A revolution in philosophy: the rise of conceptual engineering

130

by **Suspended Reason** 2nd Jun 2020

Philosophy of Language | Meta-Philosophy | Rationality | World Modeling | Frontpage

Almost a decade ago, Luke Muehlhauser ran a series "Rationality and Philosophy" on LessWrong 1.0. It gives a good introductory account, but recently, still dissatisfied with the treatment of the two groups' relationship, I've started a larger "Meta-Sequence" project, so to speak, treating the subject in depth.

As part of that larger project, I want to introduce a frame that, to my knowledge, hasn't yet been discussed to any meaningful extent on this board: *conceptual engineering*, and its role as a solution to the problems of "counterexample philosophy" and "conceptual analysis"—the mistaken if implicit belief that concepts have "necessary and sufficient" conditions—in other words, Platonic *essences.* As Yudkowsky has argued extensively in "Human's Guide to Words°," this is *not* how concepts work. But he's far from alone in advancing this argument, which has in recent decades become a rallying cry for a meaningful corner of philosophy.

I'll begin with a history of concepts and conceptual analysis, which I hope will present a productively new frame, for many here, through which to view the history of philosophy. (Why it was, indeed, a "diseased discipline°"—and how it's healing itself.) Then I'll walk through a recent talk by Dave Chalmers (paper if you prefer reading) on conceptual engineering, using it as a pretense for exploring a cluster of pertinent ideas. Let me suggest an alternative title for Dave's talk in advance: "How to reintroduce all the bad habits we were trying to purge in the first place." As you'll see, I pick on Dave pretty heavily, partly because I think the way he uses words (e.g. in his work with Andy Clark on embodiment) is reckless and irresponsible, partly because he occupies such a prominent place in the field.

Conceptual engineering is a crucial moment of development for philosophy—a paradigm shift after 2500 years of bad praxis, reification fallacies, magical thinking, religious "essences," and linguistic misunderstandings. (Blame the early Christians, whose ideological leanings lead to a triumph of Platonism over the Sophists.) Bad linguistic foundations give rise to compounded confusion, so it's important to get this right from the start. Raised in the old guard, Chalmers doesn't understand why conceptual engineering (CE) is needed, or the bigger disciplinary shift CE might represent.

How did we get here? A history of concepts

I'll kick things off with a description of human intelligence from Jeurgen Schmidhuber, to help ground some of the vocabulary I'll be using in the place of (less useful) concepts from the philosophical traditions:

> As we interact with the world to achieve goals, we are constructing internal models of the world, predicting and thus partially compressing the data history we are observing. If the predictor/compressor is a biological or artificial recurrent neural network (RNN), it will automatically create feature hierarchies, lower level neurons corresponding to simple feature detectors similar to those found in human brains, higher layer neurons typically corresponding to more abstract features, but fine-grained where necessary. Like any good compressor, the RNN will learn to identify shared regularities among different already existing internal data structures, and generate prototype encodings (across neuron populations) or symbols for frequently occurring observation sub-sequences, to shrink the storage space needed for the whole (we see this in our artificial RNNs all the time).

The important takeaway is that CogSci's current best guess about human intelligence, a guess popularly known as *predictive processing*, theorizes that the brain is a machine for detecting regularities in the world—think similarities of property or effect, rhythms in the sense of sequence, conjunction e.g. temporal or spatial—and compressing them. These compressions underpin the daily probabilistic and inferential work we think of as the very basis of our intelligence. Concepts play an important role in this process, they are bundles of regularities tied together by family resemblance, collections of varyingly held properties or traits which are united in some instrumentally useful way which justifies the unification. When we attach word-handles to these bundled concepts, in order to wield them, it is frequently though not always for the purpose of communicating our concepts with others, and the synchronization of these bundles across decentralized speakers, while necessary to communicate, inevitably makes them a messy bundle of overlapping and inconsistent senses—they are "fuzzy," or "inconsistent," or "polysemous."

For a while, arguably until Wittgenstein, philosophy had what is now called a "classical account" of concepts as consisting of "sufficient and necessary" conditions. In the tradition of Socratic dialogues, philosophers "aprioristically" reasoned from their proverbial armchairs (Bishop 1992: The Possibility of Conceptual Clarity in Philosophy's words—not mine) about the definitions or criteria of these concepts, trying to formulate elegant factorings that were nonetheless robust to counterexample. Counterexample challenges to a proposed definition or set of criteria took the form of presenting a situation which, so the challenger reasoned, intuitively seemed to *not* be a case of the concept under consideration, despite fitting the proposed factoring. (Or of course, the inverse—a case which intuitively seemed like a member but did not fit the proposed criteria. Intuitive to *whom* is one pertinent question among many.)

The legitimacy of this mode of inquiry depended on there being necessary and sufficient criteria for concepts; if such a challenge was enough to send the proposing philosopher back to the drawing board, it had to be assumed that a properly factored concept would deflect any such attacks. Once the correct and elegant definition was found, there was no possible member (*extension*) which could fit the criteria but not feel intuitively like a member, nor was there an intuitive member which did not fit the criteria.

Broadly construed I believe it fair to call this style of philosophy *conceptual analysis* (CA). The term is established as an organizing praxis of 20th century analytic philosophy, but, despite meaningful differences between Platonic philosophy and this analytic practice, I will argue that there is a meaningful through-line between them. While the analytics may not have believed in a "form" of the good, or the pious, which exists "out there," they did, nonetheless, broadly believe that there were sufficient and necessary conditions for concepts—that there was a very simple-to-describe (if hard-to-discover) pattern or logic behind all members of a concept's extension, which formed the goal of analysis. This does, implicitly, pledge allegiance to some form of "reality in the world" of the concept, its having a meaningful structure or regularity in the world. While this may be the case at the *beginning* of a concept's lifespan, entropy has quickly ratched by early childhood: stretching, metaphorical reapplication & generalization, the over-specification of coinciding properties.

(The history I'm arguing might be less-than-charitable to conceptual analysis: Jon Livengood, philosopher out of Urbana-Champaign and member of the old LessWrong, made strong points in conversation for CA's comparative merits over predecessors—points I hope to publish in a forthcoming post.)

But you can ignore my argument and just take it from the *SEP*, which if nothing else can be relied on for providing the more-or-less uncontroversial take: "Paradigmatic conceptual analyses offer definitions of concepts that are to be tested against potential counterexamples that are identified via thought experiments... Many take [it] to be the essence of philosophy..." (Margolis & Laurence 2019). Such comments are littered throughout contemporary philosophical literature.

As can be inferred from the juxtaposition of the Schmidhuber-influenced cognitive-scientific description of concepts, above, with the classical account, conception of concepts, and their character, was meaningfully wrong. Wittgenstein's 1953 *Investigations* inspired Eleanor Rosch's Prototype Theory which, along with the concept "fuzzy concepts," and the support of developmental psychology, began pushing back on the classical account. Counterexample philosophy, which rested on an unfounded faith in intuition plus this malformed "sufficient and necessary" factoring of concepts, is a secondary casualty in-progress. The traditional method for problematizing, or disproving, philosophical accountings of concepts is losing credibility in the discourse as we speak; it has been perhaps the biggest paradigm shift in the field since its beginning in the 1970s.

This brings us up to our current state: a nascent field of conceptual engineering, with its origins in papers from the 1990s by Creath, Bishop, Ramsey, Blackburn, Graham, Horgan, and more. Many, though far from all, in analytic have given up on classical analysis since the late 20th C fall. A few approaches have taken their place, like experimental conceptual analysis or "empirical lexicography" à la Southern Fundamentalists, where competent language speakers are polled about how they use concepts. While these projects continue the descriptive bent of analysis, they shift the method of inquiry from aprioristic to empirical, and no longer chase their tail after elegant, robust, complete descriptions. Other strategies are more prescriptive, such as the realm of conceptual engineering, where philosophers are today more alert to the discretionary, lexicographic nature of the work they are attending to, and are broadly intentional within that space. Current work includes attempting to figure out valid grounds by which to judge the quality of a "conceptual re-engineering" (i.e. reformulation, casually used—re-carving up the world, or changing "ownership rights" to different extensions). The discourse is young; the first steps are establishing what this strategy even consists of.

Chalmers is in this last camp, trying test out conceptual engineering by applying it to the concept "conceptual engineering." How about we start *here*, he says—how about we start by testing the concept on itself.

He flails from the gate.

Back to the text

The problem is that Chalmers doesn't understand what "engineering" is, despite spending the opening of his lecture giving definitions of it. No, that's not quite right: ironically, it is Chalmers's inquiry into the definition of "engineering" which demonstrates his lack of understanding a to what the approach entails, dooming him to repeating the problems of philosophies past. Let me try to explain.

Chalmers:

> What is conceptual engineering? There is an obvious way to come at this. To find the definition of conceptual engineering, go look up the definition of engineering, and then just appeal to compositionality.

At first blow this seems like a joke, indeed it's delivered as a joke, but it is, Chalmers assures us, the method he actually used. Based on a casual survey of "different engineering associations" and various "definitions of engineering on the web," he distills engineering to the (elegant and aspiring-robust) "designing, building, and analyzing." Then he tweaks some words that are already overburdened—"analyze" is already taken when it comes to concepts (That's what we're trying to get away from, remember? Conceptual analysis) so he substitutes "evaluate" for "analyze." And maybe, he writes, "implementing" is better than "building." So we wind up with: *conceptual engineering is designing, implementing, and evaluating concepts.*

This doesn't seem like a bad definition, you protest, and it isn't. But we were never *looking* for a definition. That's the realm of conceptual analysis. We quit that shit alongside nicotine, back in the 80s. Alright, so what *are* we trying to do? We're trying to solve a problem, multiple problems actually. The original problem was that we had concepts like "meaning" and "belief" that, in folk usage, were vague, or didn't formalize cleanly, and philosophers quite reasonably intuited that, in order to communicate and make true statements about these concepts, we first had to know what they "were." (The *"is"* verb implies a usage mission: *description* over *prescription.*) The problem we are trying to solve is, itself, in part, conceptual analysis—plus the problems conceptual analysis tried originally to solve but instead largely exacerbated.

This, not incidentally, is how an engineer approaches the world, how an engineer would approach writing Chalmers's lecture. Engineers see a problem and then they *design* a solution that *fits* the current state of things (context, constraints, affordances) to *bring about the desired state* of affairs.

Chalmers is just an analyst, and he can only regurgitate definitions like his analyst forbearers. Indeed what is Chalmers actually figuring out, when he consults the definition of "engineering"? In 1999 Simon Blackburn proposes the term "conceptual engineering" as a description of what he's up to, as a philosopher. He goes on to use it several times in the text (*Think: A Compelling Introduction to Philosophy*), typically to mean something like "reflecting":

> We might wonder whether what we say is "objectively" true, or merely the outcome of our own perspective, or our own "take" on a situation. Thinking about this we confront categories like knowledge, objectivity, truth, and we may want to think about them. At that point we are *reflecting* on concepts and procedures and beliefs that we normally just *use*. We are looking at the scaffolding of our thought, and doing conceptual engineering.

For reasons still opaque to me, the usage becomes tied up with the larger post-CA discourse. To understand what's going on in this larger discourse, or to understand what this larger discourse *ought* to be up to, Chalmers reverse-engineers the naming. In trying to figure out what our solutions should be to a problem, Chalmers can only do as well as Blackburn's metaphorical appropriation of "engineering" fits the problem and solution in the first place. The inquiry is hopelessly mediated by precedent once again. (For future brevity, I'll call conceptual engineering a *style of solution,* or "strategy": a sense or method of approaching a problem.)

Let me try to be more clear: If the name of the strategy had been "conceptual ethics," or "conceptual revision," or "post-analytic metaphilosophy" (all real, rival terms) Chalmers's factoring of the strategy would be substantially different, even as the problem remained exactly the same. Once again, a handle has been reified.

Admittedly, the convergence of many philosophers in organizing around this term, "conceptual engineering," tells us that there is *something* in it which is aligned with the individual actors' missions—but the amount of historical chance and non-problem-related reasons for its selection obfuscates our sense of the problem instead of clarifying it.

Let us not ask, "What is the definition of the strategy we wish to design, so we may know how to design it?" Let us ask, "What is the problem, so that we can design the strategy to fit it?" *This* is engineering.

De novo & re-engineering

Chalmers:

> So I encourage making a distinction between what I call *de novo* engineering and re-engineering. De novo engineering is building a new bridge, program, concept, whatever. Re-engineering is fixing or replacing an old bridge, program, concept, or whatever. The name is still up for grabs. At one point I was using de novo versus de vetero, but someone pointed out to me that wasn't really proper Latin. It's not totally straightforward to draw the distinction. There are some hard cases. Here's the Tappan Zee Bridge, just up the Hudson River from here. The old Tappan Zee bridge is still there, and they're building a new bridge in the same location as the old bridge, in order to replace the old bridge. Is that de novo because it's a new bridge, or is it re-engineering because it's a replacement?

Remember: the insight of a metaphor is a product of its analogic correspondence. This is not the "ship of Theseus" it seems.

If we were to build an exact replica of the old bridge, in the same spot, would it be a new bridge, or the same bridge? You're frustrated by this question for good reason; it's ungrounded; it can't be answered due to ambiguity & purposelessness. *New in what way? Same in what way? Certainly most of the properties are the same, with the exception of externalist characteristics like "date of erection." The bridge has the same number of lanes. It connects the same two towns on the river.*

De novo, as I take it from Chalmers's lecture, is about capturing phenomena (noticing regularity, giving that regularity a handle), whereas re-engineering involves refactoring existing handle-phenomena pairs either by changing the assignments of handles or altering the family resemblance of regularities a handle is attached to. Refactorings are functional: we change a definition because it has real, meaningful differences. These changes are not just "replacing bricks with bricks." They're more akin to adding a bike lane or on-ramp, to added stability or a stoplight for staggering crossing.

