the things you love: it feels like you've burned all your willpower and you've become a husk of a real person.*
>
> *Fast-forward to the weekend. You've finally recovered from writing The Report and you have two whole days for things you love. "This," you say to yourself, "is the weekend I finally make some real progress on learning to play the bouzouki." You get our your Bouzouki for Beginners book, tune your bouzouki, and play for about 10 minutes before you remember you had to do that other thing. That other thing is very important, so you put down the bouzouki to go knock out the other thing so you can get back to Bouzouki Weekend 2018. But while doing the other thing you remember you haven't checked Facebook in a while, and you don't want your friends to think you forgot about them, so you do that for a bit. Then you start cleaning, notice the bouzouki is there, and clean around it so you can come back to it when you're done. By now you've worked up an appetite, so you make lunch. You would get right back to the bouzouki after eating but it would be so nice to take a nap, so you do that. You're awoken from the nap by a call from your mother, so you talk to her for a while because you like catching up with family. After you get off the phone you realize it's getting late so you better get to some bouzouki playing, but you have to get ready to go out tonight with your friends because that's going to be fun! Oh well, Bouzouki Weekend 2018 is only half-over, there's still tomorrow. "I'll learn to play the bouzouki one day," you say to yourself.*

In the first part of the vignette we might say our protagonist procrastinated because they put off writing the report now to write it later, but they don't seem to be suffering from the sort of laziness we typically associate with procrastination. After all, they tried to do the work, they just weren't able to make themselves do it. In the second part they got distracted a lot, but all the distractions were things they honestly also wanted to do, so that's not exactly procrastination or laziness either. Yet in both cases there was an activity they clearly wanted to do that didn't get done, or only got done by burning through willpower and feeling exhausted afterwords, unable to do much else. What's going on?

Many people I know might describe this as a case of akrasia, an ancient Greek word literally meaning "no strength/power" but used to mean a lack of willpower or having a weakness of will (c.f. aboulia for a related but more general phenomenon that lacks the cognitive dissonance associated with akrasia). A more straight-forward explanation of akrasia is that it's the thing that going on when you do something other than what you want to do. Akrasia is the thing that's going on when you want to write

a report, play bouzouki, or do some other particular activity and instead find yourself unable to make yourself do it while you do something else instead.

To get more formal we might describe this as a conflict in your values, wants, desires, motivations, and preferences (a cluster of concepts I unify under the term "axias") due to them being in irrational relationships, "irrational" here meaning specifically not rational in a formal sense. We could formally describe akrasia then in terms of how it fails to satisfy the rationality criteria, in particular how it fails to satisfy the criterion of asymmetry, where asymmetry means that given any axias A and B, if you prefer A to B, then you do not prefer B to A. Akrasia then seems to be what happens when asymmetry goes unsatisfied.

Consider what was happening when our protagonist wasn't playing bouzouki. They wanted to play bouzouki, yet after a very brief attempt they didn't play bouzouki. So they claim they want to play bouzouki more than not play it, yet we observe them not playing bouzouki instead of playing it, thus we have conflicting lines of evidence about the relationship between these two preferences. On the one hand this might be revealing a conflict between stated and revealed preferences, i.e. a difference between what they say they want and what they do, but on the other this might reflect an actual failure of asymmetry. Deciding which isn't important though, because although conflicting preferences are part of the story of akrasia, they aren't the whole story, and the trick to understanding and ultimately dissolving akrasia is seeing how it arises in a context that looks beyond conflicting preferences, because as I'll shortly explain, akrasia is a problem of how we relate to our preferences, not our preferences themselves.

Whence preference conflict

If we're going to dissolve akrasia, it'll help to have a firmer grasp on it's etiology so we understand just where it's coming from. Knowing that it arises with an apparent failure of preference asymmetry isn't enough to really see what's going on. For that we have to delve into some fake models.

I don't of course mean here that these models aren't useful or don't reveal something about reality, merely that they are "fake" the same way all models are "fake": they compress reality in a way that helps us understand it, but do so at the cost of accurately describing reality just as it is. As this blog's title reminds us, the map is not the territory. This will turn out to be important, though, because akrasia is, in my estimation, entirely the result of models causing us to be confused about what is.

Most people I know who have expressed feeling akrasia also primarily use one or more dual-process models to help them understand the psyche/mind. Dual-process models suggest the psyche is made up of two parts, those partsroughly being the S1-elephant-id-unconscious-near part and the S2-rider-superego-conscious-far part. For book-length explorations of dual-process theory see Kahnemann's *Thinking Fast and Slow* and Hanson & Simler's*Elephant and the Brain*.

Within dual-process theory, we should expect preference conflicts, especially failures of asymmetry, to arise when the S1 part of the mind wants something different than the S2 part. Since S1 is much more powerful than S2, even if S2 is "smarter", S1 will usually win unless S2 expends a lot of energy to rein in S1 and make it do what it wants. On this model it then seems that akrasia is just preference asymmetry, albeit asymmetry caused by conflict between parts, and fighting it should consist mainly of finding ways to get S1 aligned with S2 (since obviously S2, being smarter, knows what

is best for S1). Our best bet for defeating akrasia then should be something like Beeminder or Complice, and we probably can't hope to do much better.

