High Actuation Spaces

Summary (or *hook*, really.)

Align *what* now? A surprisingly contentious question.

Yes, naively extrapolating currently available intelligences (like humans and GPTs) in order to understand superintelligence could be construed as negligence. Knowing how to “align” the ones that exist now does not straightforwardly result in safety of the ones to be developed later.

But what do we have in lieu? Some cluster of epistemic proxies to superintelligence that chant: *optimization, convergence, selection theorem, dominated strategy, coherence*. They attempt to look directly near the *end* of the trajectory of advancing intelligence, refusing to fall for anything in the interim.

These proxies are all pumped to the absolute limit, lest we commit a naive imperfection similar to overindexing on existing systems; lest we miss by a hair the hardest problem or the super-est intelligence or the sharpest left turn.

All excellent considerations. It can be dangerous to learn only from what's already here. But it's also dangerous to learn from what's not even here—whence all the rationalist imperatives to smash into reality.

Say you observe a trend like "the substrate of fiat currency keeps getting cheaper and thinner", you wouldn't want to extrapolate it to "the currency of the future will not require a substrate *at all*". That's a bad limiting operation. Sufficiently advanced technology is indistinguishable from magic, but not *actually* magic.

So let's fix that.

This project is an investigation into building a science of almost-but-not-actually magical regimes. Spaces where actuation is extremely cheap and fast, but *not* free and instantaneous.

This is not fantasy-work. This is true, for example, already for your mental world, your imagination—a very real thing, fully embedded in this world. With some tilting, you can also see it in, for example: biochemical signalling, the formation of social structures, decision theory.

The hope is to be able to articulate many general and often counterintuitive facts and confusions about *mindlike entities in general*, including ones that exist already—a non-spooky *model phenomenology & ethnomethodology*, or *prosaic* agent foundations—and apply it to fundamental problems in the *caringness* of an AI, like value-loading/ontological identification/corrigibility.

You might call this a “deconfusion” project along the above lines. However, a good place to get a more *classical agent-foundationy* picture (albeit slightly dated) of the kind of objects & contexts of investigation is the post Steam.

The non-summary

[If you want to skim on a first pass, read just this **Intro***, browse the* **Examples** *section right after, and take a look at the list under* **Output**.

If that intrigues you, check out the **Plan** and **Skill requirements** *to assess your fit*.

I’d then recommend reading **Background** *and* **Examples** *more thoroughly, alongwith the live doc with more examples, tables, and notes on the outputs*]

Intro

> *“Mathematics is the part of physics where experiments are cheap.”*
> –Vladimir I. Arnold

If you're reading this, there are some regions of the universe that you seem to have remarkably precise control over. Consider that if I ask you to imagine a zebra-painted Macbook, the neural circuitry within you—the neurons, the chemical composition, the entire apparatus—will literally, physically rearrange itself to make an image appear in your experience, almost instantaneously.

You might say that the *actuation* of certain things within your brain is relatively immediate, or abundant, or cheap, or responsive, or sensitive, or chaotic, or informal, or alive, or even "meaningless fantasy". There might be many more adjectives that are apt, some of them inextricably connected, some of them quite surprising. I'll generally use the tentative term **high-actuation space** for such regimes. (This is obviously a very contextual term, certainly agent-relative. As we'll see, it is indeed *radically* contextual.)

As the world attempts to move into a more animated and more obedient-seeming world (at least, in the meantime, before possible catastrophe), more of your larger, "ordinary" surroundings will rapidly come to also be characterized this way. More and more of "the physical world" (which label is subtly dualistic, and this agenda intends to correct that, in-part) will feel and behave like "the mental world", bringing in all the wonders and terrors along with it, amped up to eleven and probably asymmetrically.

The aim is to study the radical implications of taking **an embedded view of these high-actuation spaces**, rather than conveniently ignoring or idealizing them away.

There are several direct lines of strong relevance to AI alignment, but they are not obvious at first glance. The most readily apparent and most important connection is to issues around corrigibility. Investigating **the structure of the *arena* where the formation, stabilization, and correction of (the pointers to) value** happens is not only critical, but perhaps *all* of the work.

(It might be confusing that the above paragraphs moved from a mind-likeness of the external world to the inner world of a mind. This will come together a little more in the Examples section below.)

Given that this problem (how to point the AI's caring anywhere at all) is considered by many, including the author, to be about the hardest part of technical alignment work, this will probably be enough motivation for many. For others, we allude to some heavily in the overlap, without comment for the sake of brevity: self-location/situational awareness, deception, (embedded) abstraction, values-change, shard formation, coordination/multi-polarity, model psychology, decision theory, substrate arguments.

Examples

I'm hesitant to offer canonical examples before the background section below because it's easy to mistake them as being either a) perfectly physical or perfectly platonic and therefore clear or b) as arbitrary or contingent or constructed and therefore irrelevant. Either way, not in need of some weird "high-actuation" science. More about this dichotomy in the Background section.

But it's probably best to have a few concrete things to think with anyway. Just keep in mind that the "space" in "high-actuation space" isn't quite the right word; it's too observer-independent. "High-actuation lens" or "high-actuation frame" is also close, but is too observer-*dependent*.

The examples might seem to span/alternate from mundane to utterly fantastical. This is a feature, not a bug; the spectrum could be tilled for insightful correspondences.

I've unpacked the first two examples sentence by sentence to make the generalization easier. The others are left as an exercise :)

(The list of examples and a table of correspondences and interdependencies are updated ongoingly here.)

1. Ordinary actuation

Take the simplest possible actuation: moving around your hands and fingers. You have enormous flexibility and can orient them in a literal infinity of poses, effortlessly. Yet, there isn't an "equilibrium" posture they must be in or are tending towards, no "actually" "honest" gesture to settle in at "the end". Scientific or engineering principles of movement aren't foregrounded except in the edgecases of injury or disability or infancy or unusual athleticism. These sorts of structural constraints for how to move your body do continue to circumscribe the domain of movement, but most of the space available remains unexplored — primarily shaped by social norms around you. Consider in particular waving your hands or offering a shake or a hug; all of them might denote similar things, but have importantly varying connotations. And they're unusually hard to innovate on; the ones in existence are kinda sticky... although it does happen (like elbow bumps coming out of pandemics), a slowly co-created culture like the development of a shared accent. It's extra hard to disrupt not only because we implicitly harmonize with each others' movements by default, but also because we explicitly *teach* each other, especially children, to inhabit the same gestures, reifying them as the "correct" way. The ones pronounced "more correct" tend to carry cute symbolism, like openness of arms being openness of friendship. The "correct" ones that are fixated become *liberating constraints*, for creativity atop these more crystallized forms, like a customized homie handshake. Movements that don't have cultural steam are barely noticed, are just noise, lack legitimacy, almost lack existence; twitches. But really, no dances of hands and fingers have any inherent value or meanings if you focus on them in isolation; there is no fundamental discovery of insultingness to be had in modularly exploring the middle-ness of a middle finger. It's largely arbitrary, with no compact originator, yet often with profound impact anyway.