Why do I nitpick a metaphor? Because the cognitive tendency it exhibits is characteristic of philosophy at its worst: getting stuck up on distinctions that don't matter for those that do. If philosophers formed a union, it might matter whether a concept was "technically new" or "technically old" insofar as these things correlate with the necessary (re)construction labor. Here, what matters is changing the *function* of concepts: what territories they connect, and which roads they flow from and into; whether they allow cars or just pedestrians. "Re-engineering" an old concept such that it has the same extensions and intensions as before doesn't even make sense as a project.

Abstracting, distinguishing, and usefulness

At this point, we have an understanding of what concepts are, and of the problems with concepts (we need to "hammer down" what a concept is if we want to be able to say meaningful things about it). It's worth exploring a bit more, though, what we would want from conceptual engineering—its commission, so to speak—as well as qualities of concepts which make them so hard to wield.

Each concept in our folk vocabulary has a use. If a concept did not have a use, if it was not a regularity which individuals encountered in their lives, it would not be used, and it would fall out of our conceptual systems. There is a Darwinian mechanism which ensures this usefulness. The important question is, what *kind* of use, and at what *scale*?

For a prospective vegetable gardener shopping at a garden supply store, there is a clear distinction between *clay-based soil* and *sand-based soil*. They drain and hold water differently, something of significant consequence for the behavior of a gardener. But whether the soil is light brown or dark brown likely matters very little to him, we can suppose he makes no distinction.

However, for a community of land artists, who make visual works with earth and soil, coloration matters quite a bit. Perhaps this community has evolved different terms for the soil types just like the gardeners, but unlike the gardeners may make no distinction between the composition of the soil (clay or sand) beyond any correspondences with color.

A silly example that illustrates: concepts *by design* cover up some nuanced differences between members of its set, while *highlighting* or bringing other differences to the fore. The first law of metaphysics: no two things are identical, not even two composites with identical subatomic particle makeups, for at the very least, these things differ in their locations in spacetime; their particles are not the same particles, even if the types are. Thus things are and can only be the same in *senses*. There is a smooth gradient between analogy and what we call equivalence, a gradient formed by the number of shared senses. We create our concepts around the distinctions that matter, for us as a community; and we do so with a minimum of entropy, leaving alone those distinctions that do not. This is well-accepted in classification, but has not as fully permeated the discourse around concepts as one might wish. (Concepts and categories are, similarly, pairings of "handles" or designators with useful-to-compress regularities.)

Bundling & unbundling

In everyday life, the concept of "sound" is both phenomenological experience and physical wave. The two are bundled up, except when we appeal to "hearing things" (noises, voices) when there is a phenomenological experience without an instigating wave. But there is never a situation which concerns us in which waves exist without *any phenomenological experience whatsoever.* Waves without phenomenology—how does that concern us? Why ought our conceptual language accommodate that which by definition has nothing to do with human life, when the function of this language is describing human life?

Thus the falling tree in the empty forest predictably confounds the non-technical among us. The solution to its dilemma is recognizing that the concept (here a folk concept of "sound") bundles, or conflates, two patterns of phenomena whose *unbundling*, or distinction, is the central premise (or "problem") of the paradox. Scientists find the empty forest problem to be a non-problem, as they long ago performed a "narrow-and-conquer" method (more soon) on the phenomenon "sound": sound is sound waves, nothing more, and phenomenological experience is merely a consequence of these waves' interaction with receiving instruments (ears, brains). They may be right that the falling tree obviously meets the narrowed or unbundled scientific criteria for sound—but it does *not* meet the bundled, folk sense.

(Similarly, imagine the clay-based soil is always dark, and sand-based soil always light. Both the gardeners and land artists call dark, clay-based soil *D1* and light sand-based soil *D2*. If asked, *"Is dirt that is light-colored, but clay-based, D1 or D2?"* the gardners and land artists would ostensibly come to exact opposite intuitions.)

All this is to say that concepts are bundled at the level of maximum abstraction that's useful. Sometimes, a group of individuals realizes this level of abstraction covers up differences in class members which are important to separate; they "unbundle" the concept into two. (This is how the "empty forest" problem is solved: *sound as waves* and *sound as experience.*) I have called this the "divide and conquer" method, and endorse it for a million reasons, of which I'll soon name a fistful. Other times, a field will claim their singular sense (or sub-sense, really), which they have separated from the bundled folk whole, is the "true" meaning of the term. In their domain, for their purposes, it might be, but such claims cause issues down the line.

The polysemy of handles

In adults, concepts are generally picked up & acquired in a particular manner, one version of which I will try to describe.

In the beginning, there is a word. It is used in conversation, perhaps with a professor, or a school teacher, a parent—better, a friend; even better, one to whom we aspire—one whom we want, in a sense, to become, which requires knowing what they know, seeing how they see. Perhaps on hearing the word we nod agreement, or (rarer) confess to not knowing the term. If the former, perhaps its meaning can be gleaned through content, perhaps we look it up or phone a friend.

But whatever linguistic definition we get will become meaningful only through correspondence with our lived reality—our past observations of phenomena—and through coherence with other concepts and ideas in our conceptual schema. Thus the concept stretches as we acquire it. We convert our concepts as much as our concepts convert us: we stretch them to "fit" our experiences, to fit what previously may have been a vaguely felt but unarticulated pattern, now rapidly crystallizing, and this discovery of a concept & its connection with other concepts further crystallizes it, distorts our perception in turn with its sense of *thingness*; the concept begins to stretch our experience of reality. (This is the realm of Baader-Meinhof & weak Sapir-Whorf.)

When we need to describe something which feels adjacent to the concept as we understand it, and lack any comparatively better option, we will typically rely on the concept handle, perhaps with qualifications. Others around us may pick up on the expansion of territory, and consider the new territory deservingly, appropriately settled. Lakoff details this process with respect to metaphor: our understanding of concreta helped give rise to our abstract concepts, by providing us a metaphorical language and framework to begin describing abstract domain.

Or perhaps we go the other way, see a pattern of coinciding properties which go beyond the original formulation but in our realm of experience, seem integral to the originally formulated pattern, and so we add these specifications. One realm we see this kind of phenomenon is racial stereotyping. Something much like this also happened with Prototype Theory, which was abandoned in large part out of an opposition to its *empirical bent*—a bent which was never an integral part of the theory, but merely one common way it was applied in the 70s.

All of this—the decentralization, the historical ledger, the differing experiences and contexts of speakers, the metaphorical adaptation of existing handles to new, adjacent domains—leads to fuzziness and polysemy, the accumulation of useful garbage around a concept. Fuzziness is well-established in philosophy, polysemy well-established in semantics, but the discourses affected by their implications haven't all caught on. By the time a concept becomes entrenched in discourse, it describes not one but many regularities, grouped—you guessed it—by family resemblance. "Some members of a family share eye color, others share nose shape, and others share an aversion to cilantro, but there is no one single quality common to all" (Perry 2018).

Lessons for would-be engineers

The broader point I wish to impart is that we do not need to "fix" language, since the folk concepts we have are both already incredibly useful (having survived this long) and also being constantly organically re-engineered on their own to keep pace with changing cultures, by decentralized locals performing the task far better than any "language expert" or "philosopher" could. Rather, philosophy must *fit* this existing language to its own purposes, just as every other subcommunity (gardeners, land artists...) has done: determine the right level of abstraction, the right captured regularities and right distinction of differences for the problem at hand. We will need to be very specific and atomic with some patterns, and it will behoove us to be broad with others, covering up what for us would be pointless and distracting nuance.

Whenever we say two things are alike in some sense, we say there is a hypothetical hypernym which includes both of them as instances (or "versions"). And we open the possibility that this hypernym is meaningful, which is to say, of use.

Similarly, for every pair of things we say are alike in some sense, there will also necessarily be difference in another sense—in other words, these things could be meaningfully distinguished as *separate* concepts. If any concept can be split, and if any two instances can be part of a shared concept, then why do the concepts we have exist, and not other concepts? This is the most important question for us, and the answer, whatever it turns out to be, will have something to do with *use*.

Once again we have stumbled upon our original insight. The very first question we must ask, to understand what any concept ought to be, is to understand what problem we are trying to solve, what the concept—the set of groupings & distinctions—accomplishes. The concept "conceptual engineering" is merely one, and arguably the first, concept we should factor, but we cannot be totally determinate in our factoring of it: its approach will always be contingent on the specific concept it engineers, since that concept exists to solve a unique problem, i.e. has a unique function. Indeed, that might be all we can say—and so I'll make my own stab at what "conceptual engineering" ought to mean: the re-mapping of a portion of territory such that the map will be more useful with respect to the circumstances of our need.

E-belief: a case study in linguistic malpractice

Back in the 90s, Clark and Chalmers defined an *extended belief*—e.g. a belief that was written in a notebook, forgotten, and referenced as a source of personal authority on the matter—as a belief proper. It is interesting to note that this claim takes the inverse form of traditional "counterexample philosophy" arguments: *despite native speakers not intuitively extending the concept "belief" to include e-belief, we advocate for it nonetheless.*

Clark thinks the factoring is useful to cognitive science; Chalmers thinks it's "fun." The real question is *Why* didn't *they call it e-belief?* which is a question very difficult to answer for any single case, but more tractable to answer broadly: claims to redefining our understanding of a foundational concept like "belief" are interesting, and contentious, a territory and status grab in the intellectual field, whereas a claim to discover a thing that is "sort of like belief, or like, sorta kinda one part of what we usually mean by 'belief' but not what we mean by it in another sense" doesn't cut it for newsworthiness. Here's extended belief, aided by note-taking systems and sticky notes: "Well, you know, if you wrote something you knew was false down in a notebook, and then like, forgot the original truth, you'd 'believe' the falsehood, in one sense that we mean when we use the word 'believe.'" I'm strawmanning its factoring—it describes a real chunk of cognition, of cognitive enmeshment in a technological age, and the way we use culture to outsource thinking—but at the end of the day, one (self-)framing—e-belief is belief proper—attracts a lot of glitz, and one framing doesn't. Here's Chalmers:

> Andy and I could have introduced a new term, "e-believe," to cover all these extended cases, and made claims about how unified e-belief is with the ordinary cases of believing and how e-belief plays the most important role.

Yeah, that would have been great.

> We could have done that, but what fun would that have been? The word "belief" is used a lot, it's got certain attractions in explanation, so attaching the word "belief" to a concept plays certain pragmatically useful roles.

He continues:

Likewise the word "conceptual engineering" Conceptual engineering is cool, people have conferences on it... pragmatically it makes sense to try to attach this thing you're interested in to this word.

He's 80% right and 100% wrong. Yes, there is a pragmatic incentive to attach your carving to existing carvings, to try to "take over" land, since contested land is more valuable. It's real simple: urban real estate is expensive, and this is the equivalent of squatters rights on downtown apartments. Chalmers and Clark's *factoring* of extended cognition is good, but they throw in a claim on contested linguistic territory for the glitz and glam. These are the natural incentives of success in a field.

That it's incentivized doesn't mean it's linguistic behavior philosophers ought to encourage, and David ought know better. If two people have different factorings of a word, they will start disagreeing about how to apply it, and they will apply it in ways that offend or confuse the other people. This is how bad things happen. Chalmers wrote a 60-page, 2011 paper on verbal disputes about exactly this. I'm inclined to wonder whether he really *did* take the concept from LessWrong, where he has freely admitted to have been hanging out on circa 2010, a year or two after the publication of linguistics sequences which discussed, at length, the workings of verbal disputes (there referred to as "tabooing your words"). The more charitable alternative is that this is just a concept "in the water" for analytic philosophy; it's "bar talk," or "folk wisdom," and Chalmers was the guy who got around to formalizing it. His paper's gotten 400 citations in 9 years, and I'm inclined to think that if it were low-hanging fruit, it would've been plucked, but perhaps those citations are largely due to his stardom. The point is, the lesson of verbal disputes is, you have to first be talking about the same thing with respect to the current dimensions of [conversation or analysis or whatever] in order to have a reliably productive [conversation or analysis or whatever]. Throwing another selfish -semous in the polysemous "belief" is like littering in the commons.

The problems with narrowness (or, the benefits of division)

I've written previously on various blogs about what I call "linguistic conquests"—epistemic strategies in which a polysemous concept—the product of a massive decentralized system of speakers operating in different environments across space and time, who using metaphor and inference have stretched its meaning into new applications—is considered to have been wrestled into understanding, when what *in fact* has occurred is a redefinition or refactoring of the original which moves it down a weight class, makes it easier to pin to the mat.

I distinguished between two types of linguistic conquest. First, the "narrow and conquer" method, where a specific sub-sense of a concept is taken to be its "true" or "essential" meaning, the core which defines its "concept-ness." To give an example from discourse, Taleb defines the concept *rationality* as "What survives, period." The second style I termed "divide and conquer," where multiple sub-senses are distinguished and named in an attempt to preserve all relevant sub-senses while also gaining the ability to talk about one *specific* sub-sense. To give an example from discourse, Yudkowsky separates *rationality* into epistemic rationality—the pursuit of increasingly predictive models which are "true" in a loose correspondence sense—and instrumental rationality—the pursuit of models which lead to in-the-world flourishing, e.g. via adaptive self-deception or magical thinking. (This second sense is much like Taleb's: rationality as what *works*.)

Conquests by narrowing throw out all the richly bundled senses of a concept while keeping only the immediately useful—it's *wasteful* in its parsimony. It leaves not even a ghost of these other senses' past, advertising itself as the original bundled whole while erasing the richness which once existed there. It leads to verbal disputes, term confusion, talking past each other. It impoverishes our language.

Division preserves the original, bundled concept in full, documenting and preserving the different senses rather than purging all but the one. It advertises this history; *intended* meaning, *received* meaning—the qualifier indicates that these are hypernyms of "meaning," which encompasses them both. Not just this, but the qualifier indicates the *character* of the subsense in a way that a narrowed umbrella original never will. Our understanding of the original has been improved even as our instrumental ability to wield its subsenses grows. Instead of stranding itself from discourse at large, the divided term has *clarified* discourse at large.

