Note: short version above long version

The Sun is Good

Incommensurable World Models and the Path to Meta-Modernity

Kevin Carlson

What is the world? Is it "everything that is the case?" Does it have different strata, whose realities are grounded in mutually incompatible ways, or just one? When we describe the world to each other, don't our choices of language privilege certain answers to these questions, and foreclose on the possibility of certain others?

I won't waste much time with monism (i.e. the answer “just one”), even when many fellow Western tech-and-science types seem to take it as not only beyond the pale but actually laughable to quibble with the monistic assumption of rank materialism. As Aristotle saw long ago, "things are said to be in many ways"; *pace* him, for reasons I can't quickly explain, the following millennia of careful research into the ways in which things are said to be have recently been overwritten with a quick snicker at Cartesian dualism followed by a jump to, at best, Dan Dennett, or maybe early Wittgenstein.

Yet, the simple fact remains that things are said to be in many ways, and eliminativisms and materialisms of all kinds are foolish, harmful when taken seriously, and ineffective in following and influencing the unfolding of the world.

Schopenhauer's View

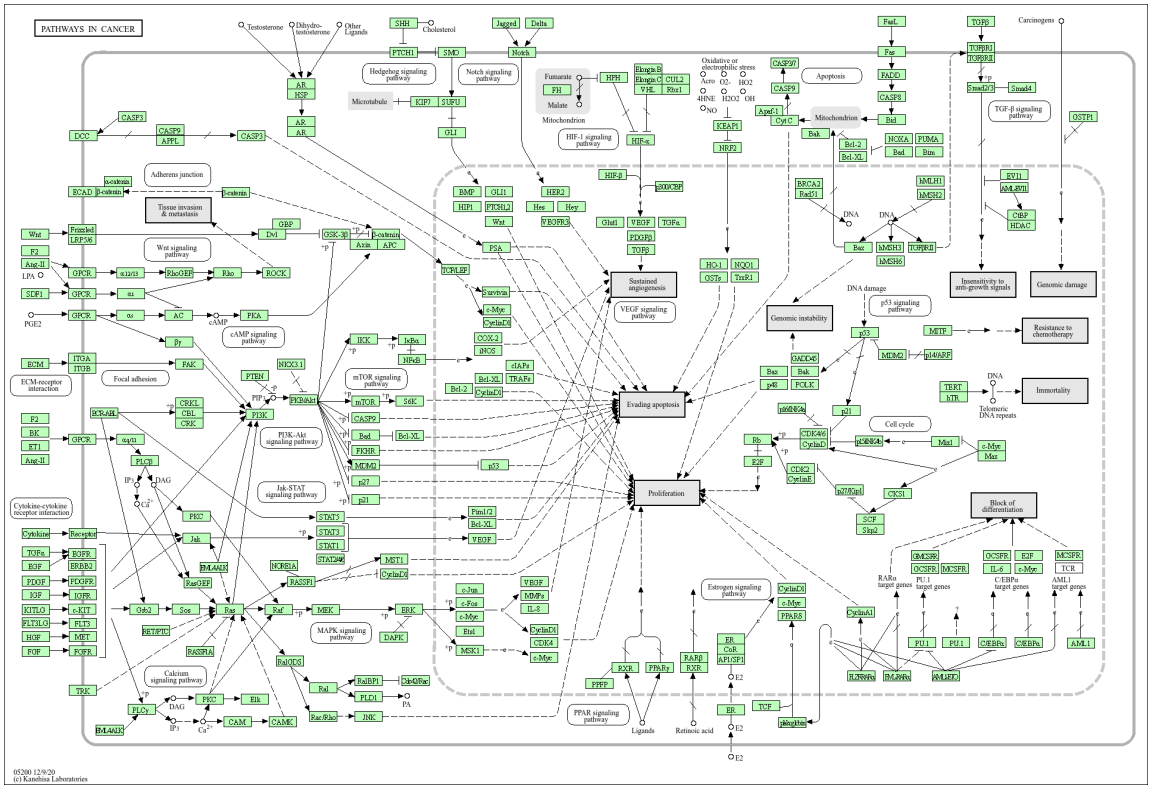
A favorite answer of mine to the questions up top is Schopenhauer's. He says things are said to be in two ways: first, as Will (capitalized *á l’Allemagne*), in which there are no *things* at all, really, just the surging, chaotic flow of pure Being that lies behind all appearances.

The second, which Schopenhauer calls the world as representation, is the familiar world of subjects-regarding-objects. What's interesting is that, even in this world of representation, Schopenhauer subdivides into four distinct ways in which things are grounded, or in which things can be: cognitive objects, grounded by reason (if all men are pigs and Socrates is a man, then Socrates is a pig), material objects (through causation), moments and places in time and space (grounded by other nearby places and times), and the ground of motive (the connection between intention and action).

Models, for real

Why am I inflicting this recapitulation of Schopenhauer on you? Well, let me get back to that…First: I’m interested in world modeling—what it is, how to do it, and whether AI can help.
Here are some examples of world models:

- GDP = C+I+G+X
- If GDP' > 0, then all in all the economy is doing well.
- God is love.



- The vibes are off with that guy.
- The rich are, by and large, talented, hardworking, and probably way cooler than you.
- The rich are, by and large, greedy, lucky, and probably ugly.

A world model is at bottom nothing more or less than a representation in the sense of Schopenhauer: a network of objects: a subject organizing (part of) the world to itself. Modernity is the state of a civilization that can develop precise language and intrepid norms of discourse to permit more-or-less free competition between publicly expressed world models. Postmodernity is the observation that grand public world models seem to often throw the baby (the illegible, fractally complex, organic system determining individual flourishing) out with the bathwater (naive adherence to tradition).

The fundamental problem of today's civilization is finding what comes after postmodernity, which correctly teaches that legible public models lead to runaway harm beyond the edges of the model, but then gets mired in inescapable negativity. Some *individuals* find a way out of nihilism to what might be called a meta-modern attitude of understanding and using modern systems (plural) without *identifying* with them, but our civilization hasn't advanced to this way of being at scale. (If you like this paragraph, you'll love David Chapman's writings on meaning and metarationality.)

The World Is Not An Object

My point in juxtaposing Schopenhauer's metaphysics and the dream of meta-modern world modeling is this: a world model is a claim, made by some particular subject, about what is to be attended to (perhaps, what is to receive *care*), which is pragmatically indistinguishable from a claim about what is real. The middlebrow attitude is that a model approximates the world. This is wrong, because the world is not an object. There is no "ground truth" to compare models to, not because we're solipsists, but because the world-in-itself is a chaotic surging mess in which no grounds are admitted.

The world is not an object. The only objects are our models of the world. This doesn't mean models can't be wrong—if you believe "vaccines cause autism," I can produce observations that (*defeasibly!*) contradict it. But it wouldn't be by "looking at the ground truth"; I'd observe that your qualitative model is inconsistent with what seem to me to be reasonable quantitative models, and we'd debate, presumably, until exhaustion.