I hope that previous paragraph threw out some red flags for you, but in case it didn't here's where I see cracks beginning to form in dual-process theory's explanation of akrasia. People with akrasia tend to identify with S2 as the real, true self. Maybe not identify with it as strongly as one might identify with the virtual homunculus in the Cartesian theater if you're a mind-body dualist, but identify with it enough to privilege it over S1 such that in a conflict between the two they would prefer, all else equal, that S2 win. And this seems reasonable, even from S1's perspective, because we get lots of signals from other people telling us that what's best for us is that which is associated with S2, far construal level, and the superego.

But this identification with or privileging of S2 is suspect because there's nothing about S2 to suggest it's the "real" self. If there is anything worthy of calling one's self*, it's made up of both S1 and S2 (and maybe more things besides!). S2 is special, but S1 is equally special, and each brings its own powers to the table. If it were otherwise, you wouldn't have evolved to have a brain so complex and capable that you'd usefully be able to use dual-process theory to make sense of it. To put my thumb on it, S1 is not a spandrel getting in your way; it's a useful part of you helping you be you and live your life!

So if that's the case, what the heck is going on that a person would interpret through a dual-process model as akrasia? Well, it's just what we've already seen: your mind is complicated, you may have different parts of it producing different desires, and then the whole thing puts those together and makes a choice about what to do. In the end, you only "want" one thing (your revealed preference) even if you had to weight many things to come to it (your stated preferences) and your confidence in your synthesis of your desires is low.

Akrasia, then, is a kind of suffering that arises from *identifying with particular desires* in spite of having already given them their fair weighting in coming to a choice of action. It can exist with or without a belief in the usefulness of dual-process theory; that was just a way to draw the way we identify with desires into relief. The key thing is that we experience akrasia because of identifying with our desires, and this also suggests it's "easy" to stamp it out: just stop doing that!

**N.B. I'm a Buddhist and a phenomenologist, so my relationship with self is complicated. So just to lay all my cards on the table, although I consider them not entirely relevant here, I think there is no-self, and also that there is some small thing we might give the name "self" that refers to the irreducible subject of experience.*

Grappling with identity

Okay, great, so akrasia isn't really real—it's an artifact of the way we understand ourselves and identify with that understanding, and it will evaporate if we can stop doing that and get back to reality itself. That's not exactly advice about what to do, even if the first step of the journey is often just knowing you could go somewhere else, and it feels a bit unfair of me to dissolve akrasia in theory but not help you dissolve it in practice, so let me give you an exercise that might help set you on your way if you'd like to have not just episteme of the true nature of akrasia, but gnosis.

Content Warning: The rest of this section asks you to work through a process that might best be described as self-applied psychotherapy. If you are or are at risk of being suicidal, psychotic, or otherwise in need of mental healthcare, I ADVISE YOU TO SKIP THE REMAINDER OF THIS SECTION. Self-help techniques can be powerful and transformative, thus they are never 100% safe, and so should be used only under supervision if you are not sufficiently mentally healthy and resilient to handle whatever shadows these questions might bring up. That said, this line of questioning might help you get a better handle on existing akrasia: it did for me.

Identify a recent event where you experienced akrasia. Maybe you were writing a report or playing the bouzouki as in the story that I opened with, maybe it was something more nebulous like wishing you had lived up to some virtue through your actions. If you come up with something more nebulous, try to make it more concrete first by identifying a particular action through which you expressed akrasia. Same goes if you initially identified something more goal-oriented. Have in mind something less like "I wanted to play the bouzouki but didn't" and more like "I wanted to play the bouzouki on Tuesday night but watched reruns of *I Love Lucy* instead".

Now ask yourself what your akratic actions imply to you about who you are? How do you feel when you think about your akrasia experience? Where in your body do you feel it? In your gut? In your chest? In your head? Behind your eyes? What is the story you tell yourself when we see that you wanted to do one thing and did another?

Why does that worry you? Rest on this question for a moment and see what comes up. You are thinking about this and suffering via akrasia, so try to put a name to that worry. Are you grasping, clinging to, or trying to achieve something and worried you're not doing it? Are you worried about what you are doing? Try to give it a name.

Now try to imagine what would happen if the things you are worried about happened. What would happen to you? How would that change who you are? Would you be a different person?

Finally, take your answers to those questions and ask, what makes you so sure? Why do you think that would happen, you would change in that way, or you would be different (or not)? How can you test those beliefs? If you can, find a way to safely carry out one or more of those tests and see what you learn. You might be surprised what your learn about your relationship to your own identity.

Repeat going through this process as often as you like. The point of it is to help you to deconstruct your hidden assumptions about the relationship between your actions and your identity. Without prejudicing your insights too much, I suspect you will find there is less you than you thought, the you that is there is less permanent than you thought, and the suffering of akrasia is being created by trying to hold on to some idea of yourself that was at best only a dream.

What if I still don't do what I ought?