Unpacking:

- *enormous flexibility; orient them in a literal infinity of poses; effortlessly*
    - This is the high-actuation, the ability to cheaply instantiate
- *isn't an equilibrium posture; no "actually" "honest" gesture to settle in at "the end"*.
    - There is an ongoingness, no doneness or final fulfilment in a high-actuation space, very unlike a map that is intended to end up perfectly matching a fixed territory
- *Scientific or engineering principles of movement aren't foregrounded*, *primarily shaped by social norms*
    - Deep understanding of the substrate tends to be less relevant in high-actuation spaces, though not irrelevant in most contexts to understand their typical use or meaning
- *might denote similar things, but have varying connotations.*
    - Meanings are shaped by connotations rather than in what is crisply denoted; high-actuation spaces have a very different notion of "correspondence" and "inference" that is unlike standard truth-correspondence and "noise" becomes "signal".
- *hard to innovate on; harmonize with each others' movements*

    - There is a kind of constraining happening anyway in a high-actuation space, but they have somewhat “circular” or reflexive reasons rather than idealized rationality justifications or physical law constraints
- *explicitly teach each other, reifying them as the “correct” way*
    - What started out as being merely descriptive, turns prescriptive, and vice versa, without any clear boundary in a high-actuation space
- *tend to carry cute symbolism*
    - There are “superficial” correspondences that are somehow meaningful anyway, and thereby influence the dynamics and stabilization of what occurs in a high-actuation space
- *fixated become liberating constraints, for creativity atop these more crystallized forms.*
    - The fact that some of these high-actuation abilities end up being constrained anyway, form the low-actuation solidity need to craft the structures at the next level of organization,
- *Movements that don’t have cultural steam are barely noticed, are just noise, lack legitimacy, almost lack existence.*
    - This is the tendency to neglect things that aren’t considered “solid” or “determined” or to have “steam”, because they’re still high-actuation
- *have [no] inherent value or meanings in isolation; no fundamental discovery in modularly exploring*
    - Modularity, like staticness, is a luxury of low-actuation
- *largely arbitrary, with no compact originator, yet often with profound impact anyway.*
    - The flexibility and the meaninglessness “underneath” does not translate to null impact

2. Emerging ontologies/babbling

When working to get out of your current frames of thought, you don't use simple deduction. If that were sufficient, it would simply be an implication *within* your old frame. In actually breaking free and getting to truly new points of view, you instead often use aesthetic judgement, sometimes saying irrational things (from the POV of the old ontology). The sharp modular conceptualization might need to be retired, and you might need to go back into messier places. That doesn't mean a complete free-for-all. You'll play with terms and still reject an old term---not because it's wrong, but because it has sticky frames and connotations that you're moving away from, guided by a sense of things working together in a more loose form than consistency. While working with very unsystematized thoughts, "coalitional reasoning" and “free association” among ideas can be necessary; sometimes you’ll need to be defensive against too strong a tide of “logic” or other formal pressure from extant systems of truth or meaning, seemingly counter to a principle like “that which can be destroyed by the Truth, should be” while simultaneously avoiding legibility bias. Things that are usually coupled become uncoupled, and vice versa. But also, old concept handles become more like icons or symbols to what is actually here—the new deeper joints of reality—and so both have and don’t have meaning. Eventually, you'll baptize some

anchors to what you're exploring, somewhat arbitrary within an isomorphism class, slightly optimized for memorability/memetic fitness, that then become the ingredients of your meaning-making activity. They ossify with repeated use, until your next transformative experiences. Each time, it's unclear where exactly the new ideas come from when the ground itself needs to be pulled away.

Slightly less handholdy unpacking for this one:

- *don't use simple deduction; often use aesthetic judgement*
    - The usual notions of "truth" become secondary to working with more dense connectedness in the actual space of the objects of study, orienting with smells more than precise sight
- *modular conceptualization might need to be retired; into messier places.*
    - Retreat to a high-actuation space
- *doesn't mean a complete free-for-all.*
    - Flexible, but not infinitely so
- *not because it's wrong, but because it has sticky frames and connotations*
    - Reasons for rejection can be mainly how a conceptualization subtly invokes particular tools rather than what it sharply denotes; it's not a clear division of signal and noise in a high-actuation space
- *a sense of things working together in a more loose form than consistency;"coalitional reasoning" and "free association"*
    - The equivalent of more reflexive/circular reasoning rather than entirely derived grounding
- *coupled become uncoupled and vice versa*
    - Determination is overcome and recreated in different ways
- *defensive against "logic" or other formal pressure from extant systems of truth; counter to a principle like "that which can be destroyed by the Truth, should be".*
    - The method of determination from existing systems can be misleading, involves overcoming of the usual tendencies of determination
- *Old concept handles become more like icons; both have and don't have meaning*
    - There are few "buttons" to make something happen in a high-actuation space; more symbols, to summon the relevant connotations and energies to work with
- *baptize some anchors; somewhat arbitrary; become the ingredients of your meaning-making activity*
    - It's useful to crystallize and have some things count as being determined in order to dance with them, to orient in a high-actuation space
- *slightly optimized for memorability/memetic fitness; ossify with repeated use*
    - The logic of the substrate (in this case, how things tend to hold in attention and memory) can play a part, even though it would be "cheating" or "silly ritualism" from a low-actuation space perspective
- *unclear where exactly the new ideas come from; ground pulled away*
    - It's harder to deny that there is no compact source of ideation when confronting the high-actuation; but that doesn't imply a lack of value of what's birthed

3. Currency

It's hard to say what exactly is so optimal about having famous people printed on banknotes that becomes the central obsession of civilization. The answer "nothing" is a good one. What causes us to continue to believe in them, then? Only that other people believe in it — and they have really only the same reason to. It's easy to fall for this hyperstition quite hard and "goodhart" on making only your bank balance go up. It's also easy to forget that things that haven't been formally economically tracked and monetized might still be valuable. But our best economic theories of value still have "subjective" right in their name. "Currency" is a good word, because it connotes both the contingent narratives that are currently in vogue and the pull of that momentum on us. Certainly the substrate of currency, the token itself, is like information, highly flexible. This is most obvious in an extremely optimized, highly thinned currency, like cryptocurrency, requiring little historical or institutional significance to power it, although a connection to/exchange with dominant fiat currencies is a must in the interim, even if the intention is for them to fade. They're quite volatile investments, partly because it's easy to spin up a new token with no real grounding in a matter of minutes... and yet a new cryptocurrency equivalent to an entrenched one would not immediately jump to the same market cap, despite the merits being the same. On merit, even though some chain selection rules are better designed than others, there's still a question one could ask of each of them, such as "how did we coordinate on *the rule itself*?" Or even, have we actually *achieved* coordination, in full? Regardless of the kind of currency, once somewhat established, they're a very powerful steering force despite being nearly empty of worth in themselves.