A Story About the Sun

Let me tell a story about a day when Alice and Bob, a couple, were discussing the ideal frequency of sunscreen use. Alice thought people should use sunscreen every day. Bob opposed this, citing studies showing most deadly skin cancer risk comes from childhood sunburns, with moderate sun exposure possibly beneficial (via Vitamin D). Alice responded citing anecdotes of contacts who tanned and later died of skin cancer.

This seemingly rational debate concealed deeper models. Alice's unstated world model factors included "if Bob really cares about me, he'll listen without being annoying about statistical details." Bob's hidden model factors included "I'm highly sensitive to the perception of being forced into something, even when the point is correct" and deeper still, something like "the Sun is Good"—a quasi-religious view.

This pattern—where legible, objective modeling on the surface masks emotional, relational, and spiritual paradigms below—is not special to small-scale, intimate interactions, but characterizes human disagreements generally. Consider NAFTA debates: quantitative economic models may mask deeper motivations about the decay of childhood hometowns, manufacturing jobs' reality-grounding nature, or—of course—attitudes toward foreigners.

We often have this cargo cult of public conversation where we, complicated bundles of stories and feelings, drape ourselves in black cloth and hold out crisp graphs as the only focus of discussion, hoping no one will look under the drape at the messy, emotional, willful grounds of our complete world model. One clear sign of reaching into meta-modernity would be when the sad, angry, hopeful creatures under the cloth start explaining their real models in all their incommensurability, and only then seek common ground as necessary.

A case at larger scale: in problems like "design a zoning policy for Berkeley," I want everyone with a stake to potentially contribute. There are challenges:

1. Normal people would rather socialize than specify policy preferences
2. Models range from "Berkeley should never change" to complex quantitative projections
3. There are many stakeholders with stakes of many, contested, sizes

Sortition (jury-duty style participation) could help with problems 1 and 3, but number 2 is deeper. How do you funge the infungible, commensurate the incommensurable? Current options are:

1. Establish a top-down modeling paradigm producing scalar ratings that can be mechanically aggregated
2. Contributions are in plain text, recombined unpredictably and illegibly by decision-makers to produce a conclusion

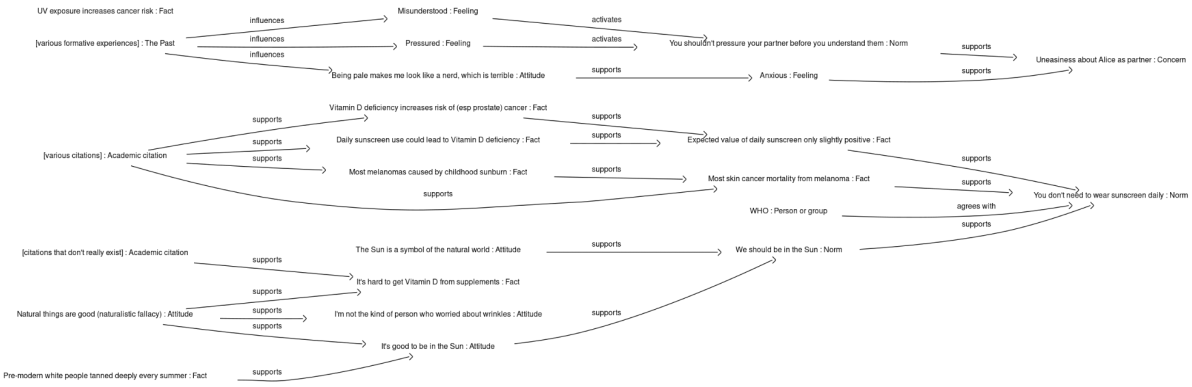
Both have problems: Option 1 excludes those who can't express themselves in the approved language, leads to overconfidence, and lacks consensus mechanisms to update. Option 2 determines the course of civilization by who yells loudest nearest the leader.

What Is To Be Done?

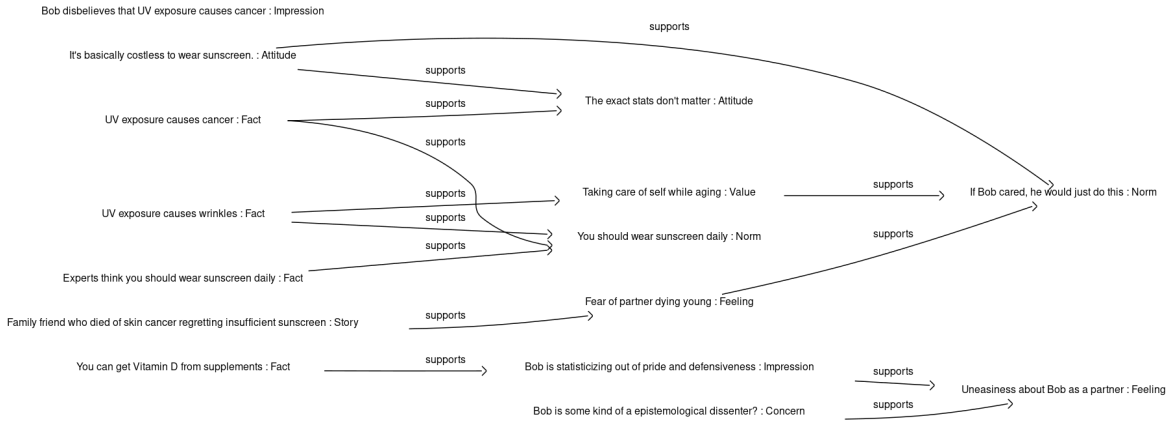
People haven't really learned to communicate their world models to each other; modernity was built on narrow exceptions that postmodernity has shown unsatisfactory. How do we get people explaining honestly, fully, how they see the world? This problem has spiritual aspects—getting people to notice how they see the world may require a lot of meditation, for instance. But that part really isn't my brief.

What do I offer, then? Structures for models: organizing, translating, comparing them. In the sunscreen example, the conversation would have improved if Alice and Bob had made their models explicit. Both could have seen the sensitive emotional points they were approaching instead of debating factual details. I think this would really help, modulo the UX problem of figuring out how to develop these models in the midst of a real, human conversation.

Bob's world model (click hamburger menu-> new analysis -> visualization to view graph scrollably):



Alice's world model:



For those UX problems at the very least, AI support is a critical activator, smoothing over what would otherwise often become a grind of manual model refinement disincentivizing the kind of deep iterated reflection that seems worthwhile even for this kind of quotidian discussion. We need low barriers to entry but high legibility and reliability. My suggestion, instantiated in the CatColab software being built at Topos and used to make the models shown above, is not to pick a single modeling language but a single way of *constructing* languages, formal enough to describe translations between approaches. AI can help by scanning possibilities and showing candidate jumps in the space of models. If candidate models have to

be expressed in a well-specified language, that solves some AI slop problems. If humans pick from candidate refinements, rather than relying on the AI to provide a consensus world model out of its broad averaging over humans-whose-writing-is-on-the-internet, that solves more problems.