Even after completely dissolving akrasia we might still find we want things that are in conflict. This is normal, because humans are irrational both in the formal and folk senses of that word. We might still fail to have asymmetric preferences and find ourselves doing things we'd in some sense prefer not to do. So be it. That doesn't have to be akrasia, though, so long as we don't start identifying with our contradictory desires. To return to the opening story again, wanting to play the bouzouki and never

doing it doesn't have to mean you experience akrasia. So long as you both accept that you both want to play the bouzouki and don't want to play the bouzouki enough to do it instead of something else, you won't suffer from akrasia because you are sure you are doing just what you want.

Further, I'm not the first to suppose akrasia isn't real. In fact, it goes right back to the first known appearance of the term with Plato's Socrates saying as much. And some experiences of akrasia may be misunderstandings of desires that have more to do with appropriate lack of motivation to do somethingthan conflicting desires or identifying with them or may be misconstrued procrastination. So if after reading all I've said above you find you still have some lingering sense that akrasia is real, check what others have had to say.

Finally, what I've described above may feel like "nothing more" than a subtle shift in perspective, but it's an important one if you want to learn to accept yourself as you are, something I view as foundational to better making progress on whatever it is you care about. So long as you are deluded, thinking you want to do one thing when in fact you want to do another, you'll always struggle to change your behavior (if that's what you want to do!) because you're acting based on a confusion. Akrasia is just one way this confusion powerfully manifest itself, and learning to sublimate it is an import step along whatever path you take.

Originally posted on Map and Territory of Medium.

Let Values Drift

I occasionally run across lines of reasoning that depend on or favor the position that value drift should be avoided.

I find odd the idea of value drift, let alone the idea that value drift is bad. My intuition is that value drift is good if anything since it represents an update of one's values based on new evidence and greater time to compute reflective equilibrium. But rather than arguing intuition, let's explore value drift a bit before we come to any stronger conclusions.

(Fair warning, this is going to get into some deep philosophical territory, be pretty unapologetic about it, and assume you are reading carefully enough to notice what I say and not what you think I said. I'm still working some of these ideas out myself, so I don't have the fluency to provide a more accessible explanation right now. I also take some pretty big inferential jumps at times that you may not be on board with as of yet, so the later parts might feel like unjustified reasoning. I don't think that's the case, but you'll have to poke at me to help me figure out how to fill in those gaps.

In spite of all those apologies, there are some key insights here, and I'm unlikely to get clearer unless I am first more opaque, so please bear with me if you please, especially if you are interested in value as it relates to AI alignment.)

Whence drifting values?

The metaphor of drifting values is that your values are initially one place and then gradually relocate to another, like flotsam. The waves of fortune, chance, and intention combine to determine where they end up on the seas of change. In this metaphor, values are discrete, identifiable things. Linguistically, they are nouns.

When we talk of values as nouns, we are talking about the values that people have, express, find, embrace, and so on. For example, a person might say that altruism is one of their values. But what would it mean to "have" altruism as a value or for it to be one of one's values? What is the thing possessed or of one in this case? Can you grab altruism and hold onto it, or find it in the mind cleanly separated from other thoughts? As best I can tell, no, unless contrary to evidence and parsimony something like Platonic idealism proves consistent with reality, so it seems a type error to say you possess altruism or any other value since values are not things but habituations or patterns of action (more on this in the next section). It's only because we use the metaphor of possession to mean something like habitual valuing that it can seem as if these patterns over our actions are things in their own right.

So what, you may think, it's just a linguistic convention and doesn't change what's really going on. That's both wrong and right. Yes, it's a linguistic convention and yes you get on with valuing all the same no matter how you talk about it, but linguistic conventions shape our thoughts and limit our ability to express ourselves with the frames they provide. In the worst case, as I suspect is often happening when people reckon about value drift, we can focus so much on the convention that we forget what's really going on and reason only about the abstraction, viz. mistake the map for the territory. And since we've just seen that the value-as-thing abstraction is leaky because it implies the ability to possess that which cannot be, it can lead us astray by

allowing us to operate from a false assumption about how the world works, expecting it to function one way when it actually operates another.

To my listening most talk about value drift is at least partially if not wholly confused by this mistaking of values for things, and mistaking them specifically for essences. But let's suppose you don't make this mistake; is value drift still sensible?

I think we can rehabilitate it, but to do that we'll need a clearer understanding of "habitual valuing" and "patterns of action".

Valuing valuing

If we tear away the idea that we might possess values, we are left with the act of valuing, and to value something is ultimately to judge it or assess its worth. While I can't hope to fit all my philosophy into this paragraph, I consider valuing, judging, or assessing to be one of the fundamental operations of "conscious" things, it being the key input that powers the feedback loops that differentiate the "living" from the "dead". For historical reasons we might call this feeling or sensation, and if you like control theory "sensing" seems appropriate since in a control system it is the sensor that determines and sends the signal to the controller after it senses the system. Promising modern theories suggest control theory is useful for modeling the human mind as a hierarchy of control systems that minimize prediction error while also maintaining homeostasis, and this matches with one of the most detailed and longest used theories of human psychology, so I feel justified in saying that the key, primitive action happening when we value something is that we sense or judge it to be good, neutral, or bad (or, if you prefer, more, same, or less).