4. Mind

This is obviously the main example and the thing we care about in the context of alignment. Instead of attempting to prematurely fold up the to-be-developed thesis here, I'll only say for now that the perspective taken for the mind (apart from that of ontological change above) is that of values as *internal currencies*, with a similar subtle dynamic of "substrate and spirit" as above. I'll only allude to some possible radical implications to explore and justify: fuzzing together of terminal and instrumental values, validity of wishful thinking (a la "iconicity" in the examples above), mixing of prescriptive and descriptive, looser notions of reflective stability, deep flexibility of values, ongoingness rather than fixity, optimization as self-undermining, "irrational" considerations for self-modification, prescriptive understanding of ontological identification.

Wherever you're tempted to use a "but dominated strategy!" or similar idealized-agency argument, I'd counter with "you're not looking at the high-actuation space of this". Part of the claim is that the staticized, finalized, formalized parts are

not the main aspect, certainly not the entirety, of the pointers to value from within an agent.

Background

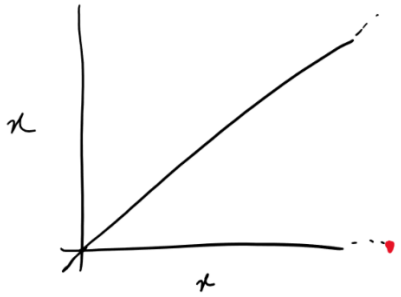
There are two relevant fallacies/misleading frames that each make up their own posts. But they're essential background for much of the discussion, so a quick introduction follows.

Warning: this section is written in an opinionated tone. I think that serves a function. If you're either very sold or very triggered, that's a sign of a great fit!

Chart Sleight

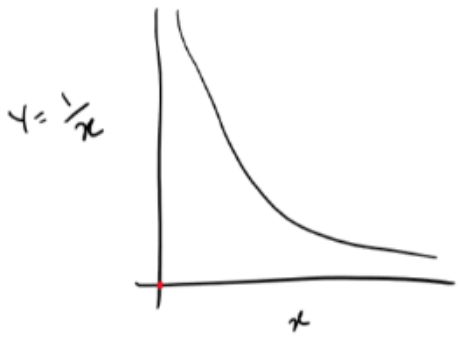
[Excuse the extremely hasty drawings!]

This is the simple identity graph:



It's defined everywhere, and I've marked with a red point the "point at infinity" that it isn't defined on in the typical real line (rather than the extended real line, for example).

We also look at the hyperbolic function 1/x:

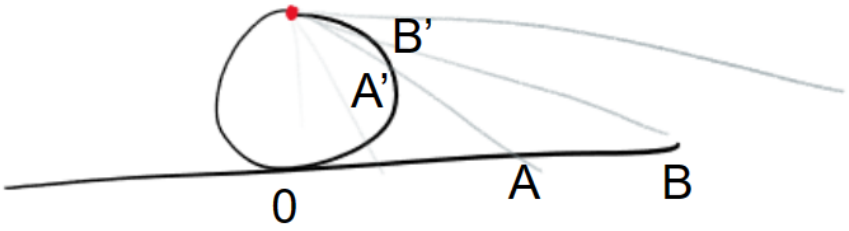


In this case, there's a (hopefully visible) red dot at zero, because this function isn't defined there. Although, quite dually, it seems tame "at" infinity.

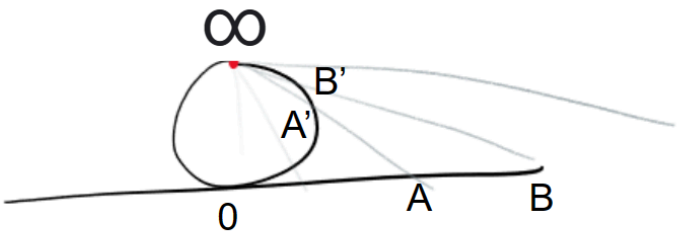
If you've taken a course in differential geometry or slightly advanced complex analysis, you'll know another way to look at these two as being the same thing from opposite sides: as being charts centered on opposite poles to cover the 1D manifold that is the circle.

Ignore the above paragraph if you have no idea what it means. Instead, notice this cool fact: you can map each point on a circle to points on a line. How? Look at this image:

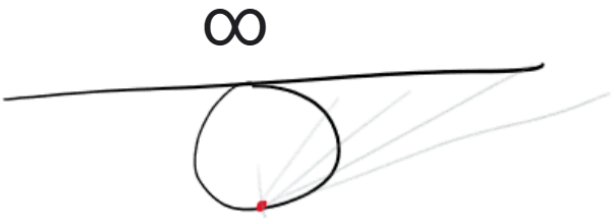
The circle is placed above the line, with its South pole touching the line. For any point on the circle that you'd like to map to the line, draw a segment connecting it to the North Pole, ie. the top red point on the circle. Where that segment intersects the line, is the corresponding line-point for the circle-point. So in the diagram below, A <-> A' and B<-> B'.



Notice this assigns a one-to-one correspondence for every point except the top point of the circle, which is parallel to the real line below and so remains unmapped (like a point at infinity; hence the red dot):

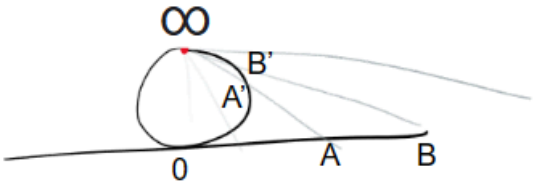


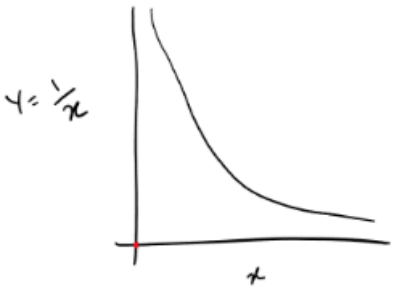
We could have another one-to-one mapping as the following though, where we shift the circle to be underneath the line, with the North pole touching the line and the South pole acting as the segment-endpoint:

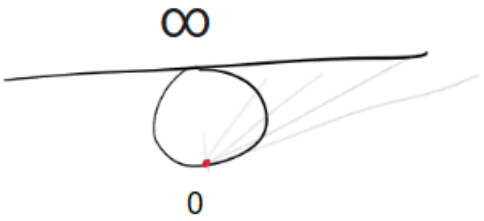


:

Now every point on the circle is mapped except the other pole. If you compare the two mappings, you'll notice that the point x on the line gets sent to the point 1/x on the previous line (true only if the line ran through the middle of the circle, to be mathematically precise, but oh well). The second "viewpoint" is like an inversion of the more normal first one, a hyperbolic transform. In fact, you could set up a table like this:







The point is that even though you can't cover the circle with just one mapping (if you don't allow a point at infinity), you can do it with two viewpoints or mappings, called *charts*, that together form an *atlas* if they're compatible in the right ways. An atlas is great: as long as we are careful to track which chart we're working on at a time—"work on" here could mean familiar operations like using coordinates or taking derivatives—we can work on any point of the circle consistently as if it were just a line. This is the secret sauce to working with more curvy surfaces that look "locally" flat. When you're taking the "hyperbolic" graph/chart, you're sort of looking at everything "inverted", but that's fine, even useful, to work with the "point at infinity".