Chalmers, for his part, sees no difference between "heteronymous" and "homonymous" conceptual engineering—his own terms for two-word-type maneuvers (he gives as an example Ned Block factoring "access consciousness" from consciousness) and one-word-type maneuvers. One must imagine this apathy can only come from not having thought the difference through. He gives some nod—"homonymous conceptual engineering, especially for theoretical purposes, can be very confusing, with all these multiple meanings floating around." Forgive him—he's speaking out loud—but not fully.

Ironically, divide-and-conquer methods are, quite literally, the solution to verbal disputes, while narrow-and-conquer methods, meanwhile, are, while not the sole cause of verbal disputes, one of its primary causes. Two discourses believe they have radically different stances on the nature of a phenomenon, only to realize they have radically different stances on the factoring of a word.

Another way of framing this: you must always preserve the full extensional coverage. It's no good to carve terms and then discard the unused chunks—like land falling into the sea, lessening habitable ground, collapsing under people's feet. I'm getting histrionic but bear with me: If you plan on only maintaining a patch of your estate, you must cede the rest of the land to the commons. Plain and simple, an old world philosophy.

(Division also answers Strawson's challenge: if you divide a topic into agreeably constituent sense-parts, and give independent answers for each sense, you have given an accounting of the full topic. Dave, by contrast, can only respond: "Sure, I'm changing the topic—here's an interesting topic.")

A quick Q & A

I'm going to close by answering an audience question for Dave, because unfortunately he does not do so good a job, primarily on account of not understanding conceptual engineering.

> Paul Boghossian: Thanks Dave. Very useful distinctions. [Note: It's unclear why Chalmers' distinctions are useful, since he has not indicated any uses for them.] To introduce a new example, to me one of the most prominent examples of de novo engineering is the concept genocide... Lemkin noticed that there was a phenomenon that had not been picked out. It had certain features, he thought those features were important for legal purposes, moral purposes, and so on. And so he introduced the concept in order to name that. [He's on the money here, and then he loses it.] That

> general phenomenon, where you notice a phenomenon, of course there are many phenomena, there are murders committed on a Tuesday, you could introduce a word for that, but there, I mean, although you might have introduced a new concept, it's not clear what use is the word. So it looks as though... I mean, science, right? I mean...

Paul is a bit confused here also. Noticing phenomena in the world is not something particular to science; the detection of regularity *is cognition itself*. If we believe Schmidhuber or Friston, this is the organizing principle of life, via error minimization and compression. "Theorizing" is a better word for it.

And yet, to the crux of the issue he touches on: why don't we introduce a word for murders committed on a Tuesday? You say, well what would be the point? Exactly. This isn't a very hard issue to think through, it's intuitively quite tractable. Paul *also* happens to mention *why* the concept "genocide" was termed. He just had to put the two together. "Genocide" had legal and moral purposes, it let you argue that the leader of a country, or his bureaucrats, were culpable of something especially atrocious. It's a tool of justice. That's why it exists: to distinguish an especially heinous case of statecraft from more banal ones. When we pick out a regularity and make it a "thing," we are doing so because the thingness of that regularity is of use, because it distinguishes something we'd like to know, the same way "sandy soil" distinguishes something gardeners would like to know.

Philosophy of Language 10 | Meta-Philosophy 7 | Rationality 8 | World Modeling 8 | Frontpage

Mentioned in

117 Situating LessWrong in contemporary philosophy: An interview with Jon Livengood
60 Looking Deeper at Deconfusion
49 [Simulators seminar sequence] #1 Background & shared assumptions
30 Book Review: Philosophical Investigations by Wittgenstein
30 AI Alignment, Philosophical Pluralism, and the Relevance of Non-Western Philosophy
Load More (5/14)

50 comments, sorted by top scoring

[-] **Kaj_Sotala** 3y < 20 >

> As part of that larger project, I want to introduce a frame that, to my knowledge, hasn't yet been discussed to any meaningful extent on this board: conceptual engineering, and its role as a solution to

> the problems of "counterexample philosophy" and "conceptual analysis"—the mistaken if implicit belief that concepts have "necessary and sufficient" conditions—in other words, Platonic essences.

After reading the essay, I'm still confused by what conceptual engineering actually is. Is it a claim about how humans use language in general, a philosophical technique like conceptual analysis is, or both?

(You seemed to attack Chalmers for trying to offer a definition for conceptual engineering, but a brief definition for the concept was exactly what I found myself hoping for. I think you are saying that you don't want to offer one because terms don't have necessary and sufficient definitions so offering a definition goes against the whole spirit of the approach... but you also note that we learn words by seeing a specific example and expanding from that, so I wouldn't think that it would be contrary to the spirit of the approach to offer a brief definition and then expand from that once the readers have something to hang their minds on°.)

[–] **Suspended Reason** 3y < 9 >

Yes, so the premise of Chalmers's lecture, and many other texts being published right now in conceptual engineering (a quickly growing field) is to first treat and define "conceptual engineering" *using* conceptual engineering—a strange ouroboros. Other philosophers are doing more applied work; see Kevin Scharp's version of conceptual engineering in his work on truth, or Sally Haslanger's version of it, "ameliorative analysis." But broadly, Chalmers's tentative definition is fine as a generic-enough umbrella: constructing, analyzing, renovating, etc. Right now, really anything in the ballpark of what "conceptual engineering" intuitively connotes is a fine description.

One place to start, as Cappelen does in his monographs on the subject, is with Nietzsche's *Will to Power*, so I'll quote that here:

> Philosophers … have trusted in concepts as completely as they have mistrusted the senses: they have not stopped to consider that concepts and words are our inheritance from ages in which thinking was very modest and unclear. … What dawns on philosophers last of all: they must no longer accept concepts as a gift, nor merely purify and polish them, but first make and create them, present them and make them convincing. Hitherto one has generally trusted one's concepts as if they were a wonderful dowry from some sort of wonderland: but they are, after all, the inheritance from our most remote, most foolish as well as most intelligent ancestors. …What is needed above all is an absolute skepticism toward all inherited concepts.

Might add to the main post as well for clarity.

EDIT: Also, to be clear, my problem is not that Chalmers attempts to offer a definition. It's that, when presented with an intellectual *problem*, his first recourse in designing a solution is to *consult a dictionary*. And to make it worse, the concept he is looking up in the dictionary is a *metaphor* that a scholar twenty years ago thought was a nice linguistic turn of phrase.

[–] **romeostevensit** 3y < 3 >

How to Philosophize with a Hammer and Chisel

[–] **peak.singularity** 2y < 1 >

As long as it's not with a bulldozer...
https://medium.com/s/story/peterson-historian-aide-m%C3%A9moire-9aa3b6b3de04

[–] **Crotchety_Crank** 2y 12

I wonder if anyone ever reads the ancients anymore. Because around here I frequently hear the word "Platonic", applied to a naive straw man of formalism that I'm not sure anyone ever believed, let alone Plato's character of Socrates. That caricature has its genesis in "A human's guide to words", which characterizes it by relying on (of all things) Aristotelean syllogism. It should go without saying that this is confused.

It irks me when supposed philosophers so badly misunderstand the history of their own field, and diagnose and psychologize philosophy's past with the ailment of "aprioristically" and simplistically reasoning based on "necessary and sufficient conditions" alone. Here is Aristotle, in the preface to the Nicomachean Ethics, the most commonly read of any of his works:

Now fine and just actions, which political science investigates, admit of much variety and fluctuation of opinion, so that they may be thought to exist only by convention, and not by nature...We must be content, then, in speaking of such subjects and with such premisses to indicate the truth roughly and in outline, and in speaking about things which are only for the most part true and with premisses of the same kind to reach conclusions that are no better. In the same spirit, therefore, should each type of statement be received; for it is the mark of an educated man to look for precision in each class of things just so far as the nature of the subject admits; it is evidently equally foolish to accept probable reasoning from a mathematician and to demand from a rhetorician scientific proofs.

The guy whose logical system is supposedly the source of all this aprioristic confusion was very content to characterize in general terms. He explicitly understood that everyday concepts are "fuzzy," or "inconsistent," or "polysemous", and explicitly cautioned against over-rationalizing them. So please understand before you criticize, and appreciate the ancients before slurring them. (No, you get no credit for hiding behind Bishop's words; you should be able to stand behind your citations, rather than credulously repeating the ignorance of another.)

Why is this a problem? There's worse exegesis of the ancients out there. And there are enough people unjustifiably worshiping Plato, that a few people uncharitably misunderstanding him on a minor forum doesn't upset the balance. This is a problem because if you don't understand Aristotle, you can't understand where he failed, and why. "All previous philosophy neglected the vagueness of ordinary concepts; we will correct this" is the exact kind of arrogance that will lead directly to a whole set of problems that have already been encountered (by Kripke, by Kant, by Epictetus, by Aristotle). And on the way, due to your dismissiveness of prior work, you'll incorporate a whole bunch of new jargon like "polysemous" that will make it harder for you to relate to the existing literature, and obscure the connections to all the good answers from the past to the questions you're investigating.

Ever wonder why there are twelve different definitions for the same idea in philosophy? It's because every so often, someone like this comes along pledging to solve philosophy anew, throwing out the old terms and confusions and starting fresh. The end result is always just to add another set of terms and confusions to the pile. To make it even more explicit: You will worsen the problem you are trying to solve.

I suppose I shouldn't be especially upset. Welcome to the grand tradition of philosophers misunderstanding the history of philosophy; there is nothing new under the sun. But somehow, I am especially upset by this, and a part of me I'm not proud of will experience especially intense schadenfreude as it watches this nascent fad fall into the same old philosophical pitfalls that the author slept through when they were covered in Intro to Ancient Phil.

[–] **Suspended Reason** 2y 15

Hey Crotchety_Crank,

Your name does suit you. I have in fact read (AFAIK good translations of) Plato and the Sophists! Very little Aristotle, and you're correct I fell asleep once or twice during an ancient phil course. Not, however, during the Plato lectures, and my prof—a hot young philosopher recently tenured at NYU—presented a picture of Platonic forms that agrees with my own account. I don't at all mean to imply that reading is the only correct interpretation, but it's a common and perhaps dominant one—several credible sources I've encountered call it the "standard" reading. A few eclectic notes in response to more specific points of contention:

- It may well be that Socrates did not believe in sufficient and necessary conditions—he is part fictional creation, so we can't of course know for sure, but he obviously carries out his dialogues in a way that can be interpreted as challenging a view of e.g. the Good or the Just as having a clear definition. This, however, is a very different question from what Plato, or later philosophers who followed Plato's footsteps, believed, as you well know.
- Depending on how one interprets Plato's language, specifically his description of the realm that forms exist in, and what it means for a form to exist, one can, perhaps, charitably understand Plato as not implying some "essence" of things. (OTOH, it also doesn't seem an accurate reading to say Plato saw these concepts as existing in the mind—so it's not clear *where* the hell he thinks they dwell. This question takes up hundreds if not thousands of pages of anguished scholarly writing.) But, important to note—as soon as one believes in an essence, "sufficient and necessary conditions" follows naturally as its prerequisite.
- It doesn't actually matter so much what Plato intended; what counts, pragmatically speaking, is how he was interpreted, and Neoplatonism + Christian metaphysics clearly believe in essences; their philosophical doctrines ruled the West for over a millennium.
- It is clearly false to say that "sufficient and necessary" is a strawman that no one ever believed. Logical positivism, conceptual analysis, and the history of analytic all explicitly contradict this claim.
- Whether or not individuals explicitly pay lip service to "sufficient and necessary," or a concept of essences, is also besides the point; as I have argued, the mode of analysis which has dominated analytic philosophy the past century rests *implicitly* on this belief.

I see you're brand new here, so a head's up: discursive norms here veer constructive. If you believe I'm wrong, please make an argument for your alternate interpretation instead of casting ad hominems. Your last line is a sick diss—no hate! much respect!—but sick disses don't hold much water. Other than a quotation by Aristotle, who is not mentioned in this post anywhere, there is no textual support in your comment for your interpretations of Plato, Socrates (though I agree), or any of the other listed philosophers.

Here is the Stanford Encyclopedia entry on Wittgenstein's life:

> Family resemblance also serves to exhibit the lack of boundaries and the distance from exactness that characterize different uses of the same concept. Such boundaries and exactness are the definitive traits of form—be it Platonic form, Aristotelian form, or the general form of a proposition adumbrated in the *Tractatus*. It is from such forms that applications of concepts can be deduced, but this is precisely what Wittgenstein now eschews in favor of appeal to similarity of a kind with family resemblance.

Note that Wittgenstein was an avid reader of Plato; he cited the philosopher more than any other, but viewed his own approach as a radical break. (He did not read Aristotle! Interesting!) It seems possible to me that Wittgenstein himself, the authors of the SEP article, and the editors who peer-reviewed it, have fundamentally misunderstood not just Platonic forms but Aristotelian forms, and therefore, the entire legacy of Wittgenstein's work. But that is a serious case to build, and it's unclear why I should take your word for it over theirs without any presentation of evidence.

Your claims here go against major philosophical reference sources, many dominant interpretations of Platonic forms, and the interpretations of many well-informed, well-read philosophers of language past. They contradict various histories of the discipline and various historical dilemmas—e.g. Bertrand Russell, famous for writing one of the most definitive histories of philosophy, is sometimes seen as "solving" the Sorites paradox (an ancient Greek philosophical problem) by arguing that natural language is vague. I'm sure other historic philosophers have made similar interventions, but if this appeal to vagueness was as obvious and widely understood as you claim, it's unclear to me why the Sorites paradox would have staying power, or why Russell's solution would be taken seriously (or why he'd bother resolving it in the first place).

I'm sincerely interested in engaging—I *do* think the story is more complicated than this piece lays out. But arguments must take a side, and summaries must exclude nuance. If you're interested in a good-faith discourse I'm game.

[–] **Crotchety_Crank** 2y < 12 >

Thanks, I picked the name myself. This is a new account because I haven't commented before, but I'm long familiar with this community and its thought - and its norms. Given those norms, I probably should have cooled off a bit before posting that comment. Let me try again. I apologize in advance for the length of the below, but charity takes more work and therefore more words.