The key to all this is to use AI to help clarify people's world models, not press people to align with algorithm-generated models. We must preserve and enrich individuals' and communities' abilities to see the world clearly and express their vision as it is, enhanced, never constrained, by new technological affordances.

The world, modeling

Kevin Carlson[1]

What is the world? Is it "everything that is the case?" Does it have different strata, whose realities are grounded in mutually incompatible ways, or just one? When we describe the world to each other, don't our choices of language privilege certain answers to these questions, and foreclose on the possibility of certain others?

I won't waste much time with monism, here, even when many fellow Western tech-and-science types seem to take it as not only beyond the pale but actually laughable to quibble with the assumption of rank materialism. In fact, as one as concrete-minded as Aristotle noted long ago, "things are said to be in many ways"; for reasons that I can't quickly or fully explain, the following millennia of careful research into the ways in which things are said to be have recently, in much of the Anglophone world at least, been overwritten with an education consisting of a quick snicker at Cartesian dualism followed by a quick jump to a proper philosopher, like Wittgenstein (which Wittgenstein? Well…better be the one I already linked, not the one I'm about to link) or Dan Dennett.

And yet, the simple fact remains that things are said to be in many ways, and with the utmost respect for the lately departed, eliminativisms and materialisms of all kinds are foolish, harmful when taken seriously, and always ineffective in following and influencing the unfolding of the world.

The good earth

In how many ways are things said to be? A favorite answer of mine is Schopenhauer's. He says that, at the top level, things are said to be in *two* ways: the first way, as *will*, in which there are no *things* at all, just the surging, chaotic flow of pure Being (you're allowed to randomly capitalize nouns when you're writing translated German, don't get nervous) that lies behind all appearances. One would think that this would be a real "whereof one cannot speak, thereof one must remain silent", which is why Schopenhauer only wrote a thousand pages or so of (exquisite) multilingual prose about this side of the world.

The other side, which Schopenhauer calls the world as *representation*, is much less existentially threatening. This is the familiar world of subjects regarding objects bumping into each other. What's

[1] Kevin is a translational scientist at the Topos Institute in Berkeley.

interesting is that, even here, Schopenhauer subdivides into four sharply distinct ways in which things are said to be, or in more transparent language, in which things can be *grounded*.

1. The grounding of *cognitive* objects is probably the most familiar to us: if I think that Socrates is a man and that all men are pigs, then I think that Socrates is a pig, as the old chestnut goes, and the latter object of my cognition is *grounded* by the former two. (What does materialism have to say about…*any* of this? "Well, like, you can reduce your beliefs about Socrates to patterns of neuronal activation subject to physical la—"please be so for real right now.)
2. According to Schopenhauer, following Kant, and being generally pretty smart, the proper way in which *material* objects are said to be is in terms of *causation*. The material world is the world in which things bump into each other, in other words. (If you're about to get all fussy about how everything is really a quantum probability cloud and nothing ever bumps into anything, well, you sound fun! That said, since you asked, Schopenhauer is way ahead of you: your problem is that you're attempting to think of an *object* as the *Ding an sich*, the thing in itself, floating off in "reality" all on its own. Schopenhauer's answer is "no object without a subject." There's no such thing as thinking of a chair-in-itself; all you can think of is a *representation* of a chair, an image, and that image is presented to a subject. Now you might be mad that we're talking rank idealism, but this is not Fichte (to which hardcore idealist Schopenhauer is hilariously mean). The subject doesn't *create* the object; they *co-create* each other, constituting each other through a *relation* that is the more fundamental reality than either of its endpoints. (Are we talking about category theory yet?) And he/they/we is/are not saying that the Universe didn't exist before humans came around, either, at least not in a stupid way: things-in-themselves are a whole different paradigm of existence, where there are no subjects nor objects, and you can stick your pre-life cosmology in the world as Will. Anyway, I say all that to clarify that in the physical world *as presented to your perception*, or indeed to your cat's (because you can have *perception* without *cognition*) things totally bump into each other. I hope any "but, akshually" urge is satisfied for the moment!)
3. Another way in which things are said to be: what is a moment in time? A moment is obviously not a physical object, it can't bump into something else. It's not a concept, though perhaps (probably not) you can conceive *of* it. So a moment in time isn't grounded by causation, nor by judgment. It's grounded by *other moments in time*, those just before and just after it.[2] You can do something analogous for space, and this leads to Kant and Schopenhauer's interested much somewhat maligned ideas about mathematical epistemology: the good Germans say we can come to know arithmetical facts through our intuition of time—consider the principle of mathematical induction, the old "and so on", to be convinced of this—and geometrical ones through our intuition of space.[3]

---

[2] I think it's reasonable to imagine that you can take an *arbitrarily small range* of such moments, so a single moment in time becomes something like a floating tangent vector, or the incarnation of an arrow pointing "that-a-way."

[3] It's widely believed that these ideas are out of date, since these guys didn't even know about non-Euclidean geometry, or whatever, but that's confused. During the re-founding of mathematics on set theory around 1900, we learned how to reify conceptions of space and of number that would have been unimaginable to mathematicians of Schopenhauer's day. But, as JP Mayberry observed, the import of this is not so much to show that Kant was *wrong* in his belief that our built-in intuition of space is the fundamental source of our ability to perceive truths about geometry (i.e. of 2- and 3-dimensional Euclidean space), but to allow us to *extend* that geometric intuition to spaces of high or infinite dimension or those with various other exotic natures. It's still true that mathematicians figure things out using their spatial and temporal intuition at least as much as via conceptual reasoning. The conceptual reasoning is critical, but there's a classic Bourbakist failure mode where a student learns that you're supposed to reason from axioms and proceeds to completely refuse to do mathematics via any but the logical paradigm of being. This is bad and does not work (maybe for Alonzo Church, but we are a fallen people these days and need to take our insights wherever we find them.)