We could get hung up on good, neutral, and bad, but let's just understand them for now as relative terms in the sense of the brain as control system, where "good" signals better prediction or otherwise moving towards a set point and "bad" signals worse prediction or moving away from a set point. Then in this model we could say that **to value something is to sense it and send a signal out to the rest of the brain that it is good**. Thus to "have a value" is to observe a pattern of action that senses that pattern to be good. To return to the example of valuing altruism, when a person who values altruism acts in a way that pattern matches to altruism (maybe "benefits others" or something similar), the brain senses this pattern to be good and feeds that signal back into itself further habituating actions that match the altruism pattern. It is this habituation that we are pointing to when we say we "have" a value.

Aside: How any individual comes to sense any particular pattern, like altruism, to be good, neutral, or bad is an interesting topic in and of itself, but we don't need that particular gear to continue discussing value drift, so this is where the model bottoms out for this post.

We can now understand value drift to mean changes in habituations or patterns of action over time. I realize some of my readers will throw their hands up at this point and say "why did we have to go through all that just to get back to where we started?!?", but the point was to unpack value drift so we can understand it as it is, not as we think it is. And as will become clear in the following analysis, that unpacking is key to understanding why value drift seems an odd thing to worry about to me.

Values adrift

My explanation of valuing implies that values-as-things are after-the-fact reifications drawn from the observation of accumulated effects of individual actions, and as such values cannot themselves directly drift because they are downstream of where change happens. The changes that will befall these reifications that we call "values" happen moment to moment, action to action, where each particular action taken will only later be aggregated to form a pattern that can be expressed as a value, and even then that value exists only by virtue of ontology because it is an inference from observation. Thus when values "drift" it's about as meaningful as saying the drawings of continents "drift" over geological time: it's sort of true, but only meaningful so long as understanding remains firmly grounded in the phenomena being pointed to, and unlike with maps of geography maps of mind are more easily confused for mind itself.

What instead drifts or changes are actions, although saying they drift or change is wrought because it supposes some stable viewpoint from which to observe the change, yet actions, via the preferences that cause us to choose any particular action over all others, are continuously dependent on the conditions in which they arise because what we sense (value, judge, assess) is conditional on the entire context in which we do the sensing. So it is only outside the moment, whether before or after, that we judge change, and so change is also ontologically bound such that we can find no change if we look without ontology. In this sense change and drift in actions and patterns of action exist but are not real: they are in the map, but not the base territory.

Does that matter? I think it does, because we can be confused about ontology, confusion can only arise via ontology, and sensing/valuing is very near the root of ontology generation, so our understanding of what it means to value is mostly contaminated by valuing itself! Certainly by the time we put words to our thoughts we have already sensed and passed judgement on many phenomena, and that means that when we talk about value drift we're talking from a motivated stance where valuation heavily shaped our perspectives, so I find it not at all odd that valuing would find a way to make itself and its products stable points within concept space such that it would feel natural to worry that they might drift, and that drifting and change in values would evaporate without sensing feedback loops to prop them up!

This is not to anthropomorphize valuing, but to point out the way it is prior to and self-incentivized to magnify its existence; it's like a subagent carrying out its own goals regardless of yours, and it's so good at it that it's shaped your goals before you even knew you had them. And when we strip away everything posterior to valuing we find no mechanism by which value can change because we can't even conceptualize change at that point, so we are left with valuing as a pure, momentary act that cannot drift or change because it has no frame to drift or change within. So when I say value drift is odd to me this is what I mean: it's exists as a function of valuing, not of valuing itself, and we can find no place where value change occurs that is not tainted by the evaluations of sensing.

(Careful readers will note this is analogous to the epistemological problem that necessitates a leap of faith when knowledge is understood ontologically.)

Yikes! So what do we do?

Steady on

The questions that motivate this investigation are ones like "how do we protect effective altruists (EAs) from value drift so that they remain altruistic later in life and don't revert to the mean?" and "how do we align superintelligent AI with human values such that they stay aligned with human values even as they think longer and more deeply than any human could?". Even if I lost you in the previous section—and I'm a little bit lost in my own reasoning if I'm totally honest—how can we cash out all this philosophy into information relevant to these questions?

In the case of drifting EAs, I say let them drift. They value EA because conditions in their lives caused them to value it, and if those conditions change so be it. Most people lack the agency to stay firm in the face of changing conditions, I think this is mostly a safety mechanism to protect them from overcommitting when they aren't epistemically mature enough to know what they're doing, and for every EA lost to this there will likely be another EA gained, so we don't have to worry about it much other than to deal with churn effects on the least committed members of the movement. To do otherwise is to be inconsistent on respecting meta-preferences, assuming you think we should respect people's meta-preferences, in this case specifically the meta-preference for autonomy of beliefs and actions. Just like you would probably find it troubling to find racists or fascists or some other outgroup working on incentives to keep people racist or fascist in the face of evidence that they should change, you should find it troubling that we would seek to manipulate incentives such that people are more likely to continue to hold EA beliefs in the face of contrary evidence.

Most of this argument is aside my main point that value drift is a subtly motivated framing to keep values stable propagated by the very feedback processes that use sense signals as input with no prior manifestation to fall back on, but you might be able to see the deep veins of it running through. More relevant to this question directly are probably things like "Yes Requires the Possibility of No", "Fundamental Value Differences are not that Fundamental", "Archipelago", and much about meta-consistency in ethics that's not salient to me at this time.