Fairness to the Ancients

I think we're talking past one another. Plato was definitely a Platonist, and he definitely employed counterfactual reasoning. Congratulations to your Ancient Phil professor on achieving tenure; I studied under others (I won't say who or where for privacy reasons), and they likewise taught that Plato believed in essences. I was not trying to imply that I think otherwise. I simply don't think that the thing Eliezer attacked in "A human's guide to words" was, in fact, Platonism; I think it was a straw man. And I took you to also be putting up that straw man, and associating it with all philosophers before Wittgenstein.

You did not cite Aristotle; I brought him in because you cited "a human's guide to words" as a paradigmatic example of a good argument against "Platonic essences." And yet, that sequence is not really arguing against Platonic essences, it's arguing against misapplying Aristotelean syllogism. Eliezer attacks the idea that the logical validity of "Socrates is a man, all men are mortal, therefore Socrates is mortal" entitles you to conclude things with certainty in the real world. Eliezer attributes that view to "the Greek philosophers", calling them "fond of certainty."° He ridicules this view often throughout the sequence. I think the passage I quoted in my original comment shows this to be a straw man of (among others) Aristotle. Aristotle acknowledges that when your premises are uncertain, your conclusions will be too; and that seeking certainty about uncertain or ill-defined concepts is a fool's errand. For that matter, I would say every Greek philosopher I am aware of would have acknowledged this, and many wrote about the problem!

The other citation that seemed unduly dismissive of the ancients was your citation to Bishop as saying that philosophers "aprioristically" reasoned from their armchairs prior to the 1900s. For the life of me, I can't find that in Bishop 1992 (ctrl+F "aprior" and "armchair", 0 results); if you can cite more specifically, I would appreciate it. I would almost certainly have qualms with any assertion of his saying "[X idea] wasn't considered before [Y date]", if he did in fact say anything along those lines.

I definitely agree that Plato was a Platonist; I'm not going against philosophical consensus on that front. What I took you to be doing was taking the label "platonism", attaching it to Eliezer's straw man, and then saying that philosophers prior to 1900 all believed it and therefore have nothing to contribute.

I took you to agree with Eliezer because you cited him, and I *really strongly dislike* his mischaracterization of Aristotle, and *even further* dislike the fact that he takes that view and attributes it to "the Greeks", whom he slurs together. I took you to be reproducing that straw man, attaching the name "platonism" to it, and generalizing that view to an even wider range of philosophers who endorsed nothing like it. I still think the article as written can create that impression, but it sounds like that wasn't your intent, and I'm sorry for jumping the gun into what amounted to an attack on your intelligence.

I'll stand by my assertion that "a human's guide to words" straw mans the ancients. Again, virtually none of the Greeks agreed with the view he attributed to them, and for that matter, attributing just about anything to "the Greeks" is bound to be wrong, given the vast differences between the diverse thinkers in the ancient Hellenistic world. I took my irritation at Eliezer's ignorance about the ancients, unfairly assumed you agreed in full with his assessments and characterizations because of your citation of that sequence, and extended that irritation towards you, thinking to myself, "*as a philosopher, this person should know better!*"

Points for Further Discussion

Finally, I want to thank you for taking the time to write a response to an ill-tempered crank; I hope I've acquit myself honorably enough in this follow-up to receive another. If you'd like to continue the conversation to more productive discussion of conceptual engineering itself, rather than disputing the ancients, I'd be interested to hear your thoughts on the following propositions (which are, of course, derived directly from ancient thinking):

1. Counterfactual reasoning (/"Conceptual Analysis") *is* the primary tool which has been used to demonstrate the vagueness of concepts, so disposing of it is dangerous to any project which is premised on the vagueness of concepts. It is one *extremely* useful tool (among others) for engineering and streamlining useful conceptual frameworks which align well with language.
2. A good account of concepts should include how concepts change. For better or for worse, concepts change when people argue about them - often counterfactually. This means that a project which sets out to understand concepts, but neglects to include counterfactual reasoning as an element of the project, may run into some very hard times very fast. "Conceptual engineering," as laid out in the article above, is not (yet?) equipped with the necessary tools for this.

[–] **Suspended Reason** 2y 9

Thanks for the thorough reply! This makes me want to read Aristotle. Is the Nichomachean preface the best place to start? I'll confess my own response here is longer than ideal—apologies!

Protagoras seems like an example of a Greek philosopher arguing against essences or forms as defined in some "supersensory" realm, and for a more modern understanding of concepts as largely carved up by human need and perception. (Folks will often argue, here, that species are more or less a natural category, but species are—first—way more messy constructed than most people think even in modern taxonomy, second, pre-modern, plants were typically classed not first and foremost by their effects on humans—medicine, food, drug, poison.) Still, it's hard to tell from surviving fragments, and his crew did get run out of town...

I say:

> For a while, arguably until Wittgenstein, philosophy had what is now called a "classical account" of concepts as consisting of "sufficient and necessary" conditions. In the tradition of Socratic dialogues, philosophers "aprioristically" reasoned from their proverbial armchairs

Do you think it would be more fair to write "philosophy [was dominated by] what is now called a classical account"? I'd be interested to learn why the sufficient & necessary paradigm came to be called a classical account, which seems to imply broader berth than Plato alone, but perhaps it was a lack of charity toward the ancients? (My impression is that the majority of modern analytic is still, more or less, chugging ahead with conceptual analysis, which, even if they would disavow sufficient and necessary conditions, seems more or less premised on such a view—take a Wittgensteinian, family resemblance view and the end goal of a robust and concise definition is impossible. Perhaps some analytic still finds value in the process, despite being more self-aware about the impossibility of some finally satisfying factoring of a messy human concept like "causality" or "art"?) One other regret is that this piece gives off the impression of a before/after specific to philosophy, whereas the search for a satisfying, singular definition of a term has plagued many fields, and continues to do so.

Like I said, I haven't read Aristotle, but Eliezer's claim seems at most half-wrong from a cursory read of Wikipedia and SEP on "term logic." Perhaps I'm missing key complications from the original text, but was Aristotle not an originator of a school of syllogistic logic that treated concepts somewhat similarly to the logical positivists—as being logically manipulable, as if they were a formal taxonomy, with necessary and sufficient conditions, on whom deduction could be predicated? I've always read those passages in HGtW as arguing against naive definition/category-based deduction, and for Bayesian inference or abduction. I also must admit to reading quite a bit of argument-by-definition among Byzantine Christian philosophers.

Frustratingly, I cannot find "aprioristically" or "armchair" in Bishop either, and am gonna have to pull out my research notes from the archive. It is possible the PDF is poorly indexed, but more likely that line cites the wrong text, and the armchair frame is brought up in the Ramsey paper or similar. I'll have to dive into my notes from last spring. Bishop does open:

> Counterexample philosophy is a distinctive pattern of argumentation philosophers since Plato have employed when attempting to hone their conceptual tools... A classical account of a concept offers singly necessary and jointly sufficient conditions for the application of a term expression that concept. Probably the best known of these is the traditional account of knowledge, "X is knowledge iff X is a justified true belief." The list of philosophers who have advanced classical accounts... would not only include many of the greatest figures in the history of philosophy, but also highly regarded temporary philosophers.

This is not, however, the same as saying that it was the only mode across history, or before Wittgenstein—ceded.

Glad to step away from the ancients and into conceptual engineering, but I'd love to get your take on these two areas—Aristotle's term logic, and if there are specific pre-moderns you think identify and discuss this problem. From your original post, you mention Kripke, Kant, Epictetus. Are there specific texts or passages I can look for? Would love to fill out my picture of this discourse pre-Wittgenstein.

On the conceptual analysis/engineering points:

1. I have wondered about this too, if not necessarily in my post here then in posts elsewhere. My line of thought being, "While the ostensible end-goal of this practice, at least in the mind of many 20th C practitioners—that is, discovering a concise definition which is nonetheless robustly describes all possible instances of the concept which a native speaker would ascribe—is impossible (especially when our discourse allows bizarre thought experiments a la Putnam's Twin Earth…), nonetheless, performing the *moves* of conceptual analysis is productive in understanding the concept space. I don't think this is *wrong*, and like I semi-mentioned above, I'm on your side that Socrates may well have been in on the joke. ("Psych! There was no right answer! What have you learned?") On the other hand, having spent some time reading philosophers hand-wringing over whether a Twin Earth-type hypothetical falsifies their definition, and they ought to start from scratch, it felt to me like what ought to have been non-problems were instead taking up enormous intellectual capital.

If you take a pragmatist view of concepts as functional human carvings of an environment (to the Ancients, “man is the measure of all things”), there would be no reason for us to expect our concepts’s boundaries and distinctions to be robust against bizarre parallel universe scenarios or against one-in-a-trillion probabilities. If words and concepts are just a way of getting things done, in everyday life, we’d expect them to be optimized to common environmental situations and user purposes—the minimum amount of specification or (to Continentals) “difference” or (to information theory) “information.”

I’m willing to cede that Socrates may have effectively demonstrated vagueness to his peers and later readers (though I don’t have the historical knowledge to know; does anyone?) I also think it’s probably true that a non-trivial amount of insight has been generated over many generations of conceptual analysis. But I also feel a lot of insight and progress has been foreclosed on, or precluded, because philosophers felt the need to keep quibbling over the boundaries of vagueness instead of stopping and saying, “Wait a second. This point-counterpoint style of definitions and thought experiments is interminable. We’ll never settle on a satisfying factoring that solves every possible edgecase. So what do we do instead? How do we make progress on the questions we want to make progress on, if not by arguing over definitions?” I think, unfortunately, a functionalist, pragmatist approach to concepts hasn’t been fleshed out yet. It’s a hard problem, but it’s important if you want to get a handle on linguistic issues. You can probably tell from OP that I’m not happy with a lot of the conceptual engineering discourse either. Many of it is fad-chasing bandwagoners. (Surprise surprise, I agree!) Many individuals seem to fundamentally misunderstand the problem—Chalmers, for instance, seems unable to perform the necessary mental switch to an engineer’s mindset of problem-solving; he’s still dwelling in definitions and “object-oriented,” rather than “functionalist” approaches—as if the dictionary entry on “engineering” that describes it as “analyzing and building” is authoritative on any of the relevant questions. Wittgenstein called this an obsession with generalizing, and a denial of the “particulars” of things. (Garfinkel would go on to talk at length about the “indexicality” or particulars.) Finding a way to deal with indexicality, and talk about objects which are proximate in some statistical clusterspace (instead of by sufficient and necessary models), or to effectively discuss “things of the same sort” without assuming that the definitional boundaries of a common word perfectly map to “is/is not the same sort of thing,” are all important starts.

2. I can’t agree more that “a good account of concepts should include how concepts change.” But I think I disagree that counterfactual arguments are a significant source of drift. My model (inspired, to some extent, by Lakoff and Hofstadter) is that analogic extension is one of the primary drivers of change: X encounters some new object or phenomenon Y, which is similar enough to an existing concept Z such that, when X uses Z to refer to Y, other individuals know what X means. I think one point in support of this mechanism is that it clearly leads to family-resemblance style concepts—“well, this activity Y isn’t *quite* like other kinds of games, it doesn’t have top-down rules, but if we call it a game and then explain there are no top-down rules, people will know what we mean.” (And hence, Calvinball was invented.) This is probably a poor example and I ought to collect better ones, but I hope it conveys the general idea. I see people saying “oh, that Y-things” or “you know that thing? It’s kinda like Y, but not really?” Combine this analogic extension with technological innovation + cultural drift, you get the analogic re-application of terms—*desktop*, *document*, *mouse*, all become polysemous.

I’m sure there are at least a couple other major sources of concept drift and sense accumulation, but I struggle to think of how often counterfactual arguments lead to real linguistic change. Can you provide an example? I know our culture is heavily engaged in discourses over concepts like “woman” and “race” right now, but I don’t think these debates take the character of conceptual analysis and counterfactuality so much as they do arguments of harm and identity.

[–] **Crotchety_Crank** 2y < 10 >

Thanks for the reply. I'll try to reply comprehensively, sorry if I miss anything. To start with - Aristotle.

What Aristotle Taught

> Was Aristotle not an originator of a school of syllogistic logic that treated concepts somewhat similarly to the logical positivists?

I'm going to break this into two parts - the part about logic, and the part about concepts. Logic first. Aristotle indeed wrote six works on logic and reasoning, which are most often collectively called the Organon. Most of it is developing a valid system of syllogistic logic. The really nice part about syllogistic logic is that correct syllogisms are indisputably valid (but not indisputably *sound*). Aristotle is totally clear about this. He showed - correctly - that logic, correctly applied, makes your conclusions *as true as* your premises (i.e. logic is *valid*); but that alone still doesn't entitle you to certainty about your conclusions, as you can't trust your premises any more than you could from the start (i.e., validity is not soundness).

In *The Parable of Hemlock°*, ctrl+F "the Greeks." Eliezer's issue isn't with syllogism. It's with something different: the assertion that "all men are mortal" *by definition*. Aristotle says nothing of the sort, least of all in the Organon; he just uses the statement as a hypothetical premise to demonstrate the form of valid syllogism, the same way you might use a sample like "all frogs are green, Harold is a frog, Harold is green" as a lesson of validity in a logic class, regardless of whether purple dart frogs exist. The text that most clearly shows this is the *Topics*, where Aristotle characterizes good arguments as constructed by using syllogism (as characterized in the earlier works of the Organon) or enthymematic syllogism, especially when the syllogism begins from established beliefs (endoxa) as premises. Explicitly, these endoxa like "all men are mortal" are not certain or guaranteed to be true; but they are better than wild speculation, especially if you are trying to persuade someone. So Eliezer's attack on the Greeks is off base, mistaking the assertion of validity for the assertion of soundness.

There's nothing wrong with syllogistic logic, as long as you don't make too much of it. Eliezer's top-line conclusion is that "logic never dictates any empirical question [with certainty]"; I think you would be *extremely* hard-pressed to find a sentence in Aristotle which disagrees, and Eliezer's clear imputation that they *did* disagree is ignorant and uncharitable. Logic is a useful tool for reasoning from premises you are reasonably confident in, to conclusions you can be similarly confident in.