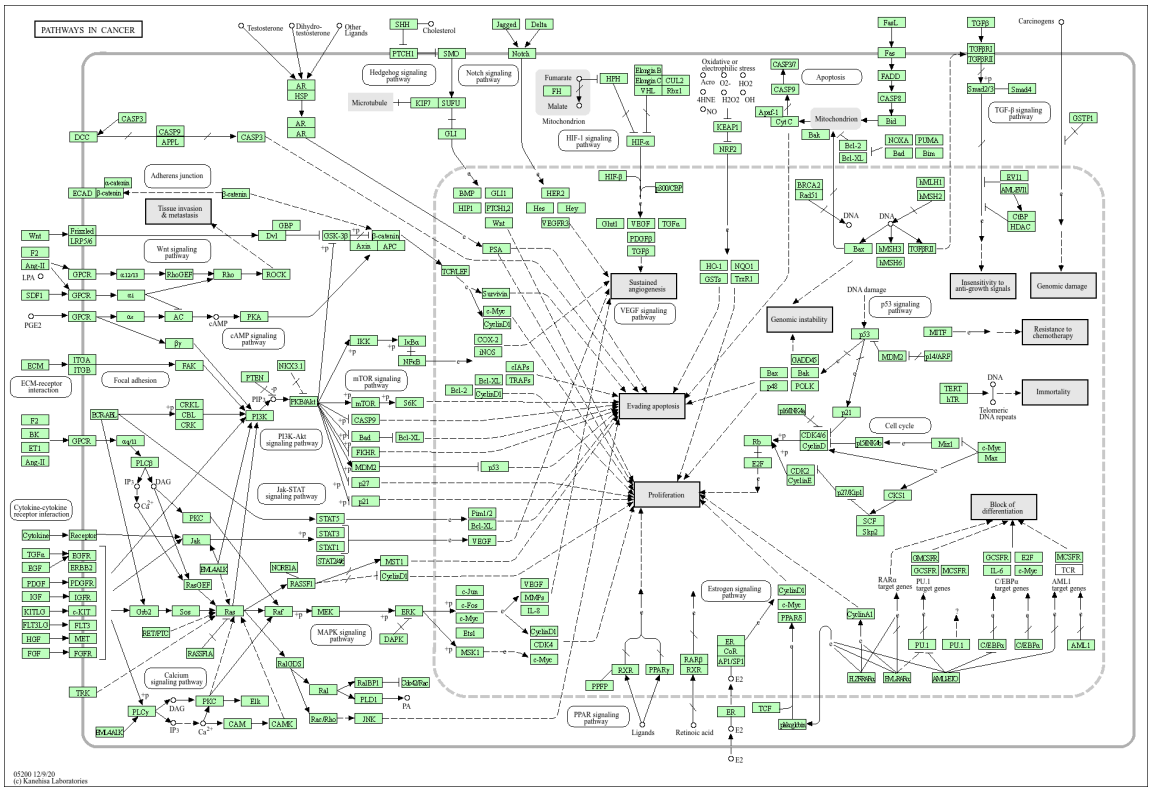
4. While I've worked myself into recapitulating Schopenhauer's whole doctoral thesis, I may as well finish with what he considers the *final* of the four ways in which things are said to be, the four forms of the principle of sufficient ground: the ground of *motive*. This is the connection between his early work and his main work on Will and Representation. If I move my arm, that motion is grounded in the fact that I *decided* to move my arm. Annoying philistines will once again insist that this motion should be given a physical, causal grounding, and I once again promise to not physically smack upside the head any such philistine who can actually successfully reduce my intention to do so to the original neuronal activations.[4]

Models, for real

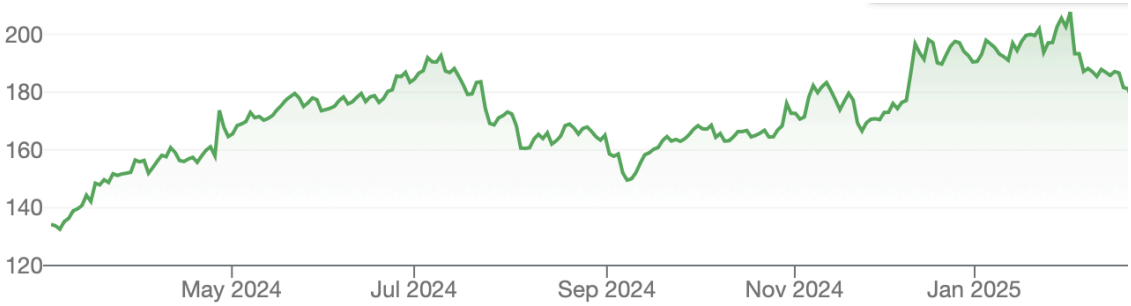
Why am I inflicting this doubtless hilarious but fundamentally straightforward recapitulation of Schopenhauer's thought on you, anyway? Well, I'm interested in world modeling—what it is, how to do it (and with regard to Cosmos readers' interests, whether and how we can or should expect AI to be able to help.) Here are some examples of world models:

- GDP = C+I+G+X
- If GDP' > 0, then all in all the economy is doing well.
- Richmond is next to El Cerrito is next to Berkeley is next to Oakland is next to Hayward is next to Fremont is next to San Jose is next to the Peninsula is next to SF, bridges go from SF to Marin, Marin to Richmond, Oakland to SF, Hayward to the Peninsula, and one more I forget.
- God is love.

[4] To indulge in a bit of psychologizing about these unfortunates, they seem to be subconsciously frightened of a slippery slope from admitting the obvious fact that only a minority of the objects that appear to us as subject are grounded in physicality to, like burning Huguenots at the stake or something. But, guys, God has been dead for longer than the wars of religion ever lasted, let's take a breath and focus our attention where the real threats are, shall we?



- Every employee has a manager, every employee and every manager has a location, and the location of an employee's manager coincides with that employee's location.



-
- The sole ethical obligation of a corporation is to maximize shareholder value.
- There is an olive tree in front of a railing, in front of a large bay of mildly salty water in front of a coastal mountain range.
- The vibes are off with that guy.
- The rich are, by and large, talented, hardworking, and probably way cooler than you.
- The rich are, by and large, greedy, stupid, lucky, and probably ugly.
- The combined oral sensation of bitter, sweet, and ethanol is pleasant.

So, there are lots of kinds of models. A world model is at bottom probably nothing more or less than a representation: a network of objects: a subject organizing (part of) the world to itself. Modernity[5] is the

[5] Modernity in European-descended civilization (I am certain there is an enormous amount of interesting stuff to expand on in this note regarding Islamic civilization, and in various ways in India and China too, but I plead incapacity to even attempt to speak about Asian history) begins to develop very early in the so-called Middle Ages, compared to what you might think based on a superficial historical education. At

state of a civilization that can develop sufficiently precise language and sufficiently intrepid norms of discourse so as to permit a snowballing process of competition between publicly expressed world models that can solve certain problems people have. Postmodernity is the state of a civilization observing that the grand public world models seem, in many ways, to throw the baby (of the flourishing of the individual, the family, the city, the earth) out with the bathwater (of naive adherence to traditional models.)

The fundamental problem of today's civilization is to find a way into what comes next after postmodernity, which teaches us very correctly that legible public models of the good characteristically lead to runaway feedback loops of harm beyond the edges of the model the more sincerely they're adhered to, but then gets mired in the resulting negativity and remains powerless to create any way forward. On an individual level, this is the Kegan stage (as elaborated by David Chapman) of "nihilism" that comes after "systematicity"; some individuals find a way out of the bog of nihilism, but it seems clear to me that the meta-systematic mode is not yet something that our civilization has advanced to at large scale at all. We are perhaps in the relation to meta-modernity that Europeans were to modernity in around 1300; most people have no idea about it, and those that do are forced to communicate in difficult, fuzzy, slow opaque prose about ideas we see only through a glass darkly. All that being said, I insist that it is our task to understand world modeling so much more clearly than our predecessors that we can use not only whole models, but even whole modeling *paradigms*, as fluid tools, formed and modified to suit their particular setting, just as flexibly as the moderns worked within their refined paradigms of reasoned argument from