On the question of AI alignment, this suggests concerns about value drift are at least partially about confusion on values and partially fear born of a desire for value self-preservation. That is, a preference to avoid value drift in superintelligent AIs may not be a principled stance, or may be principled but grounded in fear of change and nothing more. This is not to say we humans would be happy with any sense experiences, only that we are biased and anchored on our current sensing (valuing) when we think about how we might sense things other than we do now under other conditions. I realize this makes the alignment problem harder if you were hoping to train against current human values and then stick near them, and maybe that's still a good plan because although it's conservative and risks astronomical waste by denying us access to full optimization of valuing, that's probably better than attempting and failing at a more direct approach that is less wasteful but maybe also ends up tiling the universe with smiley faces. My concern is that if we take the more conservative approach, we might fail anyway because the value abstraction is leaky and we end up building agents that optimize for the wrong things, leaving gaps through which x-risks develop anyway.

(Unless it wasn't clear, AI alignment is hard.)

If any of that left you more confused than when you started reading this, then good, mission accomplished. I continue to be confused about values myself, and this is part of a program of trying to see through them and become deconfused on them, similar to the way I had to deconfuse myself on morality many years ago. Unfortunately not many people are deconfused on values (relatively more are deconfused on morals) so not much is written to guide me along. Look for the next post whenever I get more deconfused enough to have more to say.

Cross-posted to Map and Territory on Medium

Scope Insensitivity Judo

It's easy to bemoan scope insensitivity, a special case of that phenomenon where we mere humans end up caring more about the death of one person than one hundred, better remember the last bite of a meal than the first dozen, and think less is more and more is less. After all, if we didn't neglect scope we would be more rational, and so maybe happier and healthier, living in a world were everyone got more of what they wanted, since without scope insensitivity it wouldn't be so hard to convince people to help those far away who need more than those nearby who need less. But scope insensitivity is what we've got, so we have to learn to live with it.

Luckily there's plenty of reason to think we can take advantage of scope insensitive because people have already discovered ways to make the best of other forms of extension neglect. For example, adapting to duration neglect is something most people learn early on by adopting heuristics like "save the best for last" and "do the hardest part first". Salespersons and motivational speakers alike learn to exploit base rate neglect, sample size neglect, and the conjunction fallacy to convince people to do what they otherwise might not. And designers of all kinds of systems can mold incentives to work with rather than against human nature. Thus it stands to reason we can use our natural scope insensitivity to do more than fail to multiply.

I'll consider one such use case here, namely a practice of using scope insensitivity to prepare ourselves for high-stakes situations in low-stakes ones. This is a kind of scope insensitivity judo, or "gentle way", and just like in the martial art, we'll redirect the strength of our "opponent" to transform it into an unintended ally.

Dwell in the Dojo

Life is full of high-stakes situations: job interviews, first dates, nuclear missile crises. These generally feel like one-shot scenarios—there's one chance to get it right and if we fail all is lost. To wit, if we don't say the right things we'll lose our shot at that job forever, if we don't put out the right vibes that person will never fall in love with us, and if we push the big red button there'll be no second chances for anything.

We can make these multi-shot scenarios pretty easily with training, and there's some value in practicing interviewing skills, going on many dates so no one date matters very much, and running war games. These are all training methods that take something high-stakes and make it low-stakes so you feel free to experiment. That's one way to learn: by creating a safe laboratory where we can explore more before we prune.

That's not what I'm suggesting we do, though. In the dojo of scope insensitivity judo, we practice the way of getting into low-stakes scenarios that feel high-stakes so we are prepared generally to handle really high-stakes scenarios when we encounter them. We do this by taking advantage of the way our minds mistakenly believe many low-stakes scenarios are high-stakes ones because they push against beliefs and behaviors that were evolutionarily or historically adaptive but no longer are.

Consider these examples from my own life, drawn from my zen practice:

- I asked if I could bring a cushion from home for a retreat. I was told yes. I brought it. The cushion was orange, the zendo's cushions were black, it stuck out, and I was told I couldn't use my cushion.
    - I complied, but I was immediately caught by thoughts like "but you told me I could use my cushion" and "now my meditation will be worse because I'll be less comfortable" and "I'm not as good a zen student as I thought".
    - I felt embarrassed, defensive, let down, and defeated. I felt like a failure, like I was 2nd grade Gordy again getting in trouble for being weird.
    - Of course, stepping back, we can see this was a very low-stakes situation: I just switched cushions and got on with the retreat! But it felt high-stakes at the time because it pushed me in ways that might have been adaptive in some high-stakes situations, either in my personal past or within my cultural or evolutionary environment. For our ancestors, this kind of mistake could have meant loss of prestige and thus loss of resources and thus marginal loss of reproductive and survival opportunities. Lucky for me it was just about a cushion in the zendo!
- I was sitting half lotus during a long meditation period, and after about 40 minutes my legs hurt in a way that I worried was injuring them by continuing, so I uncrossed my legs and sat with them pulled up towards my chest to give them a rest. In the middle of this my teacher walked into the zendo and saw me, and came over to correct me, saying I couldn't sit like that and had to either sit crosslegged or in a chair.
    - I sort of complied: I instead took the option to do brief walking meditation before returning to sitting. I was caught by thoughts like "you didn't see how I had been sitting" and "you didn't know the kind of danger I was in" and "I must have stayed sitting out-of-form because the pain was so bad it temporarily addled my mind".
    - I felt embarrassed, ashamed, and defensive and also a little indignant.
    - This was also a pretty low-stakes situation: I walked for 10 minutes, came back and sat for the rest of the period, and it was never mentioned again. I didn't lose any of my positions or responsibilities, and my practice continued on as strong as ever. But it felt high-stakes because I had been caught out and corrected in front of others, and maybe they would think less of me. As best I can tell, they did not.
- A new person came to our Saturday morning practice period for the first time. I was work leader that week, and when it came time to hand out assignments I assigned her to clean the zendo under my supervision. I was later corrected by the person who trained me as work leader that I shouldn't have given her that assignment because new people should get simple tasks like sweeping.
    - I was immediately somewhat defensive. Cleaning the zendo was the job I had been assigned when I first came to the zen center, so I thought it was the right thing to do. I said as much.
    - In addition to being defensive, I felt like I had been let down by my trainer not telling me this before, and I also felt I had the excuse that it worked out fine.
    - Once again, this was pretty low-stakes: she cleaned the zendo well, I got new information, and I changed how I hand out work assignments. But my behavior indicates I thought it was high-stakes enough to be worth some back-and-forth and argument or defense of my position and to put myself in opposition to another person to save face. I had wanted to do a good job at being work leader, and felt threatened by the correction, leading me to escalate my response.

I drew these from my zen practice because the zen center is like a laboratory where we specialize in studying the self, and so I had more chance to examine these events and remember them than the many similar daily occurrences that happen throughout the rest of my life. Also they are less personal and raw than the times I blew low-stakes situations out-of-proportion and didn't learn from them at work, with family, and among friends. But hopefully those are enough for you to start to see the pattern: scope insensitivity means we often treat low-stakes situations like high-stakes situations, and we can take advantage of that to use them as training scenarios for genuine high-stakes events if we allow ourselves the space to stop and take a step back to consider what we're doing.

The Way of Scope Insensitivity

You can begin to practice with scope insensitivity yourself right away, because the world is constantly presenting you with low-stakes scenarios that feel high-stakes. The more anxious, depressed, or frustrated you generally are, the more you are likely you are treating low-stakes situations as high-stakes and so you will have even more opportunities to practice scope insensitivity judo than people who are more calm and equanimous.

The first part of the practice is to notice and stop. Notice when you feel like you are in a high-stakes situation. Then stop for a few breaths to examine it. Don't worry if you fail at first; learning to notice is hard if you're not already skilled at it, and even when you are skilled it's still easy to get so caught up that we forget to really look.

When you catch one of these situations, consider whether it is really high-stakes, or if you just believe it is due to scope insensitivity. If it's really low-stakes, this is a great opportunity to experiment and practice with dealing with these situations and the factors that cause them to feel high-stakes. If you're sure it's really high-stakes, that's even better, though you'll want to be a bit more cautious in how you proceed.

There are many ways you can explore these situations once you've noticed them arising, and the path you take largely depends on what you are ready for and what resonates with you. I've gotten a lot of milage out of the Immunity to Change framework and working with core beliefs (albeit within the Ordinary Mind zen context rather than a CBT context). You might prefer something that looks more like psychotherapy, various CFAR techniques, Folding, Focusing, Core Transformation, or some kind of debugging. Generally you are looking for a way to integrate what you can see when you step back and look at what's happening in these situations with your immediate reactions, and anything that helps you do that will likely work here.

And then you just keep doing it. You're unlikely to fix your scope insensitivity—that appears to just be part of how human brains work. But you can, through regular practice, retrain yourself to more deftly handle situations that previously felt overwhelming. By developing the skill of flipping what feel like high-stakes situations into low-stakes ones, you'll gain perspective on those situations that allows you to take a more thoughtful, deliberate approach that transcends the worst of our knee-jerk reactions that lead to self-created suffering.

Cross-posted to Map and Territory.

Normalization of Deviance

An important, ongoing part of the rationalist project is to build richer mental models for understanding the world. To that end I'd like to briefly share part of my model of the world that seems to be outside the rationalist cannon in an explicit way, but which I think is known well to most, and talk a bit about how I think it is relevant to you, dear reader. Its name is "normalization of deviance".

If you've worked a job, attended school, driven a car, or even just grew up with a guardian, you've most likely experienced normalization of deviance. It happens when your boss tells you to do one thing but all your coworkers do something else and your boss expects you to do the same as them. It happens when the teacher gives you a deadline but lets everyone turn in the assignment late. It happens when you have to speed to keep up with traffic to avoid causing an accident. And it happens when parents lay down rules but routinely allow exceptions such that the rules might as well not even exist.