It's no straw man to say that Aristotle liked logic. The straw-manning comes when Eliezer asserts that "the Greeks" thought you could derive certain empirical truths from logic *alone*. (Parmenides, Spinoza, and Kant attempted this, but not Plato, Aristotle, or most philosophers.) Rather, Aristotle's logic is all about taking established pretty-good beliefs (which are not called certain, but are generally acknowledged and are the best we have to work with) and having a sure way to arrive at *exactly equally good* beliefs. Putting this in writing was an *incredibly* valuable contribution to philosophy.

Now for the part about concepts. Did Aristotle treat concepts similarly to the logical positivists? Honestly, I think not; my impression is that the average positivist was a nominalist about the question of universals, while the best summary of Aristotle's view on the topic probably heavily uses the word hylomorphism. It's kinda his own deal, like how Plato was Platonist. I don't love Aristotle's metaphysics, and I think there are powerful skeptical/nominalist critiques of hylomorphism, which is after all a formalist view of one kind or another. But I don't think Eliezer really advanced them, or understood Aristotle's (or any Greek's) phenomenology of concepts at all. For a little taste of how nuanced Aristotle's thoughts on words and concepts actually were, here's another bit from the last book of the Organon:

> It is impossible in a discussion to bring in the actual things discussed: we use their names as symbols instead of them; and therefore we suppose that what follows in the names, follows in the things as well, just as people who calculate suppose in regard to their counters. But the two cases

> (names and things) are not alike. For names are finite and so is the sum-total of formulae, while things are infinite in number. *Inevitably, then, the same formulae, and a single name, have a number of meanings*. [emphasis added]

Relevant Reading (By philosopher)

> specific pre-moderns you think identify and discuss this problem.

If we're discussing the problem of "gee whiz, in what sense do concepts exist and truthfully inhere in an ever-changing world?" Virtually all of them! Here's a short rogue's gallery, take your pick if you're intrigued by one in particular.

Plato: Plato's answer is formalism. But even (or especially) if you think that's absurd, his treatment of the *question* is incredibly valuable. Plato is *deeply aware* and *deeply disturbed* by the fact that the world around him is changeable, that appearances and naively-constructed concepts deceive, and that nothing certain can be found in them. And the core of many of his dialogues are devoted to proving exactly that. Take the *Theaetetus*, where he talks about certain knowledge. Can we get it by sense perception? Not quite, appearances can deceive. What about judgment? Fallibility would indicate no. Is it justified true belief? Perhaps, but "justification" demands prior knowledge of the thing itself, so this is invalid by circularity! Plato strongly hints at his solution of formalism, but to pave the way to it, he demolishes more standard accounts first by trying to prove the slipperiness of ordinary concepts and the inaccessibility of certainty. Skeptical accounts can find a great deal to like. (Ever wonder why J.L. Mackie's skeptical "argument from queerness" begins as a steadfast defense of Platonism as the only way to objective morality? For generations, skeptics have made hay by starting with Plato's objections to others, then attacking Plato's rehabilitative view as the final step of a deflationary account.) *Parmenides* is also recommended reading, as most of it is *criticism* of the theory of forms. But it's not for the faint of heart, you'll need some really good secondary lit - or far better, a supportive professor to read through it with. Trying to read and understand it by yourself is an aneurysm risk.

Aristotle: Often denser than Plato. But he's far more methodical and much easier to interpret, since he's not writing dialogues with Straussian readings or citing myths which he didn't believe or any of that artistic jazz. The *Nicomachean Ethics* may be a good place to see him apply his method of discourse about the natural world, but the writings of his that are most relevant to this conversation are definitely *Physics* and *Metaphysics*. (fun fact: the field was named for the book; "meta" is just Greek for "after", so "Metaphysics" just means "after physics", "more physics," or maybe "physics 2".)

Stoics: Chrysippus is your boy here. He is taken to be one of the first *nominalists* (a general term for one of the most popular non-realist views, i.e., that universal properties are words alone and not things in their own right). https://iep.utm.edu/chrysipp/#H5 has a summary you might like, and it may be the best we can do, since virtually all of Chrysippus' actual writings are not extant (his views were passed to us by way of others' summaries of them), and most other Stoics (like Epictetus or Aurelius) spent more time talking about ethics, with physics receiving more of a passing mention.

Epicureans: Really just Epicurus, as his teachings were passed down by Lucretius in *De Rerum Natura*. Virtually nothing else from this school is extant, but their influence is very significant. Steadfast materialists, atomists, atheists, and hedonists. This community would like their teachings a *lot*. I'll take this opportunity to point out a trend which is commonplace throughout ancient philosophy; Epicureanism is atheist, but the text sings paeans to gods, using them as stand-ins for abstract concepts. This is weird, but not at all rare in ancient philosophy. Anytime you see someone invoke a god or a myth, before dismissing it as superstition, see if it's useful to treat it as metaphor or conjecture instead. Remember that, for all the talk of gods and myths he engaged in, one of Socrates' two crimes that he was killed for was *impiety*.

Skeptics: You will agree with these people less than their names imply you will. They thought some weird stuff; Academic or Pyrrhonian, either way it sometimes comes off as worshiping ignorance. In any case, formalists they were *not,* and their eponymous attitude comes across in their writings, which are very clear that if there *are* in fact universals, we are either *unable* to come to know them, or even *morally forbidden to try*.

Peripatetics, Cynics, Cyrenaics and more: there are *so many ancient Greeks*. Many of them may not have written anything of value on this question, I can't say. This is the part where I confess ignorance of and wonder at the true diversity of Ancient Greek thought.

Another big gap in my knowledge is Christian and medieval thought, but I had enough friends who studied it to understand that my received caricatures of it were misplaced. Aquinas apparently contributed things to metaphysics in the vein of Aristotle. Maybe Augustine has dope metaphysics, no idea. God features prominently, of course, so know thyself and whether that's a turn-off.

Early Moderns: Spinozism is *super weird* and monist and stuff. Maybe not that. Kantianism is *incomprehensible*, even in the original German, but if you can find a good professor to walk you through it (preferably in a classroom environment), there is a reason he was so influential. The obvious suggestion is the *Critique of Pure Reason*, and it is definitely the one that is relevant here. (It's where the separation of syntheticity from prioricity comes from! I don't think it's a *good* separation, but you will need to understand what it means if you want to understand many metaphysicians after him, most of all Kripke.) I personally like *The Critique of Judgment* too.

Continentals: Another gap in my knowledge. A friend read a lot of them and said "there's no there there", but I would guess that had as much to do with that friend as the writing itself. Another said Hegel is apparently very fun "in the right state of mind" (I think they meant psychedelics. This is not an endorsement of illegal drug use.) As with other categories on this list, I will acknowledge my ignorance of whatever brilliance might be here. For what it's worth, if you are interested in critiquing the "classical" method of counterfactual reasoning - or reasoning in general - you may find allies here, even if they are strange bedfellows.

Moderns: Jumping right up to the 1900s. Meinong gets a bad rep but I still like him (do square circles exist? Maybe as much as anything else does!) Russell and Wittgenstein, you cited already. Tarski is also a great one, who created a modal logic ("T-schemas" is a search term you can start with) which is intended to be generalizable over different uses of language. Almost certainly has connections to anything philosophy of language-related. I like Carnap a whole lot, and he did a lot of philosophy of science which you may find relevant. I dislike Kripke a lot, but there's no question that his thought is an intensely relevant to any philosophy which deals directly with the idea of *meaning* (he doesn't think it's a thing, or at least, wants a deflated version of it to be the norm). He took himself to be in the tradition of Wittgenstein.

Counterfactual Reasoning

I really like, and generally agree with, your summary of how edge cases and obtuse counterexamples have pushed people to somewhat absurd conclusions. I'll provide some pushback, but first let me indulge myself in agreeing, and providing an example. My undergraduate senior paper employed an unfortunately complex variant of the trolley problem (guess how many tracks were involved?) to contest an arcane ethical principle relevant to a *facially absurd* variant of utilitarianism. It was truly approaching self-parody, and I was well aware, I just wasn't sure what other topic I had an idea about which would fill enough pages. (funnily enough, I can write more than enough pages on random internet fora, though.)

For all that ethics should be able to provide us with answers, and there should *be* answers even for corner cases... it is extremely clear to me that academic ethics has gone over the deep end. Ethical views are now defined based on cases which are often so ridiculous that *whatever* decision one would make in those situations is probably a noncentral example of ethical or unethical behavior. It's clear enough to me how we got here, given a certain kind of steadfast realism about ethics, and it's unclear what exact countervailing view I think should prevail... but somewhere, somehow, we have gone wrong.

Is the source of the problem counterfactual reasoning itself? Perhaps a certain too-strong form of it. But I also think that a mature version of "conceptual engineering" would see a lot of it employed.

> I'm sure there are at least a couple other major sources of concept drift and sense accumulation, but I struggle to think of how often counterfactual arguments lead to real linguistic change. Can you provide an example?

The example, or family of examples, that I want to give you and propose as an incredibly useful analogy here, definitely one where there are lots of examples of "concept drift and sense accumulation", is *law*. It's not exactly common usage, but legal language has a bunch of desirable features as an analogy here to apply "conceptual engineering" to. The boundaries of initially-vague concepts like "probable cause" or "slander" are often decided based on past definitions and laid-out sets of necessary and sufficient conditions in case law. But they are also subject to shift when corner cases are encountered which clearly do or don't fall into the category - previous understandings of the necessary and sufficient conditions be damned. Ultimately, the courts converge on definitions that are *useful* at the very least, and they use a number of methods to do it, counterfactual reasoning and N&S conditions being some of the tools in the toolbox. Do you think law should dispose of those tools, and do you think it would lead to better decisions if they did? My answer is "no"; I think they're great pragmatic tools in conjunction with other tools; and that makes me think that N&S conditions and counterfactual reasoning aren't the real problem here. They can be useful ways to engineer concepts, rather than just a destructive way to attack them with corner cases.

Legal language is also nice because it gives us a clear sense of an evaluative objective, a way to "grade" our engineering project - in a word, we might say "justice." (Meanwhile, to engineer common language, we might grade based on "clarity" or "intersubjectivity".) When the existing body of rules and conditions still leave room for doubt, we can employ and develop our terminology to produce results that accord with a notion of justice.

I hope you like that proposed application of the theory. Interested to hear your thoughts on whether it's fitting, or if not, why not.

[−] **Suspended Reason** 2y ‹ 5 ›

Appreciate the thorough response; there are some good recs here. I haven't read any of Chrysippus, and my knowledge of the Epicureans is limited to their moral philosophy (alongside that of the Stoics). That said, I can't help but get the feeling you're negging me a little with the references to skeptics, continentals, and professorial assistance! Fortunately or unfortunately, I'm less a rationalist than my presence here might imply—Bourdieu's symbolic capital and ethology's signaling theory are interchangeable in my book. Also fortunately or unfortunately, I'm not a uni student these days, my institutional education concluded a few years back, so I suppose I'll have to make headway on any texts solo, without professorial help.

A quick meta-note: I think there's a problem whereby people who study historic philosophy have incentives to steelman their subjects' ideas and thinking, in order to justify their study. I imagine this claim will be received with some pushback, so I'll try to break it down to less controversial parts, and we can sum them together. First, I think there are strong incentives in academia for *everyone* to constantly justify their work. Whether it's prepping for department socials, getting tenure, applying for grants, or just coming to peace internally with a lifetime dedicated to scholarship, it's hard to help this subtle narrative of self-justification. Second, I think when we read ancient texts, we're in a tricky situation. As Wittgenstein once said of Plato,

> Perhaps Plato is no good, perhaps he's very good. How should I know? But if he is good, he's doing something which is foreign to us. We do not understand.

Perhaps Witt overstates the case, but I feel like we can agree that texts are *incredibly* "gappy," as the literary theorist Wolfgang Iser says. That is, so much of texts' intended meaning resides in metonymic implication, "what can be left unsaid," contextual situation, etc—and the further we get, culturally and temporally, from these texts, the easier it is to project contemporary schemas onto philosophy past. Not to give you homework, but you may be interested in reading the interview I did with philosopher Jonathan Livengood° around the same time I wrote the piece under discussion. We talk a bit about N&S conditions, connections between Plato and positivism, but more relevant to our current discussion, we chatted about secondary sources' treatment of their subjects. He says:

> The danger is more on the side of over-interpreting, or being overly charitable to the target. I just wrapped up a grad seminar on the problem of induction, and we were looking at the historical development of the problem of induction from Hume to 1970. As I pointed out, when you look at Hume, Hume's great, he's fun to read, but he's also deeply confused, and you don't want to do the following, which is a mistake: If you start with the assumption that Hume was just *right*, and assume that, if you're seeing an error it must be an error in your interpretation—if that's your historiographical approach, you're not going to understand Hume, you're going to understand this distorted SuperHume, who knows all these things Hume didn't know, and can respond to subtle distinctions and complaints that someone living now is able to formulate. That's not Hume! Hume didn't have an atomic theory, he didn't know anything about DNA or evolution; there are tons of things that were not on his radar. He's not making distinctions we'd want him to make, that a competent philosopher today would make. There's a real danger writing secondary literature, or generating new interpretations. If you want to publish a book on Hume, you need to say something new, a new angle—what's new and also responsible to what Hume wrote? It ends up doing new philosophy under the guise of history.

I think it's hard to litigate this for specific texts, *because* of their gappiness. We'll never know, unless/even if we have rich historiographic knowledge, whether we're being overly charitable or uncharitable. I do think your Aristotle examples are compelling counter-examples to Yudkowsky's analysis, but looking at some of the other philosophers you mention as being "woke" on concepts... there I'm a little more skeptical. (Kripke I think we should strike off the list, since he's very explicitly a Wittgensteinian in thought; ditto with many continentals.)