---

the least there's highly modern intellectual thinking at Merton College, Oxford in the 14th century, while economically, British peasants were using a quantity of metal in their day-to-day lives by the high middle ages that was unprecedented in history anywhere else, as far as I know (I can't remember whether I got this from Brett Devereaux's blog, Lynn White's book, or somewhere else, sorry.) The institution of patenting an invention was well-established from Poland-Lithuania to Ireland by the 16th century, and of course the long slow battle against absolute political power begins, in its European instantiation, somehow even before the battle to *establish* absolute political power, as Magna Carta itself hearkens back to much older tribal traditions. Postmodernity begins no later than the "dark, Satanic mills", though of course one has to beware the pre/trans fallacy; I think the vision of Jerusalem is more than enough to establish Blake as already prepared to *transcend* modernity, for all the premodern aesthetic of the green hills and lambs. (One can't help but note that the vision of building Jerusalem, being classified as a modernism, really suggests Paul as the first modern; I'm sure I'm not the first one to consider this. That produces a rather neat summary of European "modernity" as simply the whole sweep of Christian dominance of the culture, from Justinian to Aquinas to Luther to Jefferson to Luther King, with "postmodernity" the countervailing force of the past centuries; note that Voltaire was a modern *par excellence* and Blake, so unapproachable to his peers, a proto-postmodern, so actually believing in Christianity is only vaguely correlated with submission to this centrality of established Christian norms to European culture.) The point here is something like the old one that the late-modern analogue of the medieval Catholic church was not so much any particular successor *church*, but rather the "church" of reason, liberty, and self-determination founded in the wake of the wars of religion and forming the prime motivation for the lives and work of the bourgeois revolutionaries of the 18th century as well as that of Marx (though certainly somewhere on the line from Trotsky to Lenin to Stalin to Mao to Pol Pot his heirs begin to fundamentally demur from that founding faith) and on, in the American context, to Wilson and Truman and MLK. In the arts, Wagner encompasses the whole tale of the descent from pre- to post-modernism in the Ring, as will as instantiating it in his own intellectual history, developing from a sympathy with naive Feuerbachian "just give the gold back to the Rhinemaidens" pastoral-fantasists proto-socialism to a hope for the intellectual vanguard represented by Wotan and the other Norse gods to lead the people to freedom to, well, the lustful destruction of the world in fire and fury as the breaking of Wotan's spear unravels all law, even that of Fate itself, perhaps returning civilization all the way past the pre-modern state of nature to the protoplasmic original surging chaos of pure Will…So, anyway, these currents coexisted; people interested in progress ought to track the idea over its long history and not slip into the fallacy of starting modernity with the Industrial Revolution, when postmodernity was already brewing, nor of measuring modernity's downfall from the appearance of poets with uncapitalized names.

first principles, describing the physical world using calculus, and assuming that every problem can be solved once somebody invents a sufficiently good new idea.

Schopenhauer on the catwalk

My point in juxtaposing Schopenhauer's metaphysics and the problem of meta-modern world modeling is just this: a world model is a claim, made by some particular subject, about what is to be attended to, which is no different, insofar as there is no object without a subject to attend to it, as a claim about *what is real*. The middlebrow attitude toward modeling is that a model is an approximation to the world ("all models are wrong; some are useful" and all that.) This is wrong and foolish, because *the world is not an object*. There is no "ground truth" to compare models to, not because we are solipsists or subjectivists, but because the world *in itself* is a chaotic surging mess in which *no grounds are admitted*. In the end, "at bottom", things don't happen for reasons, or according to temporal succession, things just *happen*, or maybe *nothing happens*, and in either case you don't want to, and luckily cannot, look too closely at the chaotic primordial ooze of the *Ding an sich*.

The world is not an object. The only objects are our *models* of the world. "Does it follow that a model can never be wrong?" No, of course not, stop asking such oppositional questions. If you have a world model "Vaccines cause autism" then I'm reasonably (though defeasibly) sure that, whatever you mean by that model, I can produce observations that contradict it. But it wouldn't be by "looking at the ground truth", whatever that's supposed to mean; I would observe that your simple, qualitative model is consistent with a certain range of quantitative models, namely of *how many* vaccines cause *how much* autism *how fast*, and prima facie, I'd observe that there's no reasonable choice of parameters making the quantitative dynamics consistent with your qualitative model. You would then propose a more refined model with some kind of exceptions or restrictions showing that your qualitative model is still plausible (I haven't done a debate like this seriously enough to know what you'll say, but I'm sure you've come prepared) and then I'll try to respond, and so on, until presumably we all get exhausted and go home, either both believing ourselves to have given the better model or at least admitting that our beliefs are not always formed from our best models of the truth (actually changing those beliefs in direct response to argument is, of course, generally far too much to ask!)

Note that the standard and usually smartest way for you to respond is *not* by claiming that some fine details of my quantitative model aren't quite right (oh, it's only the vaccines suspended in pure fluoride, the sulfide ones are fine (please don't take my advice on chemistry)) but rather by asserting a dramatically different modeling *paradigm*. A particularly common paradigmatic opposition is that between *anecdotal* and *systematic* evidence. So you say something like "I have three different friends whose babies were perfectly healthy until six months after they got the shots!" The thing with this is, I can calculate probabilities of coincidence at you until I'm blue in the face, I can imply that you or your friends are lying by questioning whether the babies were really originally healthy, but it is *almost impossible* to talk you out of something that you *know* via intimate person-to-person evidence.[6]

[6] I'm using the word "know" advisedly, here: *you* know something, even if I most certainly do not! I'm much more on the "knowledge is a felt sense of certainty" train here than the "justified true belief" train, which I think is ridden only by annoying, nebbish analytic philosophers anyway. My train implies, among other things, that knowledge is *defeasible*! That's self-evident as soon as you ask the question, and a felt-sense definition has the advantage of making it clear that updating knowledge is possible in principle, though of course the problem of just how to get someone we care about (or who's being annoying on the Internet (or who is us)) to defeat knowledge they hold is wide open.

Broadly speaking, distinct modeling paradigms, and most certainly the particular exemplar paradigms of "quantitative-statistical population-scale modeling" and "gathering anecdotes", are *incommensurable*. This is at the heart of the acceleratingly drastic inability of the Western expert class to convince the Western public of much of anything: to take an example to which I'm personally sympathetic, the whole *paradigm* of picking a number to summarize the whole economy, and then asserting that if that number go up, then everyone should have broadly positive affect about the economy, is increasingly rejected by people who find it in various ways unsatisfying and have more access to microphones than they did during the postwar consensus. This is reminiscent of the late James C. Scott's hill tribes, who responded over millennia, all over the globe, to central urban governments' world models of "everyone who lives here permanently is a subject, and subjects are obligated to pay the following taxes annually…" with their world model of "those clumsy city bureaucrats can't find me down in the holler."