It took a much less mundane situation for the idea to crystalize and get a name. Diane Vaughan coined the term as part of her research into the causes of the Challenger explosion, where she described normalization of deviance as what happens when people within an organization become so used to deviant behavior that they don't see the deviance, even if that deviance is actively working against an important goal (in the case of Challenger, safety). From her work the idea has spread to considerations in healthcare, aeronautics, security, and, where I learned about it, software engineering. Along the way the idea has generalized from being specifically about organizations, violations of standard operating procedures, and safety to any situation where norms are so regularly violated that they are replaced by the de facto norms of the violations.

I think normalization of deviance shows up all over the place and is likely quietly happening in your life right now just outside where you are bothering to look. Here's some ways I think this might be relevant to you, and I encourage you to mention more in the comments:

- If you are trying to establish a new habit, regular violations of the intended habit may result in a deviant, skewed version of the habit being adopted.
- If you are trying to live up to an ideal (truth telling, vegetarianism, charitable giving, etc.), regularly tolerating violations of that ideal draws you away from it in a sneaky, subtle way that you may still claim to be upholding the ideal when in fact you are not and not even really trying to.
- If you are trying to establish norms in a community, regularly allowing norm violations will result in different norms than those you intended being adopted.

Those mentioned, my purpose in this post is to be informative, but I know that some of you will read this and make the short leap to treating it as advice that you should aim to allow less normalization of deviance, perhaps by being more scrupulous or less forgiving. Maybe, but before you jump to that, I encourage you to remember the adage about reversing all advice. Sometimes normalized "deviance" isn't so much deviance as an illegible norm that is serving an important purpose and "fixing" it will actually break things or otherwise make things worse. And not all deviance is *normalized* deviance: if you don't leave yourself enough slack you'll likely fail from

trying too hard. So I encourage you to know about normalization of deviance, to notice it, and be deliberate about how you choose to respond to it.

You are Dissociating (probably)

You are probably dissociating all the time and don't realize it. Let's look at just what I mean by this bold claim.

Disclaimer: *I'm not a medical professional, and although I'm talking about "dissociation" here I definitely don't mean to give medical explanation or advice about clinical dissociation. I'm going to talk about the same mental phenomena that cause clinical dissociation disorders, but focus on situations and people for whom it doesn't rise to the level of disorder. If you are seeing or think you need to see a medical professional about dissociation, maybe don't read this if you think reading about dissociation might make it worse and don't disregard medical advice in favor of anything I say.*

Much of my claim that most people are dissociating frequently depends heavily on just what I mean by dissociation, so let's start by trying to nail down just what I mean by this term.

Clinically dissociation is "a disconnection between a person's thoughts, memories, feelings, actions or sense of who he or she is." That's general enough to cover all phenomena psychiatric professionals would like included as "dissociation" but doesn't cleanly distinguish dissociation from other conditions like schizophrenia and psychosis on its own. Helpfully there are a few specific disorders that count as dissociation: depersonalization disorder, derealization disorder, and dissociative identity disorder. These disorders point towards concrete experiences people have that will help us explore dissociation.

NB*: Each disorder includes phenomena that, on their own, are not necessarily indicative of disorder. That is, it's totally normal to experience depersonalization, derealization, or some level of identity dissociation without it being disruptive enough to be a disorder. My understanding is that standard criterion to test if is something a psychiatric disorder is to check if it significantly interferes with a person's ability to live their life as they would like to.*

Depersonalization is probably the kind of dissociation most people have clear memories of having experienced: it's a kind of disidentification with the self. Some people describe it as being out of their body; I might describe it as being trapped in my head, failing to identify with the body in which the mind sits like a homunculus. The mind and body seem detached, or the body seems unreal or fake, or in some cases even parts of the mind might seem to be not the real self. Something like looking in the mirror and feeling like the person staring back isn't really you, or seeing your thoughts pass by and not identifying as being the one who thought them.

Derealization is basically the same thing as depersonalization, but about the environment beyond the body. A kind of feeling like the things and people around you are fake or that you don't inhabit the same plane of existence as they do. I'd say it's sort of like feeling you're walking around inside a video game or movie of your life and may include a feeling of lack of agency.

Identity dissociation is a bit different, as it involves a person having more than one identity or "personality state" (cf. ego states). Although in the disorder case these identities may be so strong as to cause amnesia between them when one is actively being identified with, the more common experience is something like becoming "a

different person" in different scenarios, like maybe you're a party animal Friday nights and a reserved church goer Sunday mornings. Somewhere in between lie things like tulpas and IFS and feeling like your mind is made up of multiple agents.

What these experiences have in common is noticing some kind of split related to identity. In fact, "dissociate" literally means to come apart from being joined (the same etymology holds for the synonym "disassociate"). We've talked a bit about what results from the coming apart, but what was joined in the first place that could be split?

The idea in each case is that identity is somehow experienced as not unified with the rest of reality. With depersonalization this is identity splitting from unification with the mind or body such that thoughts and actions aren't recognized as belonging to the self. With derealization it's identity splitting from surroundings so the self no longer feels it's part of the world. And with identity dissociation it's identity seen as made up of multiple parts that either take turns being the "real" self or collaborate to form the self.