I think it's worth re-clarifying what I think the historic blindspots of philosophy have been, and the way I believe a style of inquiry has proven unproductive. I know my original piece is both very long, by online standards, and not especially clear structurally.

Essentially, I think that most philosophical projects which fail to appreciate the Wittgensteinian "words don't work that way" lesson will end up doing lexicographic work, not philosophy. My claim is that, with a concept like "causality" or "justice" or "beauty" (there are dozens of equally contested terms, historically), there is no "there" there. Rather, there are a multitude of continuous, analogically and historically related phenomena which are "close enough" in various ways that, with some extra specification via contextual use, these handles are pragmatically useful. If one seeks to analyze the natural language concept "causality" or "justice" or "beauty" by finding commonalities between the natural language meanings, they will end up doing primarily historic, cultural, and lexicographic work, because these word-bundles are in no way atomic, they are in no way essential. In another culture, or another language, there might be twelve types of causality or justice or beauty. They might conflate justice and beauty as a single term. How, then, does it make any sense to treat these, implicitly, as if they were natural kinds, that is, to look (as many 20th C philosophers do), for an explanation of causality that is robust to all native-English usages, but also has some deep underlying quasi-essence which can be singularly studied, analyzed, and understood? Philosophers in the know today will readily admit there are no natural kinds—species were the last example to cling to, and speciation is very messy and socially constructed, as any undergrad biologist knows. There are only continuities, at least at levels higher than particles, because the world is incredibly complex, and the possible arrangements of matter functionally infinite. (I know very little about physics here, so excuse any ignorance.) Our concept of causality, as Livengood talks about in the interview, is tied up in a long cultural history of moral judgments and norms, in folk theories and historically contingent metaphors. It is not a single coherent "thing." And its bounds do not relate to intrinsic material forces so much as they do human use. Native speakers will attribute causality in a way that is pragmatic, functional, and social.

In other words, natural language is near-useless, and often counterproductive, in trying to understand natural territories. Until recently, we might remember, plant and animal species were classified by their value to humans—poisonous vs medicinal plants, edible vs nonedible, tame vs wild animals, noble vs base beasts, etc. Imagine, now, a natural philosopher attempting to hash out a concise and robust definition of "noble animals," separate from a nominalist thread like "they're all described as noble by humans," as if there were some property inherent to these organisms, separate from their long cultural and historic understanding by humans. Such a philosopher would find out, perhaps, a *bit* about human beings, but almost nothing worthwhile about the animals.

This is the situation I see with conceptual analysis. Natural language is a messy, bottom-up taxonomy built around pragmatic functionality, around cultural and social coordination, around human life. Conceptual analysis acts as if there is a "there" there—as if there were some essence of "justice" or "causality" that maps closely to the human concept and yet exists separate from human social and cultural life. I submit there is not.

(These folk might quibble they don't believe in essences, but as I remark to Jon, my opinion here is that "a classical account of concepts as having necessary and sufficient criteria in the analytic mode is in some way indistinguishable from the belief in forms or essences insofar as, even if you separate the human concept from the thing in the world, if you advance that the human concept has a low-entropy structure which can be described elegantly and robustly, you're essentially also saying there's a real structure in the world which goes with it. If you can define X, Y, & Z criteria, you have a *pattern*, and those analyses assume, if you can describe a concept in a non-messy way, as having regularity, then you're granting a certain Platonic reality to the concept; the pattern of regularity is a feature of the world.")

We might consider the meaning of textual "meaning." It can refer to an author's intention, or a reader's interpretation. It can refer to a dictionary definition, or the effect of a cause. All these are present in our language. Literary theorists spent the 20th century arguing over whether meaning just "is" unknowable author intention or diverse reader interpretation or some formal, inherent thing inside a text. (This last position is absurd and untenable, but we'll set that aside for now.) This "debate" strikes me as a debate not over the *world*, or the territory, or the nature of reality, but over whether one sense of a term ought to be standard or another. It is fundamentally lexicographic. There are many valuable insights tucked into these incessant theoretical debates, but they suffer from residing inside a fundamentally confused frame. There is no reason for one singular definition of "meaning" to exist; "words don't work that way." Many senses have been accumulated, like a snowball, around some initial core. The field ought, in my opinion, to have separated authorially intended meaning from reader-interpreted meaning, called them different terms, and called it a day. I say "ought"—why? On what grounds? Because, while in everyday linguistic use, a polysemous "meaning" might be just fine & functional, *within the study of literature*, separating intent from interpretation is crucial, and having diverse schools who use the term "meaning" in radically different ways only breeds confusion & unproductive disagreement. It is hard for me to understand why philosophers would *ever* approach the "causality" bundle as a whole, when it is clearly not in any way a singular concept.

I know many philosophers have attempted to carve up terms more technically, in ways more pragmatically suited to the kinds of inquiries they want to make (Kevin Scharp on truth comes to mind), but many, historically, have not.

Second, any philosopher who takes edge cases seriously in trying to understand natural language does not understand natural language to begin with. Because our words are functional tools carving up a continuous material space, and not one-to-one references to real, discrete objects with essences, they are optimized *for real human situations*. Much of the fretting over gendered language, or racial language, comes because there is increasing awareness of "edge cases" or "in betweens" that disrupt our clean binaries. Similarly, Pluto's ambiguous planet/non-planet status comes because it, and other bodies in our solar system, sits awkwardly between cultural categories. There is no such "thing" as a planet. There are various clusters of atoms floating around, of many different sizes and materials, and we've drawn arbitrary lines for functional and pragmatic reasons. The best piece I can recommend on this is David Chapman's "ontological remodeling" (I quibble with his use of "ontological," but it's no matter—it shows how cultural and historical, rather than inherent or natural, the concept of "planet" is.)

I'll quote the philosopher Marcus Arvan here in the hope of clarifying my own often messy thought:

> I increasingly think — and so do Millikan, Baz, and Balaguer — that [the analytic] approach to philosophy is doubly wrong. First, it is based on a misunderstanding of language. I think Wittgenstein (and Millikan) were both right to suggest that our words (and concepts) have no determinate meaning. Rather, we use words and concepts in fundamentally, irreducibly messy ways — ways that fluctuate from moment to moment, and from speaker/thinker to speaker/thinker. A simpler way to put this is that our concepts — of "free will", "justice" etc. — are all, in a certain way, defective. There is no determinate meaning to the terms "free will", etc., and thus philosophical investigation into what "free will" is will be likely to lead, well, almost everywhere. At times, we use "free will" to refer (vaguely) to "reason-responsiveness", or to "actual choices", or whatever — but there is no fact of the matter which of these is really free will. Similarly, as Balaguer points out in another paper, there is no fact of the matter whether Millianism, or Fregeanism, or whatever about the meaning of proper names is right. All of these positions are right — which is just to say none of them are uniquely right. We can, and do, use proper names in a myriad of ways. The idea that there is some fact of the matter about what "free will" picks out, or what names mean, etc., all fundamentally misunderstand natural language.
>
> And there is an even deeper problem: all of it is hollow semantics anyway. Allow me to explain. In his paper on compatibilism and conceptual analysis, Balaguer gives the following example. Two psychologists, or linguists, or whatever are trying to figure out what a "planet" is. They then debate to no end whether Pluto is a planet. They engage in philosophical arguments, thought-

> experiments, etc. They debate the philosophical implications of both sides of the debate (what follows if Pluto is a planet? What follows if it is not?). Here, Balaguer says, is something obvious: they are not doing astronomy. Indeed, they are not really doing anything other than semantics. And notice: there may not be a fact of the matter of what "planet" refers to, and it does not even matter. What matters is not what the concept refers to (what is a planet?), but rather the stuff in the world beyond the concepts (i.e. how does that thing — Pluto — behave? what is its composition? etc.).

I understand that this critique is focused on 20th C analytic, and that your comment above is focused more on the ancients. But it seems like big picture, what we're trying to figure out is, "How well-known are these problems? How widespread are philosophical practices which fall into linguistic pitfalls unwittingly?"

Showing my hand, in the nominalist/conceptualist/realist frame, it seems to me that any frame but nominalism is scientifically untenable. Various cog-sci and psych experiments have, in my opinion, disproven conceptualism, whereas the collapse of natural kinds bars, for those empiricists unwilling to believe in the supersensory realm, realism. I do want to explore nominalism more, and probably should have included at least a paragraph on it in this piece. Many regrets! I believe I felt under-educated on topic at the time of writing, but this is a good reminder to read up. From the secondary sources I've come across, it seems like the closest analogue to the emerging modern view of language, universals, natural kinds, abstract entities, etc.

(Sidenote: isn't Aristotle a realist like Plato? Or at least, in the medieval era his legacy became such? I usually see him pitted against nominalism, as one of the orthodoxies nominalism challenged.)

My big-picture understanding of the philosophical history is that a Platonic realism/formalism outcompeted more nominalist or pragmatic contemporaneous views like those of Protagoras (or perhaps the Epicureans!). The diversity of Greek thought seems incontestable, but the "winners" less so. (It's not for nothing they say all philosophy is footnotes to Plato.) Realist views go on to dominate Western philosophy up until the medieval era, bolstered by the natural incentives of Christian theology. Nominalism emerges, and claims a non-trivial number of philosophers, but never fully replaces more realist, analytic, or rationalist viewpoints. (I include rationalism because the idea of a priori and analytic both, IMO, are fatally undermined by nominalism + the messiness of natural language.) American pragmatism strikes hard against the Hegelian rationalisms of its day, but regrettably makes little long-term impact on analytic. Similarly, Wittgenstein's warnings are largely ignored by the analytic community, which continues on with conceptual analysis into the present day, as if nothing was the matter with their methods and puzzle-like riddles. (The continentals, for all their problems, did take seriously Wittgenstein's critique. Foucault's *Archaeology of Knowledge*, or Lyotard's examination of language games, or Bourdieu's dismissal of essentialism, each come to mind.) I am curious if you'd contest this.

I am still trying to understand *why* the linguistic critiques of such riddles and paradoxes, by a philosopher as well-known and widely read as Wittgenstein, have not more widely impacted the academic philosophy community. It seems you're on my side on this one, the issues with contemporary academic philosophy, so allow me to quote some speculation you might find interesting. The first cause is likely self-selection out: whereof one cannot speak, thereof one must be silent. And so it goes with graduate students pilled on later Witt. Second are problems of selection proper: knowledge regimes, and their practitioners who have invested lifetimes in them, do not cede their own follies lightly. Meanwhile, they continue to select students who confirm, rather than challenge, their own intellectual legacies—both unconsciously, because of course they believe their intellectual legacies are more correct or important, and consciously:

> A friend who was considering applying to graduate school in philosophy once told me that a professor described what the graduate programs are looking for as follows: they want someone who will be able to “push the ball forward.” The professors want to know that their graduate students will engage with the professors’ problems in a productive way, participating in the same problem-solving methods that the professors use — for example, clarifying puzzles by drawing creative new distinctions involving obscure and highly technical philosophical concepts.
>
> Needless to say, if this is the requirement for becoming a professional philosopher, then quite a few kinds of philosophers need not apply. Such as philosophers who ask questions and resist asserting answers, or philosophers who view the adoption of dogmatic philosophical positions as arbitrary and pointless. Oddly enough, any philosopher with the perspicuity to understand the futility of the puzzle-playing philosophers’ methods will probably struggle to be heard and understood in an American philosophy department today, much less employed. In effect, a kind of blindered credulousness is now a prerequisite for entering and rising in a field that is ostensibly defined by its commitment to unrelenting critical inquiry. (src)

Still, when I learned that philosophers today still take seriously one anothers' *intuitions* (and about bizarre, other-worldly counterfactuals) as sources of knowledge about *reality*, I realized that inexplicable amounts of folly can persist in disciplines. Alas.

Regarding law, that is indeed a good example of counterfactuals shaping language, though I'm not sure how much legal definitions filter into mainstream usage. Either way, legal language really is such a rich area of discussion. Textualist views, which I would previously have dismissed as naive—"there's no *inherent* or objective meaning in the words, man! Meanings drift over time!"—have some compelling pragmatic arguments behind them. For one, a Constitutional provision or Congressional law is not the product of a single designer, with a singular spirit of intent, but rather the result of a dynamic process within a committee of rivals. A bill must pass both chambers of Congress and then the Executive chair; at each stage, there will be voters or drafters with very different intentionalities or interpretations of the wording of the law being passed. Textualism, in this frame, is a pragmatic avoidance of this chaotic, distributed intentionality in favor of the one common source of truth: the actual letter of law as written and passed. How can we meaningfully speculate, in such a system, what Congress "intended," when the reality is a kludge of meanings and interpretations loosely coordinated by the text-at-hand? A second case for textualism is that is prevents bad incentives. If a lawmaker or coalition of lawmakers can create a public impression of the intent, or spirit, of a law, which exists *separate* from the actual impressions of the voting and drafting representatives, and this intent or spirit is used in court cases, an incentive is created for strategic representation of bills in order to sway future court cases. Third, a textualist might appeal to public transparency of meaning, in the vein of the Stele of Hammurabi. A population must be able to transparently know the rules of the game they are playing. Oliver Wendell Holmes: "We ask, not what this man meant, but what those words would mean in the mouth of a normal speaker of English, using them in the circumstances in which they were used ... We do not inquire what the legislature meant; we ask only what the statutes mean." How they are understood is, from this perspective, more important than the intent—since individuals will act according to the law as understood (and not as intended).

These are the steelmen of textualism—look what happens, however, when it's applied naively:

> "Well, what if anything can we judges do about this mess?" Judge Richard Posner asked that question midway through his opinion in United States v Marshall.'
>
> [...]