You can do something similar for 2 dimensions, which I won't spell out but the image might be enough:

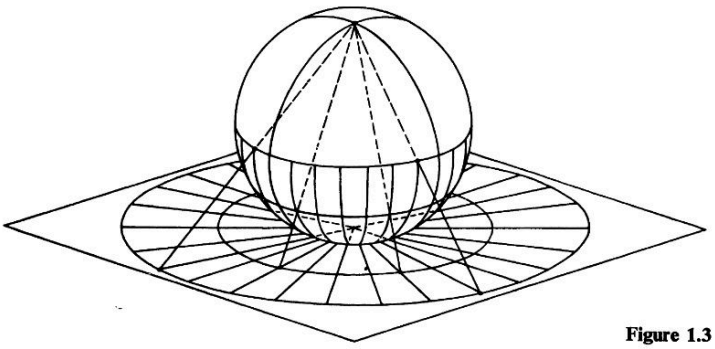


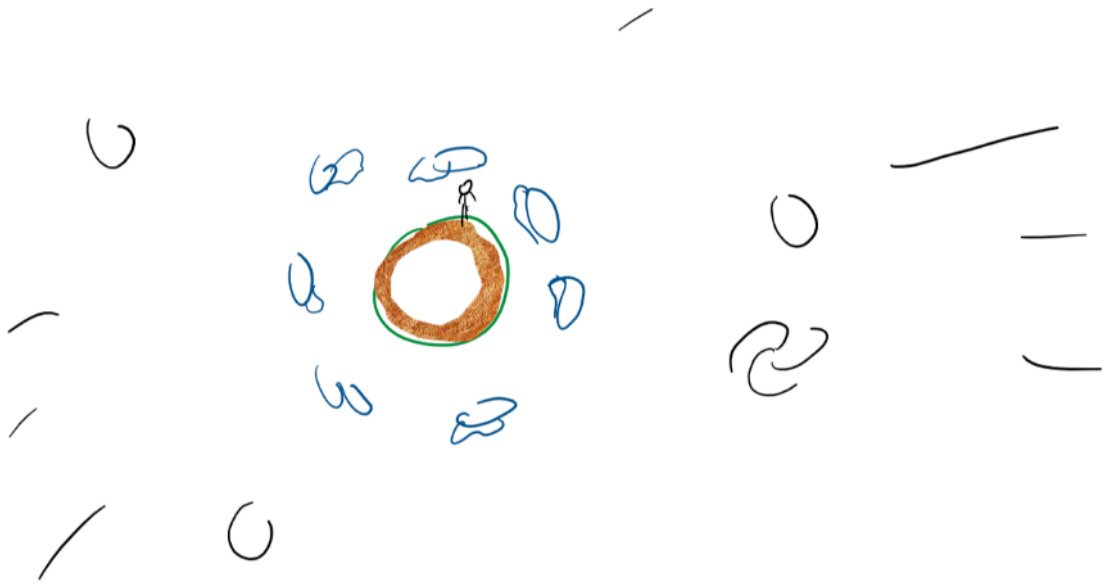
Figure 1.3

Again, with only two charts, similar to the above. Of course, this time the correspondence is to $\mathbb{R}^2$ rather than the one dimensional line. You would need more than just two charts to cover a curved space in general, but we won't worry about that.

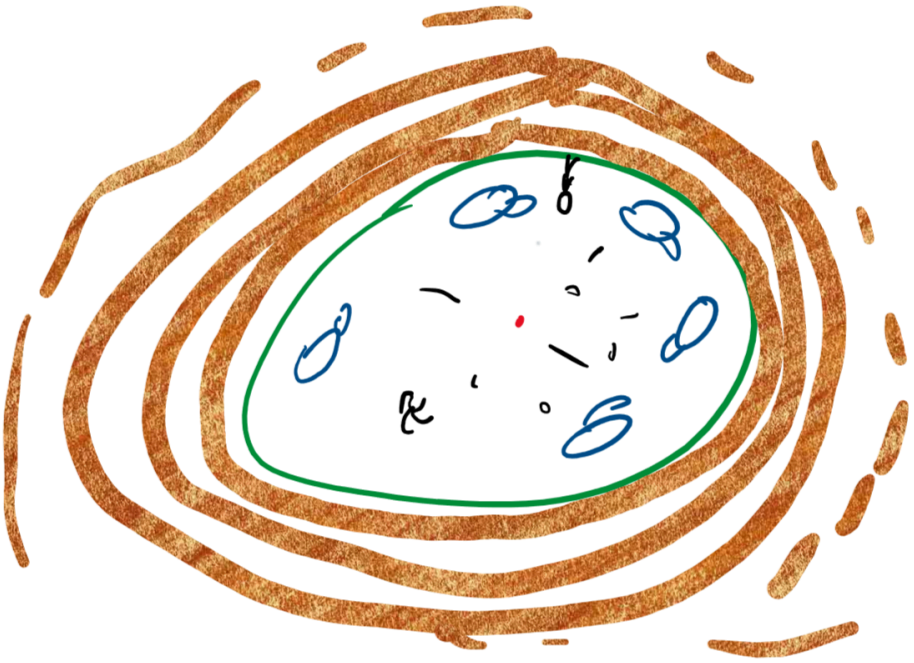
We're going to jump straight to 3 dimensions (same principle), although with the additional challenge of being quickly and badly drawn in 2D.

Specifically, you could pause to imagine what it might look like to experience your 3D space have a hyperbolic transform. Where instead of the outer world stretching out to infinity, it stretches *inwards* to infinity.

The normal picture is this, you standing on the planet and looking out at the stars and galaxies throught the clouds:



versus the hyperbolic picture:

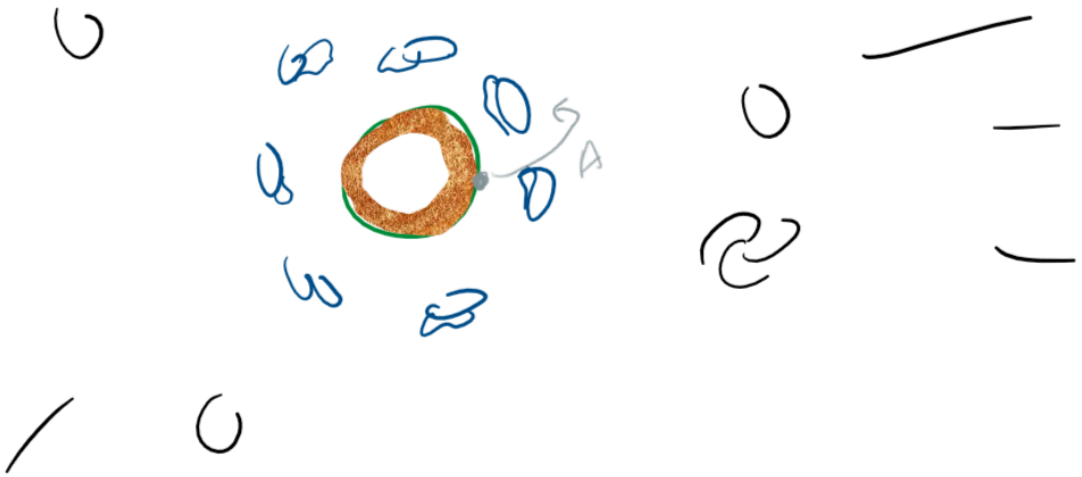


(The terrible brown lines are supposed to be the ground stretching outwards, and the point for spatial infinity is the red dot in the centre now, where all the many galaxies of the universe collect around)