It is a traditional belief of the high modern/Enlightened/professional-managerial/expert class that a modeling paradigm is *better* as it better approximates *neutrality, objectivity,* and *quantitativeness.* (It is no doubt purely coincidental that creating models with these properties depends precisely on the technical skills these (we) experts invest so much energy in developing.) This prioritization of legible, quantitative modeling as the "right" way of understanding the world is, when taken beyond its rightful limits (within which it is, again, effective beyond the wildest dreams of all premodern people), a scurrilous and harmful monism closely allied with the nonsensical and harmful monism of materialism.

A story about the Sun

Let me tell a story about Alice and Bob. Some time ago, Alice and Bob, who were in a relationship, were in a discussion about the Millennial trend of wearing daily sunscreen. Bob was broadly opposed, Alice, strongly in favor. Bob entered his PMC-paradigm argument style with aplomb, producing piles of papers assessing the vast majority of skin cancer mortality risk to melanomata arising from sunburns, especially childhood sunburns, with the risk of carcinomas from less intense sun exposure mild enough that overall cancer mortality in sunny places is probably lower overall (Vitamin D probably reduces certain cancer risks, particularly prostate cancer, skewing the facts of this question for men.) Alice was largely in the anecdotal paradigm, describing tragic stories of contacts who tanned and later died, regretting their lifelong carelessness about sunscreen. This discussion was, I'm afraid, mostly stupid, though our friends Alice and Bob are generally of the most thoughtful. Unless Bob could seriously argue that daily sunscreen use was net harmful, which he couldn't (a very moderate amount of unprotected sun exposure in the summer probably keeps your vitamin D levels sufficient all winter; in California, where Alice and Bob live, you can even get useful sun exposure *in* the winter; anyway, while Bob was sorely tempted to argue that Vitamin D supplements have bioavailability problems, that's not a well-grounded point), Bob was trying to argue that a *statistically small chance of dying* meant he *shouldn't take a certain action to avoid it*. Assuming that using sunscreen was basically cost-free, this line of argument would be obviously nonsensical. But it wasn't! What was going on here?

Well, there were parts of both characters' world models they were both avoiding discussing. On the relatively easy side: Alice was worried about wrinkles, Bob, about being pale. These issues are *relatively* easy, because, like…from a saint's-eye view, any iota of risk of death or even serious injury/costly disease isn't worth trading off against issues of appearance, and yet people often push the edges of (especially long-term) safety in, say, exercise habits for the sake of appearance. Besides, seemingly superficial preferences like these can be tied to very deep parts of the psyche: to the value placed on taking care of oneself through middle age, or respectively, to what the Germans call

Intellektuellenweicheierscheinungsangst.[7] More troublesome, Alice's model was resting heavily on the largely-implicit model factor of "if Bob cares about me he'll listen to things I care about without being incredibly annoying about the fine details" whereas Bob's was resting very heavily on the two details "I feel a strong urge to avoid doing things that feel vaguely coerced even if the object-level point is correct" and even deeper in his consciousness, something like "the Sun is Good," coming out of a general vulnerability to the naturalistic fallacy as well as a pure (again, not necessarily consciously endorsed) pagan spirituality. These aspects of their world models are largely getting below the level even of *thoughts and beliefs*, but to *emotions and vibes,* and in particular the last one is essentially in the *Religious* modeling paradigm.

I tell this story at such great length because Alice and Bob told me about it in such intimate detail, not because I'm primarily interested in using progress on world modeling to improve our small-scale interpersonal dynamics[8] but as a reasonably containable test case. The thing is, the *kind* of phenomenon I'm reporting, where one side of a conversation is attempting to communicate via a very legible, objective world-modeling paradigm and another side is working anecdotally, while quietly both are strongly motivated by highly emotional, relational, spiritual, etc. paradigms that may not be fully conscious even to themselves, is not something that's specific to romantic relationships—rather, it's *characteristic* of human disagreements about the world.

For less-intimate instance: someone who's opposed to NAFTA may do their best to produce quantitative general models of why NAFTA was harmful, and certainly there's a lot of hard and serious economic work on both sides of this question. But if they're from pre-Rust Belt western New York, perhaps they're more highly motivated by sorrow for the decline of their childhood home, for friends they've watched spiral into deaths of despair, by the sense of reality that manufacturing work gives in a way that service sector work often does not; certainly there may be less sympathetic emotional and ethical motivations regarding attitudes toward foreigners as well. The point is, we often have this cargo cult of a public conversation going on where we, complicated bundles of stories, feelings, an,beliefsl, drape ourselves in a black cloth and hold out nice, crisp graphs in front of ourselves, hoping if we can mumble sufficiently complexly about how excellently we know how to make number go up then nobody will look under the drape to see all the messy details surrounding the crisp aspects of our world model.

One clear way we'll know if this society has begun to reach the meta-modern stage of its attitude toward world modeling is when we see the sad, angry, hopeful creatures under the cloth start to poke their heads out and explain their real models in all their incommensurability and irrationality, and only *then,* and only when *necessary*, start to look for a point of common ground from which to build a common model for making a decision.

What is to be done?

So, then. People have not really learned to communicate their world models to each other; modernity is (was?) built on a few relatively narrow, if incredibly valuable, exceptions that postmodernity has thoroughly explicated as unsatisfactory. How are we going to get people, on the small scale and on the large, explaining accurately and honestly how they see the world to each other? Well, this is a big problem with many aspects. Some of them, such as getting people to *notice* how they see the world, are

[7] Roughly translatable to "dweeb-appearance-fear".
[8] Not, in turn, because this isn't a profoundly important problem, but moreso because I'm deeply suspicious of attempts to scale techniques for communication with close friends and in intimate partners; most innovations on this front seem as dehumanizing as freeing.

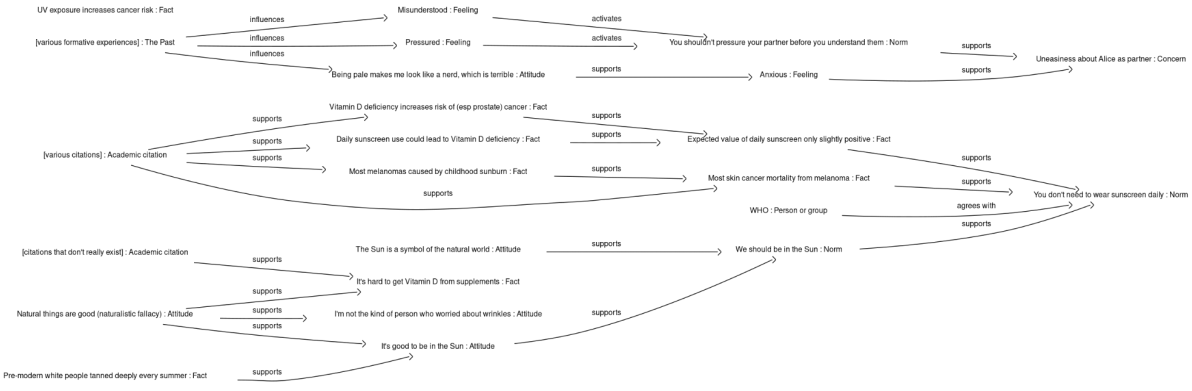
downright spiritual. Possibly you should learn to jhana, or somebody should go found a wonderful new religion. That's not my brief; I am a simple mathematician.