> The issue in Marshall was whether blotter paper impregnated with the illegal drug LSD counts as a "mixture or substance containing" LSD. The question matters because the weight of the "mixture or substance" generally determines the offender's sentence. A dose of LSD weighs almost nothing compared to blotter paper or anything else that might be used in a similar way (such as gelatin or sugar cubes). If the weight of the medium counts, a person who sold an enormous amount of pure LSD might receive a much lighter sentence than a person who sold a single dose contained in a medium. Also, the per-dose sentences for sales of LSD would bear an arbitrary relationship to the per-dose sentences for sales of other drugs, because the LSD sentences would be, for all practical purposes, a function of the weight of the medium.
>
> [...]
>
> The majority ruling held that blotters were "a mixture or substance containing" LSD, and therefore part of its weight. "Judge Posner's dissent argued that the "mixture or substance" language should be interpreted not to include the medium, because the majority's conclusion led to irrational results-indeed results so irrational that they would be unconstitutional if the statute were not construed differently."
>
> [...]
>
> Treating the blotter paper as a "mixture or substance containing" LSD produces results that are, according to Judge Posner and Justice Stevens, who dissented in Chapman, "bizarre," "crazy," and "loony."" Selling five doses of LSD impregnated in sugar cubes would subject a person to the ten-year mandatory minimum sentence; selling 199,999 doses in pure form would not.

How did the court come to this decision?

> The Supreme Court used dictionaries to define "mixture," coming to the conclusion that a blotter fit the definition ("a 'mixture' may ... consist of two substances blended together so that the particles of one are diffused among the particles of the other") and that this was sufficient for their ruling. And yet, Strauss writes, this dictionary definition has little to do with normal English use of the word mixture, which would never call a water-soaked piece of paper a "mixture" of paper and water, or a piece of paper soaked in salt water and dried, with the salt crystals remaining, a "mixture" of salt.

A man was sentenced to decades in prison over this. The truth is that Congress almost certainly did not intend to write legislation in which selling five doses of sugar-cube LSD resulted in a higher sentence than 200k pure doses. The situation eerily echoes philosophical discourses I've come across. Chalmers, for instance, looking up "engineering" in the dictionary in order to figure out the solution to analytic's problems is not nearly as harmful as the Marshall ruling. But it equally confused. The map is not the territory, as LessWrongers are fond of saying—and justice is not found in the dictionary.

Apologies for the wall of text.

[−] **peak.singularity** 2y < 1 >

"Puzzle-playing" reminds me of Kuhn's *The Structure of Scientific Revolutions :* https://samzdat.com/2018/05/19/science-under-high-modernism/

So, that's just academia for you, except it might be worse in the Philosophy department, for all the reasons that you outline ?

[-] **Suspended Reason** 2y < 1 >

Hmmm, after giving it a day, I feel like I may have unfairly or unproductively bombarded you here, so know I won't be offended if I don't get a response.

I'll try to read some of the recommendations, and perhaps in a while I can come back to this conversation with more of value to contribute.

[-] **peak.singularity** 2y < 1 >

Plato was not a "20th Century "Platonist"" ?
https://samzdat.com/2018/01/26/platonism-without-plato/

[-] **technicalities** 3y < 10 >

> Raised in the old guard, Chalmers doesn't understand...

This amused me, given that in the 90s he was considered an outsider and an upstart, coming round here with his cognitive science, shaking things up. (" 'The Conscious Mind' is a stimulating, provocative and agenda-setting demolition-job on the ideology of scientific materialism. It is also an erudite, urbane and surprisingly readable plea for a non-reductive functionalist account of mind. It poses some formidable challenges to the tenets of mainstream materialism and its cognitivist offshoots")

Not saying you're wrong about him in that lecture. Maybe he has socialised and hardened as he gained standing. A funny cycle, in that case.

[-] **romeostevensit** 3y < 7 >

Why does the compression library need to change?

One answer is that social reality is anti-inductive because signaling regimes get saturated and goodharted. A new source of bandwidth is needed, so you use a new incompatible compression scheme.

> You're frustrated by this question for good reason; it's ungrounded; it can't be answered due to ambiguity & purposelessness. *New in what way? Same in what way?*

the functions that same and new represent can be said to be underspecified and have a degree of freedom at compile time.

[-] **Cobblepot** 2y < 6 >

Hi, I'm new to this site so not sure if late comments are still answered...

The issues you raise overlap with relatively recent enthusiasm for discussing "natural kinds" in philosophy. It's a complex debate, and one you may be familiar with, but the near-consensus view in philosophy of science is that the best account of scientific categories/concepts is that concepts are bundles of properties that are/should be considered natural kinds based not on whether they are constructed or natural (a false dichotomy) but based on whether these concepts are central to successful scientific explanations. "Scientific" here includes philosophy and any other type of rational explanation-focused theorizing, and "success" gets cached out in terms of helping with induction and prediction. So the usefulness that you ask about can be grounded in the notion of successful explanations.

Here is a paper that discusses, with many examples, how concepts get divided-and-conquered in philosophy and science: https://doi.org/10.1007/s13194-016-0136-2. One example is memory—no one studies memory per se anymore; they research some specific aspect of memory.

Author, let me know if you want references for any of this.

[-] **habryka** 2y < 20 >

> Hi, I'm new to this site so not sure if late comments are still answered...

Late comments are generally encouraged around here, and we generally aim to have discussion that stands the test of time, and don't ever shut down comment threads because they are too old.

[-] **Suspended Reason** 2y < 1 >

Hey Cobblepot. Super useful link. I was not aware of that concept handle, "conceptual fragmentation"—helps fill in the picture. Not surprising someone else has gotten frustrated with the endless "What is X?" philosophizing.

It sounds to me like this idea of "successful" looks a lot like the "bettabilitarian" view of the American pragmatists, like CS Peirce—the cash value of a theory is how it performs predictively. Does that sound right to you? Some links to evolutionary epistemology—what "works" sticks around as theory, what fails to work gets kicked out.

Memory is a really good example of how necessary divide-and-conquer is to scientific practice, I think. So much of what we think of as a natural kind, or atomic, is really just a pragmatically useful conflation. E.g., there are a bunch of things that form a set *primarily because in our everyday lives they're functionally equivalent*. So to a layperson dirt is just dirt, one kind of dirt's the same as the next. To the farmer, subtle differences in composition have serious effects for growing crops, so you carve "dirt" up into a hundred kinds based on how much clay and sand is in it.

[-] **Mitchell_Porter** 3y < 6 >

"Conceptual engineering is a crucial moment of development for philosophy—a paradigm shift after 2500 years"

This claim alone gives me confidence that 'conceptual engineering' is a mere academic fad (another recent example, 'experimental philosophy'). But I confess I don't have the time to plough through all these words and identify what the fad is really about.

[-] **Suspended Reason** 3y < 6 >

Yes, I think it all depends whether you find the criticisms of Socratic dialogue, logical positivism, and "tree falls in a forest"-type questions raised on this board since the late 00s compelling.

[-] **cousin_it** 3y < 5 >

Paper by Chalmers, maybe people will find it a good intro.

Overall I agree with your point and would even go further (not sure if you'll agree or not). My feelings about colloquial language are kind of environmentalist: I think it should be allowed to grow in the traditional way, through folk poetry and individual choices, without foisting academisms or attacking "old" concepts. Otherwise we'll just have a poor and ugly language.

[-] **Suspended Reason** 3y < 1 >

I agree, and think many conceptual engineering-type philosophers would agree, about natural language. The problem is that when you're applying rigorous analysis to a "naturally" grown structure like "truth" or "knowledge," you run into serious issues. Kevin Scharp's project (e.g.) is just to improve the philosophical terms, not to interfere with mainstream use.

[-] **tristanls** 2y < 3 >

There is a description of human intelligence that may be more suitable to discussion of concepts.

> Each column in the neocortex—whether it represents visual input, tactile input, auditory input, language, or high-level thought—must have neurons that represent reference frames and locations.
>
> Up to that point, most neuroscientists, including me, thought that the neocortex primarily processed sensory input. What I realized that day is that we need to think of the neocortex as primarily processing reference frames. Most of the circuitry is there to create reference frames and track locations. Sensory input is of course essential. As I will explain in coming chapters, the brain builds models of the world by associating sensory input with locations in reference frames.
>
> -- Hawkins, Jeff. A Thousand Brains (p. 50). Basic Books. Kindle Edition.

> The hypothesis I explore in this chapter is that the brain arranges all knowledge using reference frames, and that thinking is a form of moving. Thinking occurs when we activate successive locations in reference frames.
>
> -- Hawkins, (p. 71).

> If everything we know is stored in reference frames, then to recall stored knowledge we have to activate the appropriate locations in the appropriate reference frames. Thinking occurs when the neurons invoke location after location in a reference frame, bringing to mind what was stored in each location. The succession of thoughts that we experience when thinking is analogous to the (...) succession of things we see when we walk about a town.
>
> -- Hawkins, (p. 73).

With this description of human intelligence, a concept could be unbundled into *stimulus* (sensory input) + *reference frame*.

For example, neocortex creates reference frames for specific *uses*, which informs the linkage between concepts and use. Worth highlighting here that the same *stimulus* may be linked to different reference frames in people, resulting in different use.

Additionally, the nature of how the neocortex stores information (sparse distributed representation (SDR)) informs the stretchiness of concepts, SDRs being used for representation of both the stimulus and the reference frames.

[–] **Aapje** 2y < 3 >

Chalmers' paper just seems to be an implicit defense of weaponizing language. He even uses the term "conceptual activism" and bemoans the difficulty in making others adopt his new definitions for existing words. He recognizes that "words have power" and argues that words should be redefined to use that power to "make for a more just world," like pushing through things like gay marriage, without having to change the law. He argues that "If everyone (including judges) uses 'marriage' as if it applies to same-sex marriage, then even if historical external links say that 'marriage' still refers only to unions between men and women, this will matter very little for practical purposes."

Of course, this sentence hides a clear contradiction within itself. If everyone would truly agree, then why would the judge need to rule on it?

The completely unrealistic notion that one can get everyone to agree on that new definition just serves to shield these words from criticism that this method is undemocratic and a violation of the trias politica. In practice, we see that people weaponize the claim of consensus by excluding large groups from consideration. It's often a mere tautology: everyone who matters agrees with me and disagreeing with me shows that you don't matter, because then you are unscientific, homophobic, etc.

The paper ignores Foucault's objection that language is used for social control by the powerful. Of course Chalmers may believe that those who have the most power to shape language or will have that power with the bottled persuasiveness that he dreams of, are the ones who desires deserve precedence over others and/or the majority. Yet many disagree (although ironically, neoreactionaries probably would agree).

[–] **TAG** 2y < -2 >

Do you think Foucault would disapprove of gay marriage? I can see him disapproving of marriage....

> He recognizes that "words have power" and argues that words should be redefined to use that power to "make for a more just world"

And there's still a problem if they do?

Loaded language is used when something is decided by direct popular vote, as gay marriage was in Ireland, and when it is decided by the courts, and when it is decided by the government...because loaded language is ubiquitous. Putting the "marriage" in scare quotes loads the question one way, so leaving them off loads it the other the way. There is no neutral option.

Before you condemn something as egregious, in the sense of "bad", you have to ensure its egregious in the sense of "unusual".

You're trying to persuade me, with language... that's loaded ... and I'm trying to persuade you ... with language... that's loaded.

You used the phrase "weaponised language" , which is an example of itself.

[–] **Aapje** 2y < 1 >

> Do you think Foucault would disapprove of gay marriage? I can see him disapproving of marriage...

That is completely irrelevant. I just went with the example from the paper to demonstrate why the paper is flawed, on its terms. My claim is that 'conceptual engineering' is not a neutral attempt to understand and improve concepts, but an attempt to use language as a political instrument. Foucault recognized how language functions as such, but Chalmer's papers doesn't make that explicit, which encourages bad behavior, like portraying politically-driven concept-pushing as being politically neutral and unquestionably good.

We often see people claim that their subjective beliefs are objective and then (try to) abuse their power to force others to treat those beliefs as objectively true, good or otherwise superior to other beliefs. Chalmer's paper provides backing for such behavior.

His paper is the philosophical equivalent of a political science paper that argues that the problem with politics is that people disagree and that if you could find a way to make everyone agree with the author, he could make the world much more just, so the goal should be to find a way to make everyone agree with you.

Yet in actuality, this is an extremely dangerous goal. If you set out to use your power and ability to force everyone to agree with you, then it seems far more likely that you'll end up with a dictatorship where people hide their true beliefs out of fear, than a situation where everyone is truly convinced. Furthermore, one can argue that disagreement itself is needed to examine the upsides and downsides of goals, so a society without disagreement will tend to end up chasing worse goals. If so, the very goal of actually wanting everyone to agree, is antithetical to achieving good outcomes.

> There is no neutral option.

Yes, but that is exactly my criticism of the paper. It claims that it is "relatively straightforward" to design and evaluate new concepts, which is a claim that it is possible to objectively rank concepts as being better or worse. Chalmers merely sees one issue: the "difficult social project" of concept implementation.This is like arguing that it's easy to design new products and evaluate whether they are actually better, but that the real issue is getting everyone to buy that better product. Or that it's easy to design laws and evaluate whether they are actually better, but that the real issue is getting everyone to vote for the law.

It's just an amazing level of hubris and ignorance, coupled with an authoritarian mindset. Chalmers is able to decide for us what definitions are better than the ones we are already using.The only real problem is that he sees is that he can't make us use his definitions.

[–] **Suspended Reason** 2y < 2 >

You might be interested in a post I wrote on some of the ethical problems of top down conceptual engineering: https://spilledreality.tumblr.com/post/620396933292408832/engineering-doctrines

I scold Cappelan & co a bit for exhibiting some of the high modernist tendencies James Scott critiques so well, and argue for a doctrine of non-interventionism

[–] **Aapje** 2y < 4 >

One good way to think about concepts might be as a goods with network effects in a marketplace. So there is a cost to learning a concept (the price of the concept) and the concept has to be considered useful enough for people to freely adopt it. Yet not all goods are bought freely, but they can be forced on people, as well, just like concepts can.

The more people use the same concept, the higher the network effect value, similar to how beneficial it is for people to use the same fuel in their car. Yet those network effects also reduce diversity, but not all. It's sufficiently beneficial to have separate options for diesel and gas, despite the costs of having two different fuels. And just like with goods, network effects are not homogeneous, but there tend to be 'bubbles'. Aviation fuel is died to be able to tax it separately, which works because planes don't tend to use regular gas stations nor do cars fuel up at airports. Jargon has value because of these bubbles (and the lack of understanding by people outside of the bubble can be a feature, just like people choose what goods to buy in part to keep themselves in a certain bubble).