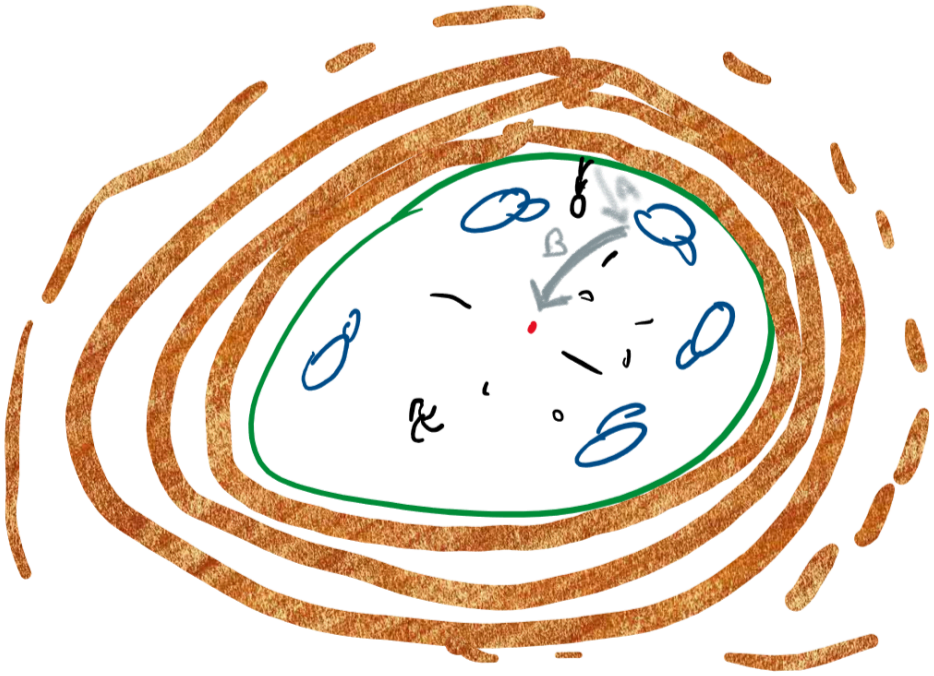
This is a strange looking world, but that's just the nature of taking an unusual chart.

But imagine if someone were excited to use the charts and had a clever idea to get to infinitely far galaxies in a short time by taking a two-legged journey:

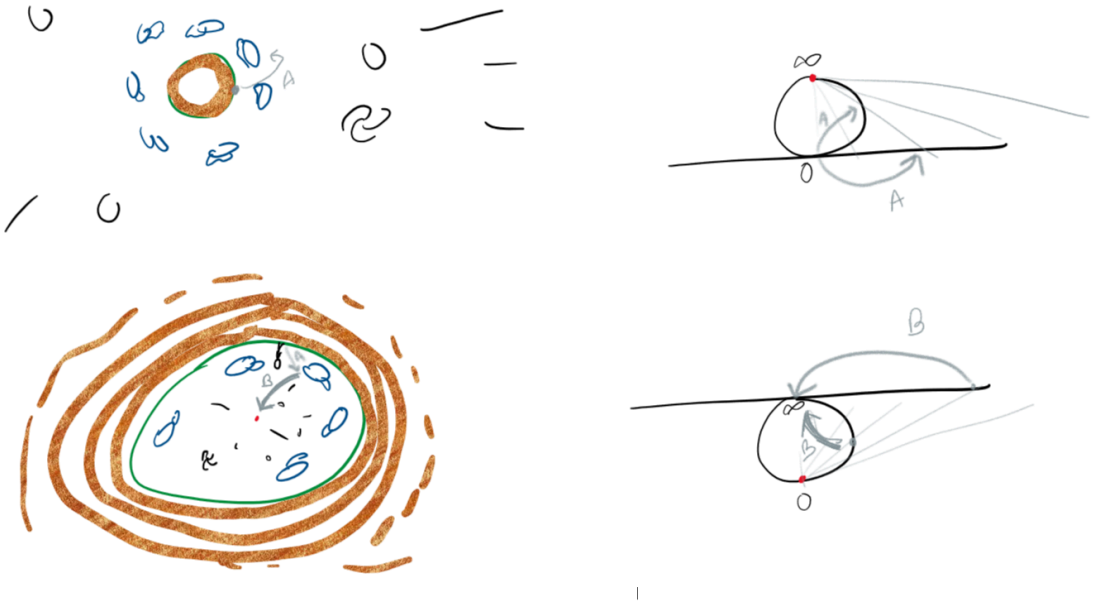
First you get to the clouds in the first chart in finite time, the first leg, say "A":



Then you switch charts (this is allowed!) and get to the point at infinity in finite time, leg “B”:



The equivalent on the 1D picture laid side-by-side to get from zero to infinity in two finite steps again, A, then B:



What’s wrong with this picture?

[break to think]

The inverted/hyperbolic world is a strange world, where it seems like it's a short distance to get to the far reaches of outer space and an infinite distance to get to the core of the earth. The problem is that *your idea of distances is also warped* in the second chart. A constant speed will start to "shrink" in the inverted chart, as you head towards the centre. Your size will start to shrink too. It *is absolutely okay* to make use of the other chart, as long you keep this in mind. It's *not* okay to mix up properties of one chart (like "normal" intuitions of distance) *while working in the other chart*.

That's what I call a **chart slip:** switching a chart and then mixing up their properties. This is relevant to "taking limits" in idealized agency, among other things.

***

Now for a practical example: the substrate of currency that was mentioned at the beginning of this doc.

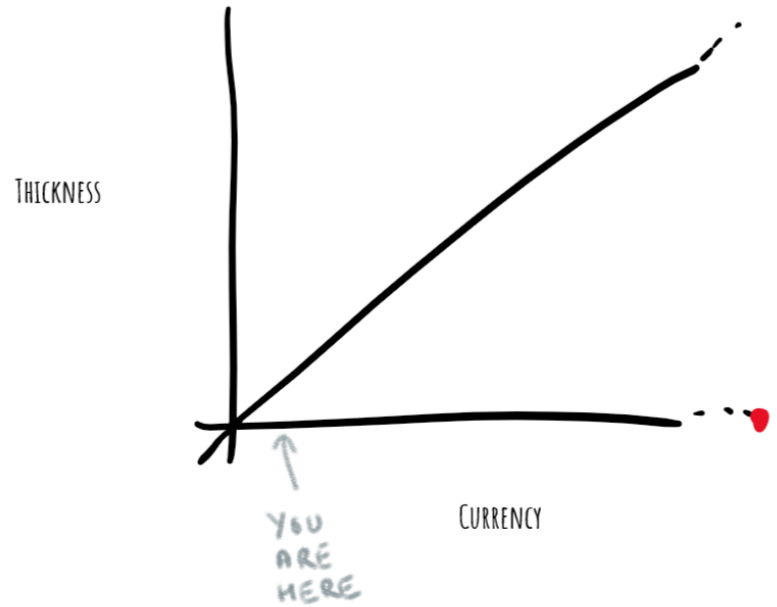
The "thinning" down of the substrate of currency from cows to coins to credit cards to crypto has been very useful, much cheaper to operate, at least for the end user.