What do I have to offer? Well, if you know anything about 20th-century mathematics, you won't be surprised to hear that I offer *structures*. Structures for your models: organizing them, translating them, putting them together, analyzing them, comparing them, and so on, structures galore, beyond your wildest imaginings! In the earlier example of SPFGate, I imagine that the conversation would have gone much better could Alice and Bob have made explicit their sufficiently complete world models, visualizable broadly as shown below. With these models in place, for instance, Alice could immediately check and see (perhaps even assisted by software) that Bob was not, in fact, denying that, in general, UV radiation causes skin cancer; both could see the heavy, deep emotional sensitive points they were approaching and focus there first as opposed to debating the fine factual details which would become so much less threatening afterwards.

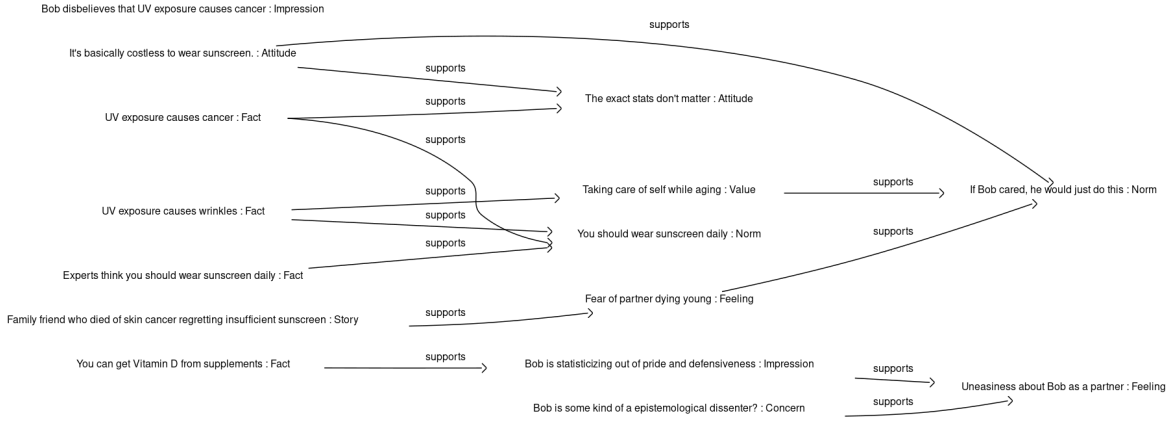
Of course, constructing these models would require some iteration, notably for Bob to zoom out from the "Calculate a confidence interval"-type paradigm I initially hyperfocused on. This iterative process could perhaps be purely dialectical between the discussants, but I'm sure you're wincing terribly at the idea of pausing an intimate and important conversation repeatedly to modify a world model, perhaps to migrate the model to a more complex paradigm, to simulate or compute something, debug a bit that doesn't parse, etc…My hope is that there's a world in which we can accomplish this kind of vision *simultaneously* with speaking to each other like humans, and indeed, even more like humans than we most often do without such technical aids.

How? On the one hand, AI may make the process of model tracking more about *checking* the model that's tracking your conversation and modifying it only when you notice you've seemed to say something different from what you wanted to; on the other, while I've illustrated these ideas with a quotidian one-on-one example, naturally in a negotiation on a major business or policy decision, or a discussion of a complex scientific or engineering model, there's more tolerance for speaking with technological supports if it leads to more valuable outcomes. All that said, you can click the links to the models for our current attempt-in-progress at realizing this vision, via the CatColab software being built with my colleagues at the Topos Institute. (I'm sorry these models are a bit hard to view, here! They're scrollable in the app.)

Bob's world model (click hamburger menu-> new analysis -> visualization to view graph scrollably):



*Alice's world [model](): *



This example hopefully does something to illustrate the possible fruitfulness of this paradigm (of, in short, *making explicit what is real to you in a situation*) in terms of clarity and efficacy of communication. Arguably, you don't need very much fancy math stuff to make a system capable of the above; anybody can build software for drawing and visualizing graphs in real-time in a small group on a reasonably constrained set of variations of what "graph" is taken to mean. But please follow me that this is both a *potted* and an *early* example. What I'd really like to show, did I have such a one worked out, is an example like "design a jet engine, predict ahead of time to the extent possible where the main decision points will be and what the testing process has to be, and maintain it and all institutional memory of its nature through its full life-cycle." Or, "design a zoning policy for Berkeley." Or…."design a new Constitution for the United States?" Problems like these are much larger-scale. They have, or should have, participants coming from wider or narrower perspectives and at many levels of technical sophistication. I want a world where *everyone* with a stake in a problem like "design a zoning policy for Berkeley" can potentially put their two cents (or two dollars) in. There are many obviously-intractable challenges to this vision. For instance:

1. Every participatory-democracy-flavored plan is impossible, because *normal people would rather hang out with their friends than specify their policy preferences*.[9]
2. Even if they would provide it, some people's world models here are just "Berkeley is nice and should never change". Others have something like "I have the following quantitative model of housing prices and availability of home health aides in Berkeley in 2035 under the following parameterized housing growth scenarios…" You obviously can't funge these models, right?
3. There are *lots* of pieces of input to collate: at least from everyone currently in Berkeley, plus from many outsiders, those in nearby cities who work in Berkeley, those who might like to live in Berkeley one day, those who have opinions about the University of California system, those who have opinions about housing policy or Bay Area dynamism in general, and so on.

You can think of more.

[9] You, dear reader, no doubt spend much of your time hanging out with your friends specifying policy preferences; but you aren't normal people!

There are some easy-ish partial solutions to 1 and 3. For a flyer, try sortition! Bring in obligatory but decently-compensated, limited-commitment jury duty-style participation.[10] Sortition largely solves 1 and substantially solves 3, so I think the biggest problem that needs solving to open up a new era of participatory, effective policy-making is 2.

How do you funge the infungible, commensurate the incommensurable? The main pre-existing options to answer this question, which after all has to be answered at-least-implicitly for any policy to ever get chosen by a group, are

1. Establish, more or less from the top down, an acceptable modeling paradigm in which to communicate preferences. Ideally this paradigm eventually produces a simple scalar rating of quality for each choice, and then we just choose the choice with the biggest number and pat ourselves on the back for being so optimal.

2. Let anybody who can get into the metaphorical room participate in plain text, which is recombined in nondeterministic and unpredictable ways in the minds of the final decision-maker(s) until they make their decision.