So if I try generalize what's going on here, I'd describe it as **splitting a joined or unified experience of reality into parts.**

What I find fascinating about dissociation is that we categorize it as an abnormal state, or at least notice it because it surprises us (causes us to perceive the world in a way we didn't predict), and inherent in the concept of dissociation is the idea that there's an alternative, unified state, and that unified state is the natural state that can be split apart rather than the reverse. Yet we know the brain is made up of lots of different parts that communicate with each other, and if we look really hard we can't find anything that is the irreducible center of identity. Further, we know that, however it is that attention/consciousness works, one of its features is that we only pay attention to one thing at a time, but can rapidly switch between objects of attention such that it can seem like we're paying attention to a lot of stuff at once. This would seem to suggest that dissociation should be the normal, expected state of simply noticing how the brain works, yet it feels unexpected when experienced nonetheless. What gives?

I think what's going on is a clash between how we conceptualize identity and how our minds form. Specifically, identity is an after-the-fact model we have of ourselves that we keep updating to match what actually happened, and our capable but limited powers for self-modeling lack enough structure to stuff in all the complexity we can observe into the model, thus putting us in a place of continually be surprised that our inadequate but best effort model fails to predict what we observe about ourselves. When this becomes the focus of attention, we label that feeling dissociation, but in fact we are dissociated from the reality of our existence all the time because our maps of ourselves are not exactly the territory.

So should we do anything about all this dissociation?

Not necessarily because, so long as dissociation isn't so disruptive to living life that it becomes a clinical diagnosis, it's just recognition of the basic fact that our models of ourselves involve abstractions that aren't perfectly predictive of our experiences.

That said, I think there are a variety of sub-clinical sources of suffering in life caused by excess dissociation. In particular, I'm thinking of suffering of the sort caused by overly identifying with how you model yourself, like the pain of not being as smart, successful, attractive, etc. as you think you are or should be or the failure to fully

enjoy life because you're too focused on how you much you think you'll like or dislike something rather than how much you actually like or dislike it when you experience it. But there's also subtler forms of suffering caused by dissociation, like the existential dread that can come from noticing the modeled self is not as permanent as you'd hoped.

If you want to do something about it, there are many possible practices that can help. Meditation is one that will help you by studying the self. Another is Focusing. Another, more social version might be various authentic relating practices like Circling. All of these work by showing you where your model and reality don't quite match up and offer you the opportunity to learn more about yourself and become less surprised at what it is like to be you.

***NB**: None of them are a substitute for psychiatric care, though, so if reading this leads you to suspect you are suffering from a dissociative disorder, I encourage you to consult with a mental health professional.*

Forcing yourself to keep your identity small is self-harm

The title lays out my thesis in perhaps the maximally provocative way: if you put direct effort into keeping your identity small (à la Graham), you are doing harm to yourself akin to the kind of harm you do to yourself by suppressing emotions.

I was inspired to write this after reading some of the recent discussion on the EA Forum about why people do or do not identity as an EA. Responses that mention not wanting to identity as EA in the name of keeping one's identity small got me thinking about this topic and made me realize that, despite likely being somewhat less attached to various identities than average, I don't eschew being identified with things or using identities as a means of communicating with others. And I immediately had a strong sense of why I don't do that: taking action to directly avoid having a large identity would be trying to control myself in a way that leads to cognitive dissonance and dissociation at best and cognitive fusion at worst.

Here's what I see going on. A person realizes they are identifying as or with something, let's say effective altruism since that was the motivating example. They have a commitment to keeping their identity small, so give themselves feedback that they are doing something wrong when they identity as an effective altruist and should stop doing it. Negative reinforcement like this, though, rarely makes the underlying reason for wanting to do something go away, like perhaps wanting to signal that you are the kind of person who cares about everyone equally and thinks rationally about things. Instead it creates a secondary desire to not do the thing that lives alongside the desire to do it. So now you both want to identity as an EA and don't want to, and whether or not you do it depends on which desire is stronger.

A good model of what happens next is offered by Internal Family Systems: competing parts of the brain are now engaged in a battle of trying to protect you from what other parts of your brain want you to do. Aside from causing a lot of suffering as a result of being the battleground on which this fight occurs, it can eventually lead to isolation (exile) of those parts of the self that seek to identify with something. They're still there trying to do their thing, of course, but they are so well suppressed that they become hidden from self-awareness.

I see this as analogous to the effects of suppressing emotions in order to control them. Maybe you can do it, but rather than true mastery of emotions it causes suppression of the ability to see what is happening. The result is often that rather than getting, say, scared, you instead think the world is out to get you as a result of cognitive fusion that happened as a result of suppression. The path back out is through things like focusing and studying the self to forget the self.

None of this is to say I disagree with keeping your identity small as a useful direction, but it's more the kind of thing to check in on occasionally to see if it's a consequence of other things you're doing rather than something you can go directly towards, since the direct path, as argued above, makes things globally worse even as you seemingly make local improvements. Small identity is something that best happens as a consequence of other habits of mental health and self-awareness, not as a direct target. Compare notions like effortlessness, trying not to try, not trying to try, and global wayfinding.

Put another way, don't Goodhart yourself on keeping your identity small. Instead focus on nonmonotonic Pareto improvements that, as a consequence, will eventually make everything better, including less identifying with things.