THICK

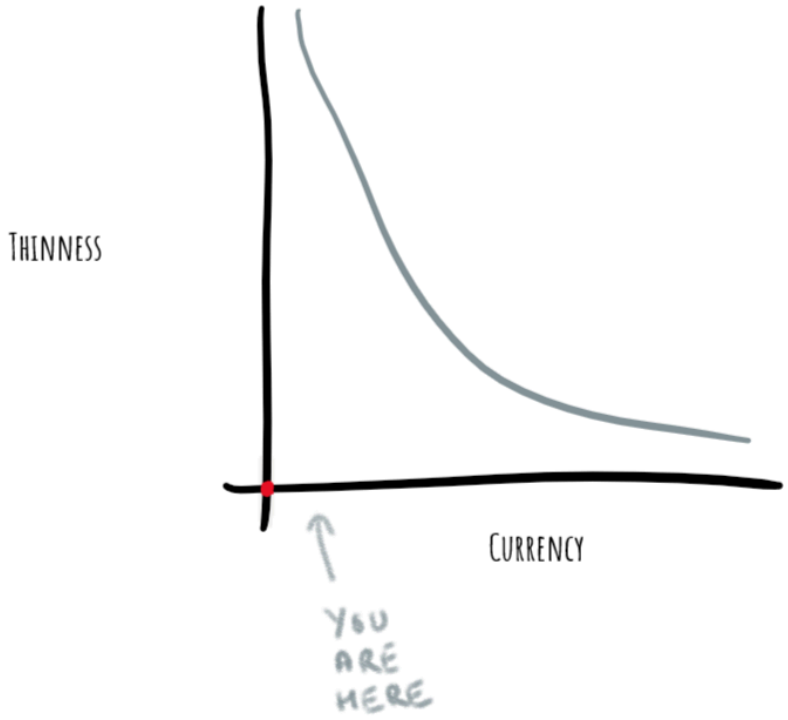
10101001001
01000101010
00011111010

THIN

But instead of the "normal" variable being "thickness" that will soon go down to literally zero as we move ever leftwards…



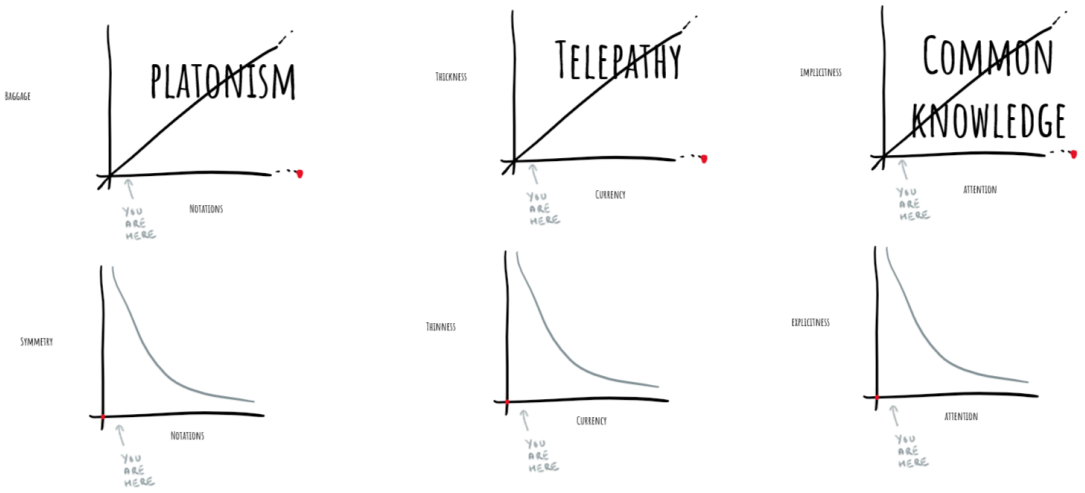
I would invite you to see it as thinness going higher and higher but getting increasingly harder to go to infinite “thinness”:



Again, there’s nothing wrong with taking either chart/view per se. The only error is in thinking that you could, within some finite but very long time in the future, get to *no substrate at all*.

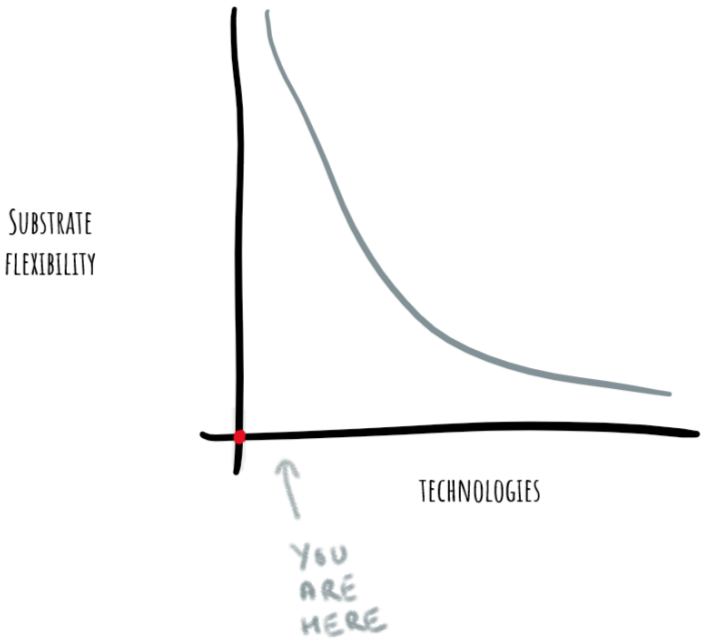
That would be a **chart slip**, because it looks like we're very close to almost zero thickness of the substrate. A world with literal telepathy (no substrate needed at all) is not the real one, though it might be fine to take that limit in some contexts. It's merely one where it is relatively extremely cheap or abundant—you can see the connection to high-actuation.

There are similar things in getting *rid of notational baggage* (versus the more "normal" *finding of symmetries*) to avoid falling for chart slip when thinking about Tegmark universes, and getting to zero implicitness, or full common knowledge (versus the more "normal" *raising of explicitness*):



Primarily, I want to **abandon *substrate independence*** as a useful term, except in rare circumstances. Instead of substrate dependence going down to zero as we get ever improving technologies to the left…

we only get to have multiplying realizability or improving *substrate flexibility* as we move to the left:



And all of the *high-actuation* investigation is in taking the substrate (and its flexibility) seriously, even as it gets *really, really flexible,* rather than rounding that off to total substrate independence.