You're more than clever enough to see the problems with these, but just as long as I'm clearly not writing to a word limit, let's spell it out for common knowledge: choice 1…

- Rules out participation of anyone who can't (learn to) express themselves in the accepted language, and of any *ideas* that can't be well-expressed in that language. (Consider how well "the Sun is kind of a god" fits into a paradigm of "statistical mortality models".) Those who are excluded from participation are both failing to have their inherent beauty as instantiations of the universe's desire to regard itself respected, and also sometimes construct guillotines.
- Tends to lead to the clever people who can use the fancy language fluently being wildly overconfident that there can't possibly be an unknown unknown *crouching with bared teeth* RIGHT BEHIND THEM! (q.v. guillotines)
- Admits no consensus mechanism to update the approved language with the times: *pace* Jefferson's 19 years, people are still being governed in part by the world models of the American founders, and indeed of the English common law.
- Leads, in many instantiations, to lots of numbers which are largely made up and/or irrelevant to the real point, which is sometimes bad—if they turn out to be wrong, then certainly, but even more often if they create the false confidence of "I have a confidence interval, all you have is the word 'bad'".

As for option 2, well…the whole *image* of modernity is that we're totally not determining the course of civilization by some stochastic process determining who yells loudest as near as possible to the leader as close as possible to the moment of decision, right?

Of course, both have advantages: option 1 gives everyone a common standard to refer to, if only they can learn the standard language (see: the common law, the calculus), and those common standards are once again at the heart of everything that worked about modernity. Option 2 is cool because it turns out that,

---

[10] The voluntary structure that's more-or-less what we have now incentivizes participation people who are boring and/or extreme and/or have nothing better to do with their time than yell at city council members, and it would be *harder* to get voluntary participation in a more structured system, since all anybody really wants to do is stand up in front of a mike and speechify about how idiotically their city council has been proceeding. This "community participation != democracy" point is widely made in YIMBYish circles.

almost tautologically, anything that needs to be expressed can be expressed in natural language, and it's very strong in some information-theoretic, Austrian economics sense to leave the channels for information flow as wide (in terms of communication style and of permitted participants) as possible.

So we want to take the best of both sides: low barrier to entry but high legibility and reliability. My suggestion is, essentially, not to pick a single modeling language but to try to pick a single way of *constructing* modeling languages, which is itself formal enough to describe partial translations between different approaches in a precise enough way that you can try to scale. The best math yet seen for this is what we're implementing in CatColab.

Now, changing modeling paradigms is tough: you don't know what to add and when: your universe gets more complicated: you have to do some thinking and some drudgery: altogether, it's most often in the "nobody's gonna do this for themselves" attractor I mentioned above. This is (finally!) a great application for AI. Neural nets are pretty good at scanning a big space and showing you some candidate jumps to make in it. If every candidate has to be expressed in a well-specified language, that solves for some proportion of the AI slop problem. And if the actual human trying to develop a world model is the one picking from candidate refinements, that solves for another proportion. "Is AI critical here?" The AI is probably load-bearing here in a pragmatic sense, but not in a principled sense: it's trying to draw out from the user what they "really think" is real, to the extent that "draw out" rather than "co-construct" is the right metaphor, the same thing that, given enough patience, they might do by themselves.

Robot Schopenhauer (still on the catwalk)

Does the AI have its own world model? Put in Schopenhauer's terms, is there a subject regarding the objects that an AI communicates about? Thus far, to whatever extent that such subjects *experience*, they seem to be superpositions of simulacra, averaged out of the mere written records of numerous human subjects. They see as real, modulo such details as prompting, very much what the median author of long-form content available online in 2025 sees as real.

This derivative quality of the subjects that may be experiencing on the inside of a language model—and to be clear, most humans are ourselves derivative most-to-all of the time, and what powerful derivations some of these language models make!—is one good reason to emphasize the AI-as-*amanuensis* image from above, as opposed to an image in which, for instance, humans are dialoging directly with the AI's world models, to whatever extent they have them, and to whatever extent there is a "they" there *to* have them. Indeed, when we are trying to gather world models in support of understanding some problem, we want to allow for as radical an independence of the models collected as is permitted by the range of human thought and feeling and the possibility of communication across the resulting inferential gaps, rather than to prematurely converge on a model which will tend to bias toward legible rationality and educated consensus; besides, or in other words, if a certain perspective really ought to be part of a conversation, there ought to be a human available to represent that perspective directly.

Why are the AIs derivative in this way? Well, reverse the question: why are humans *original*, sometimes? One fundamental reason seems to be that we have an apparently-unique ability to make analogies across realms of being. We can observe physical objects knocking into each other and start drawing parabolas; we can pass from symbolico-logical *reasoning* about the one-point compactification of the natural numbers to the simulated spatiotemporal *experience* of of a geometrically accelerating endless succession of points accumulating at infinity; we can define the number of paths through a graph in purely formal, apsychological ways and yet use our faculty of *choice* to, in a real sense, *actually take* all the different paths.

Something about this fruitful interaction among the various modes of representation is at the heart of how a human chess or go master plays—in particular, at the heart of how her play is different than that of an AI master—via partly-reflectively-opaque spatial intuition of the quality of a board arrangement. Similarly, my understanding of the ability of human mathematicians to make leaps of Taoian "post-rigorous" insight has to do with the ability to pass from formal-linguistic definitions of mathematical structures to physical-spatio-temporal representations in which, in good cases, truths about these structures can be merely *perceived*, rather than *deduced*, deduction being a process that few mathematicians make much use of except in communicating knowledge already established for themselves in another manner.

This is all, perhaps, interesting enough but I may be making a finer point than is needed, here: for Schopenhauer, a language model is a rather unprecedented form of being that is grounded entirely in *concepts*, and entirely lacks *intuitions*.[11] I want to claim that acts of profound human creativity arise from discoveries of how to translate an intuitive experience one has had—often a primarily emotional experience for artistic creation, but Newton's apple and Poincaré's omnibus also apply here—into a more stabilized form which a range of other people can then relate useful to their own intuitions. In science, mathematics, literature, these stabilized forms are largely conceptual, but it's interesting to consider that, for instance, a piece of music is *not* built out of concepts, but rather of sounds in temporal succession, perhaps a "sculpture built out of time," insofar as sounds themselves are our intuitive experience of sufficiently high-frequency periodic phenomena.

So, perhaps the current AIs are lacking in that they have no access to the lower humus of intuitive experience, having to rely on recombination of already-refined and -abstracted conceptual communication for their best attempts at creativity. All of this points, again, to the imperative that AIs must be used to help clarify and refine the world models of *people*, and not that people may be pressed to align their world models to those generated by algorithms, which to date can only vary their models within a legible and previously-expressed space, leaving a terrible risk of a second round of the failures of modernism, where a lifeless legibility is imposed not only on the functions of the state but on interpersonal relationships and even, dreadful thought, on the interiority of each individual. Such a loss would be horrifically dystopian, but glimmers of this future are both already clearly present and clearly consistent with AI that never achieves the vaunted status of ASI or even AGI, indeed, of AI that remains at the very level of capability we've already reached. Let us confidently stand for the preservation and enrichment of the ability of the individual and the community to see the world clearly and express their vision to the rest of the world, enhanced as possible by the ghost of that world as instantiated in these new technologies, but never permitting the intuitive, the specific, the embodied, emotional, personal, willful, and true to lie subservient to that ghostly hand.

[11] At least mostly! Maybe one could imagine that the sequentiality of prompt processing structures some intuitive temporal experience for a language model.