You could still take the view from the chart centred at the "ideal pole" where you do have substrate independence. But then you might miss on reality entirely:

[Diagram]

"In ideality there is no difference between two instantiations of an abstraction, but in reality there is" mirrors the "In theory, there is no difference between theory and practice. In practice, there is.",

(The attentive reader might see more than one connection to Box-inversion here. That's certainly relevant, but with perhaps a quite different purpose.)

Instead of rounding off "it can be instantiated in very, very, very, many substrates" to "it can be instantiated in any substrate", the invitation is to keep a track of the network of interconvertibility as needed. More on this later.

Simulation & Construction: The simbox fallacy

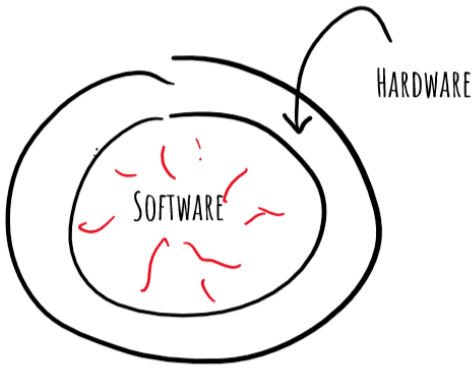
Since the previous subsection was already too long, I'll keep this one short though it deserves to be fleshed out more carefully.

The *chart slip fallacy* in the previous section was railing **against the word *"is"* in virtual-ish contexts**, in some sense, at least in statements like "is substrate-independent" or "is common knowledge". This lets us preserve **both the reality and the flexibility** of virtual-seeming/information-like entities. It isn't quite antiplatonic (because it's still okay to effectively work in the "platonic chart") but it is against mixing up of platonic and ordinary reality in clumsy ways.

This one is a railing **against the word *"in"* in virtual-ish contexts.** Like when we say "the character *in* the movie" or "the puzzle *in* the video game" or "the picture *in* my phone" or "are we *in* a simulation?" or even perhaps "values in a mind".

Briefly, this is a dangerous metaphor, at least connotationally. *Simulations are **not** boxes.*

This picture:



is an understandable but potentially misleading one. There's not a physical containment, but physical colocation, and software could alternatively be understood as a kind of a *metaphor for hardware*.

This is meant in the same way that *mathematics* and *models* and *interfaces* are metaphors. They are constructions that correspond to some aspect of what we might find ourselves in dialogue about, fun (and often tedious, intricate) analogies for some reality of the thing you want to operate. Metaphors that are so good, so precise, that it's tempting to use them in an overly convenient, somewhat slippery fashion and forget that they're metaphors.

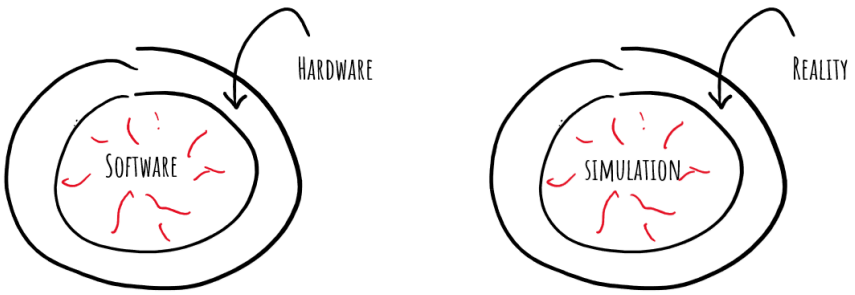
(I once said "rationalists often forget that math is a really good *metaphor* for reality; postrationalists often forget that mathematics is a *really good* metaphor for reality". It goes without saying that a "straw-" prefix is implicit for both identities in that sentence.)

When you talk about a “file” “in” your computer, a lot of things are happening at the hardware level that you’ve approximated. Calling it a “file” is a good metaphor, a skeuomorphism, for *some* of the characteristics of *some* of the things you can do with it.

“Software as metaphor for hardware” emphasizes the reality of the hardware substrate to the exclusion of the reality of the “higher” levels and so is ultimately untrue, but it does free up the picture of the container.

So if you were tempted up there to go “yeah, yeah, you’re talking about abstractions with less precise words, I know what that is”... well, I’m okay with rounding it off to your existing insights around that, so long as you escape all the subtle ways of still believing that simulations are boxes.

The idea that you could smush together a globule of reality that contains something that is “pure simulation” and not reality, is some unnecessary dualism.



Just because you’re “in” a simulation, *does not mean you’re not in reality*.

Yes, you can indeed have all kinds of barriers, physical and informational. That’s one way to construct a box, isn’t it? By surrounding something with barriers along all dimensions.

And so if I try to say “software isn’t *trapped inside* hardware, the simulatee isn’t *trapped inside* the simulation”, you might object: *but what if my eyes and ears are literally strapped to a VR kit, my legs on a VR-treadmill-thingy? What if it’s exactly like The Matrix? Am I not then exploring* only *a virtual world?*

Whatever you’re interacting with in its *most immediacy*, that *is* reality. Even if you think “your sensors” are being “fed information” that has nothing to do with your “real surroundings”---the machine you’re plugged “into”, you are plugged into *directly.* When you move your hand to press a button on the controller, or your eyes on the screen, or your attention “in” the matrix, that *really* happens. It happens *in reality.* Where else could the

interaction between “you” and the “interface” happen? Where is *all* of it embedded? Where is the computation happening?

This can be subtle, but it is also *extremely* obvious. Which of course, is what makes it subtle. The reason it can be a very stuck confusion is that you think reality “reaches” you when you fuse together your sensory experience somewhere. Or worse, and more likely, when you manage to *crystallize* a thought about reality (which, remember, is a representation).

This might still seem like not having direct access, but then you don’t have “direct access” (at least in the way demanded) to ordinary physics either. You’re in a classical hallucination, remember? And whatever eventual access you can get to past your hallucinations into the actual quantum world, a serious “virtual” physicist “inside” a video game could get to as well, from just a persistence towards uncovering the deepest regularities, extending “beyond” the more contingent/local video game laws that were cooked up for the simulatee on top of “base” reality. “Direct” in “direct access” should be retired just like “independence” in “substrate independence”.

(As a side note: what is the actual content in the questions “Are we in a simulation?” “Am I a Boltzmann brain?” “Is this ML model deceptive?”. I claim it is the question: “*am I expecting an inductive catastrophe?”* That is, “Do I expect all the learning and testing and evaluating I believe I’ve done so far to be less than useless, and not eventually all add up to what was normal before?” So there might be reason indeed to distinguish this *Anti-Egan’s Law* category, rather than calling all cases “equal access”. But the word “in” still only confuses rather than clarifies.)

> *“Tell me one last thing,” said Harry. “Is this real? Or has this been happening inside my head?”*
>
> *Dumbledore beamed at him, and his voice sounded loud and strong in Harry’s ears even though the bright mist was descending again, obscuring his figure.*
>
> *“Of course it is happening inside your head, Harry, but why on earth should that mean that it is not real?”*
> — Harry Potter and the Deathly Hallows

So the second subtle implication to all this is that *just because something is constructed, doesn’t make it unreal.* In this case, we preserve **both the reality and the contingency** of virtual-seeming entities. We don’t need to have eternal principles for them to be the “actual” real thing. There is no rhyme or reason to isolate some aspect of reality as “merely simulation” and then puzzle over how it regains reality-fluid; constructed entities (whether constructed physically or socially or mentally) and contingent laws are just as real and exciting to explore. **Without this thoroughly clear, studying the properties of high-actuation spaces might seem boring or pointless or unscientific.**

Philip K. Dick is famously quoted as saying "Reality is that which, when you stop believing in it, doesn't go away." This is a great exhibit of the confused way that we tend to believe that "reality is that which is independent of the mind's participation."

The point, of at least this background section, but really this whole subagenda, is to notice how we neglect a science of non-modularity, to the point of even calling it "unreal". High-actuation spaces are full of dependent arisings, transient constructed forms, instability in isolation, enormous flexibility of regularities. But none of that makes it unreal or unworthy of study. We could promote our tools to meet as-yet-intractable phenomena rather than demote their reality.

Accepting reality when there's such flexibility and interdependence makes it sound like the claim is there is no difference between fantasy and "reality", or social consensus and "reality".

Lack of differences is not the claim. Quite the contrary; the invitation is towards refined sciences, a fine-grained perspective of "abstraction". There really is a difference between an image of a sandwich and what you might call an "actual" sandwich. But this difference is more ordinary, like the difference between water and ice. More substrate-level, scale-level, expectation-level, *not ontology-level*.

This difference can be more subtle, because of a whole host of interdependent, mutually-reinforcing *determination biases* (lumped under "real") that are hard to notice. For example, one might be tempted to say "a picture of a sandwich isn't real; you can't eat it". But this is factually wrong. You *can* eat a picture of a sandwich, it just wouldn't nourish you or taste very good. It is utterly strange that we pack "nutrient-rich" into the word "real", but this is what happens when we have expectations automatically "coming off" of a pointer (like "sandwich") that you haven't noticed. This appears so obvious that we miss the contextuality of it, and end up pronouncing real/unreal instead of "satisfies a demanded property". High-actuation stuff happening inside your mind is similarly easy to mistake as not being legitimate or real (as in the examples). It could be useful to note this kind of contextual adaptation instead of being run by it. It is certainly useful to notice what you have been implicitly calling "real".

It is unscientific to treat low-actuation spaces as high-actuation (like hanging up a picture of a car to manifest one), or import high-actuation methodology and results into low-actuation. This is what gives you epistemic nonsense like "magical thinking". **But it is equally unscientific to do the reverse** — of restricting yourself to low-actuation methodology and results in high-actuation contexts — is the claim here. And that's an easy mistake to make when all high-actuation stuff is banished into "unreality".

Output

Some possible objects of study outputs, with a line each, follow. The links between what you've read so far and this section are developed in the subsection immediately after.

Classification
The most straightforward task is to identify the characteristics of high-actuation spaces and the domains and lenses of relevance, and make the connections rigorous, sampling from many fields.

Empiricism vs control tradeoff (or *conduction-construction opponence*)
A central piece is about how being more careful about limits might expose some of the self-undermining in monomaniacal optimization (caricaturous analogy: "Growth seems to be central to life. Cancer really embodies growth. Extreme life forms must be like cancer taken to the limit—*megacancer*." But "megacancer" doesn't exist, cancer eats itself up), possibly allowing for a crisp formalization of a tradeoff between seeing (empiricism) and doing (control).

Determination bias/Steam
The hope is to have some results similar to inductive bias, about the usefulness and restrictiveness from things gaining steam or becoming stickier and more determined than others within high-actuation spaces (and as a side-effect, have a more general formulation of issues and solution-concepts in decision theory).

Naturalized account of narrative
It may be more obvious at this stage that one of the central aspects of study is the substrate and spirit of all kinds of *currency* and how they become sticky within a collection of agents, *especially* internal currencies in one mind.

Substrate-sensitive/fine-grained information theory
Information theory/coding theory is about optimal exchange rates between fully determined entities (the codes) and probabilistically determined entities (the information channel), and having this emerge as an "effective field theory" for more generalized accounts of the relation between things of varying "determination status" in a high-actuation space; nebulosities other than just the probabilistic kind.

Terminalizing/internalizing necessity theorem
In examining with a finer-grain the process of decision-making and grounding conditional cooperation, there is a theorem-sketch to refine which makes the claim that at the heart of (self-)coordination lies the inducing of new terminal values (or at least many properties that are very close or equivalent). More generally, outlining the instrumental and terminal reasons to have terminal values-change.

Coexplication
The background methodology for much of the outputs (but not all) will be to produce *live theory.* More on this in the note on outputs linked below, but this is a purported dual to

explication, of explanations that go in the opposite direction of boiling things down to their parts to understand them.

Outline epistemic and communication challenges
There is (meta) progress to be had in just outlining why these might have been neglected, and working on making the relevant fallacies explicit.

Link: a note on outputs. >>>

Plan

We might spend the first few weeks:

- Discussing examples
- Identifying properties and elaborating on them
- Filling in correspondences and interdependences
- Embedding them in formalisms and literature
- Collecting crisp questions
- Generating answers

We might each pick up a specific example cluster suited to our backgrounds at some point in the process above.

We could then as a group lean into a rigorous investigation of a chosen few properties, or work on specific outputs that seem tractable.

Risks and downsides

Unlikely to be very dual-use, but too early to say for sure. We might have to be discerning about the visibility of outputs in the case that there is extremely surprising progress, especially if comparatively more applicable to current paradigms.

Acknowledgements

Conversations with Abram Demski (and others at MIRI) have been a central ground of creation. Thanks also to Sam Eisenstat, Evan McMullen, Linda for feedback on the draft and Clem, Steve Petersen, Nora, Martin, TJ, Dusan, Alex Flint, Ramana Kumar for conversations over the last year.

Team**Team size**

3-5 people

Research Lead
Sahil

Email: sahil@intelligence.org

Abram Demski is likely to join us for discussion.

Skill requirements

As might or might not be clear from the above, this is an excellent arena for those spiritually at the intersection:

- of math and philosophy
- of prosaic alignment/modern ML and agent foundations
- of computer science and biological/sociological lenses
- of rigor and ritual
- of material and phenomenological investigation
- of systematic and postsystematic modes
- of strong agreement and subtle disagreement with MIRI-esque (esp Nate/Eliezer) views on alignment
- of intrigue and skepticism around shard theory

For most of this, there is no specific background that's necessary, though some abilities to engage with and produce technical-ish language is likely to be. Additionally, if you find this interesting but

- are more interested in a legible engineering role than an illegible deconfusion one, you'd make a great teammate in the production of *live outputs.* How exactly this fits in is made clearer in the **Note on Outputs** subsection in the live doc. [Live Engineer role]
- are more naturally able to provide attention in the form of listening/nurturing/midwifing of ideas and want to get in on the ground floor of where this happens, you'd make a great *ground* for what is coming into being. This is not a "secondary" role in my view; that would be succumbing to channel bias. [Ground role]