Etc.

One might therefor compare centralized and top-down conceptualizing to central planning and expect to see somewhat similar downsides.

[-] **TAG** 2y < 1 >

I didn't think your comments were very relevant to the paper as I can see little about politics in it.

> Yet in actuality, this is an extremely dangerous goal. If you set out to use your power and ability to force everyone to agree with you, then it seems far more likely that you'll end up with a dictatorship where people hide their true beliefs out of fear, than a situation where everyone is truly convinced.

And what power or force is that? Actually autocratic leaders have weapons, a monopoly on physical force. Chalmers is using words. "Weaponised terminology" is still terminology, not weapons -- despite the misleading impression the (engineered) term creates.

> "....homonymous engineering can also be extremely difficult to implement, unless one is very powerful or very lucky, or in a small community..."

So Chalmers doesn't see himself as a dictator who can impose his will. And why should be? Other people can use the same "weapons" to oppose him. He doesn't have a monopoly on power.

Chalmers, or whoever else wants to re engineer a concept has to win out in the marketplace of ideas , and that's the exact opposite of authoritarianism.

> Chalmers is able to decide for us what definitions are better than the ones we are already using, the only real problem is that he can't make us use his definitions.

Everyone thinks their stuff is best, and tries to push it. That's what youve been doing.

[-] **Aapje** 2y < 1 >

> ...as I can see little about politics in it.

When introducing the evaluation stage, the paper only mentions ethics as a way to evaluate concepts: "*And then there's the evaluation stage, which plays a central role in the conceptual ethics work by people like Alexis Burgess and David Plunkett.*"

The idea that things should be evaluated first and foremost by ethics, rather than by other means, is central to the Social Justice ideology, and is extremely political. For example, papers have been removed from journals because some people consider the findings to be unethical, rather than wrong, which is a completely different standard than was used in traditional science. The examples he gives of how to evaluate concepts are mostly drawn from SJ:

"*Miranda Fricker's work on epistemic injustice and its varieties like testimonial and hermeneutic injustice, would be a paradigmatic example here of drawing out a fruitful concept. Sally Haslanger's work on gender and race is another. A key example would be her work towards the analysis of the concept of woman in terms of oppression. What Haslanger calls ameliorative analysis is conceptual engineering in the revisionary mode to serve various ends, including the ends of social justice. This ameliorative strand of conceptual engineering has been picked up by many other people in recent social philosophy. Kate Manne's revisionary analysis of misogyny is an example.*"

None of these are apolitical examples.

> And what power or force is that? Actually autocratic leaders have weapons, a monopoly on physical force. Chalmers is using words.

I'm not claiming that this paper is a weapon, but rather that it's an apologia for weaponizing concepts.

> "....homonymous engineering can also be extremely difficult to implement, unless one is very powerful or very lucky, or in a small community..."
>
> So Chalmers doesn't see himself as a dictator who can impose his will.

That's not what the quote that you give says, so you are misrepresenting his words. He actually says that one kind of conceptual engineering can be extremely difficult in some circumstances. You falsely translate that into a claim that it is impossible, which Chalmers doesn't even claim for the specific situations that he considers more difficult than other situations, let alone in general.

In the conclusions of the paper, he states that: "*Concept design and concept evaluation are relatively tractable, but widespread concept implementation is a difficult social project. As a result, conceptual engineering on a community-wide scale is difficult, but it is possible.*"

Chalmers describes this kind of social engineering as being desirable. The difference between a classical liberal and an authoritarian is that the former doesn't sees a lack of control over the behavior of others as desirable, while the latter sees it as a problem to be solved. Chalmers is the latter.

> Everyone thinks their stuff is best, and tries to push it. That's what you've been doing.

My criticism of Chalmers is not that he is trying to convince others, in which case your criticism would be correct, but that he wants to find way to control others.

Perhaps you don't understand the difference between trying to convince people and trying to control people?

[–] **TAG** 2y

> My criticism of Chalmers is not that he is trying to convince others, in which case your criticism would be correct, but that he wants to find way to control others.
>
> Perhaps you don’t understand the difference between trying to convince people and trying to control people?

I fully understand it. The thing is that you have not made the case for control over conviction, as something that is literally true. All your arguments are based on loaded language.

> I’m not claiming that this paper is a weapon, but rather that it’s an apologia for weaponizing concepts.

"Weaponised concept" is a misleading metaphor.

> In the conclusions of the paper, he states that: “Concept design and concept evaluation are relatively tractable, but widespread concept implementation is a difficult social project. As a result, conceptual engineering on a community-wide scale is difficult, but it is possible.”
>
> Chalmers describes this kind of social engineering as being desirable

Chalmers doesn't use the term "social engineering". Your argument is a slippery slope from "conceptual engineering" to "social engineering" to "control".

> None of these are apolitical examples

I said "little" , not "none".

> The idea that things should be evaluated first and foremost by ethics, rather than by other means, is central to the Social Justice ideology,

And the Christian, Buddhist , Muslim.....ideologies. "Evaluate things ethically" is not an extraordinary claim. Even classical liberalism holds that you should evaluate things and foremost by their impact on freedom, which is an ethical argument.

[–] **Aapje** 2y

> The thing is that you have not made the case for control over conviction, as something that is literally true.

I'm not sure how one could ever prove that, in the absence of hindsight and perhaps not even then. Many people didn't expect Hitler to prosecute the Jews, despite what he wrote in Mein Kampf, so as far as I can tell, even that book doesn't meet your standard of proving that Hitler wanted to prosecute the Jews in a way that is "literally true."

Of course, you can choose to always err in favor of petting animals unless that specific animal has already harmed someone, but that policy is only feasible in an environment with very few (very) dangerous animals. Your policy is not suitable to other environments.

Anyway, my claim is based on large part on the total absence of recognition in the paper that there is any potential problem with seeking to control others and instead, his claim that the only problem is that making people use your concepts is very difficult. Then there is his approval of highly problematic and extremist social justice advocates, without giving any competing political examples, which makes it highly likely that he has extremist politics himself, based in the rather simply logic that people tend to quote and approve of things they believe in themselves.

But it doesn't really matter what he believes, because the paper is an apologia for being so one-sided, in particular in the current political climate. It is pushing an open door.

> I said "little" , not "none".

Yet I was easily able to find a paragraph with him approving of three different extremely political and polarizing examples, which he chose to use instead of neutral examples. When a very high percentage of the examples is political, that is not "little."

> And the Christian, Buddhist , Muslim.....ideologies. "Evaluate things ethically" is not an extraordinary claim.

It is/was actually crucial for human development and peace that religions are/were liberalized. For example, the Peace of Westphalia liberalized Christianity in Europe, requiring believers to stop acting on their belief that only faith in their religion would ensure eternal salvation, which they considered justification for extreme warmongering.

And that something is not an "extraordinary claim," doesn't mean that it is not extremely harmful or that it can't lead to extraordinary outcomes. Hitler's antisemitism wasn't extraordinary, but the final ~~solution~~ outcome was extraordinary. However, even fairly common outcomes can be quite bad.

> Even classical liberalism holds that you should evaluate things and foremost by their impact on freedom, which is an ethical argument.

You are again failing to distinguish the object level and the meta-level. Setting global rules that allow individuals to fairly freely make their decisions based on their own ethics is fundamentally different from desiring/demanding that each of those decisions is tightly controlled directly or indirectly by a global set of ethical rules.

[–] **Chris_Leong** 3y 3

Thanks for writing this post. Better connecting the discussion on Less Wrong with the discussions in philosophy is important work.

Also, how is the idea of conceptual engineering different from Wittgenstein's idea of language as use?

[–] **Suspended Reason** 3y 5

Though I don't know much about it, I take "meaning as use" as a vague proto-version of the more explicit theories of fuzziness, polysemy, and "family resemblance" he'd develop later in his life. In some sense, it merely restates descriptivism; in another less literal sense, it's a tonal subversion of more classical understandings of meaning.

Conceptual engineering takes a very different stance from mere descriptivism; it specifically thinks philosophers ought to "grasp the language by its reins" and carve up words and concepts in more useful ways. "Useful," of course, depends on the fields, but e.g. in metaphysics, the disambiguation would be focused on evading common language traps. In that way, it's a bit like Yudkowsky's "Taboo Your Words."

Thanks for reading!

[–] **Chris_Leong** 3y < 2 >

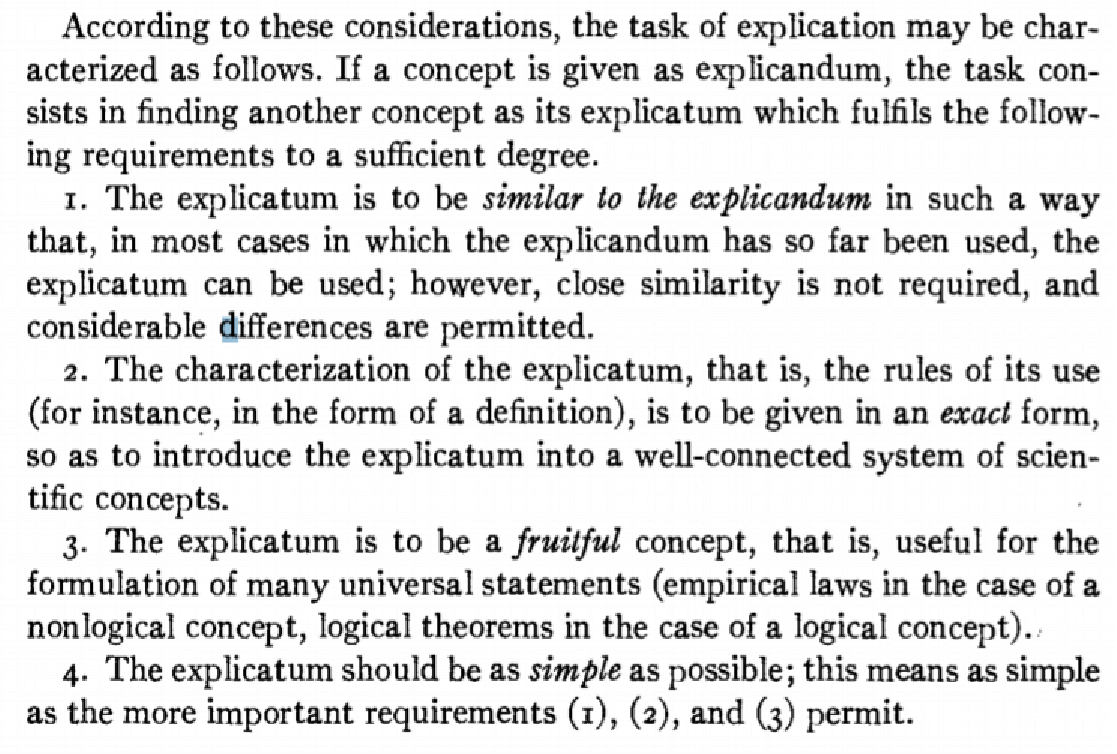
Oh, one more thing I forgot to mention. This idea of Conceptual Engineering seems highly related to what I was discussing in Constructive Definitions°. I'm sure this kind of idea has a name in epistemology as well, although unfortunately, I haven't had the time to investigate.

[–] **Stefan_Schubert** 2y < 2 >

I think that though there's been a welcome surge of interest in conceptual engineering in recent years, the basic idea has been around for quite some time (though under different names). In particular, Carnap argued that we should "explicate" rather than "analyse" concepts already in the 1940s and 1950s. In other words, we shouldn't just try to explain the meaning of pre-existing concepts, but should develop new and more useful concepts that partially replace the old concepts.

> Carnap's understanding of explication was influenced by Karl Menger's conception of the methodological role of definitions in mathematics, exemplified by Menger's own explicative definition of dimension in topology.
> ...
> Explication in Carnap's sense is the replacement of a somewhat unclear and inexact concept *C*, the *explicandum*, by a new, clearer, and more exact concept $undefined$, the *explicatum*.

See also *Logical Foundations of Probability*, pp. 3-20.



According to these considerations, the task of explication may be characterized as follows. If a concept is given as explicandum, the task consists in finding another concept as its explicatum which fulfils the following requirements to a sufficient degree.

1. The explicatum is to be *similar to the explicandum* in such a way that, in most cases in which the explicandum has so far been used, the explicatum can be used; however, close similarity is not required, and considerable differences are permitted.

2. The characterization of the explicatum, that is, the rules of its use (for instance, in the form of a definition), is to be given in an *exact* form, so as to introduce the explicatum into a well-connected system of scientific concepts.

3. The explicatum is to be a *fruitful* concept, that is, useful for the formulation of many universal statements (empirical laws in the case of a nonlogical concept, logical theorems in the case of a logical concept).

4. The explicatum should be as *simple* as possible; this means as simple as the more important requirements (1), (2), and (3) permit.

[–] **Suspended Reason** 2y 1

I've heard similar things about Carnap! Have had some of his writing in a to-read pile for ages now.

[–] **cogitoprime** 3y 2

This was a fascinating read. You may find an essay in my recent post history on the purported difference between "Greek" and "Hebrew" ways of thinking interesting.

I stumbled on this essay while reading a book that does a meta-analysis of Lakoff's "objectivism vs experientialism" and then proposes an integration as part of a larger integration project for general and domain language. It does so while also addressing implications for Machine translation and AI, but I'm neither a linguist nor an AI researcher so I find myself wondering what someone more familiar with Lesswrongian projects than I would think of it.

The integration it proposes is through the philosophy of Emmanuel Levinas, my(budding) area of expertise.

Anyways, here it is, and I'd be interested to hear your thoughts if you do read it.

https://www.dropbox.com/s/yh3ap6oeyt0lb5l/Possibility_of_Language_A_discussion_of_the_nature..._----_%28Pg_27--300%29.pdf?dl=0

[–] **Suspended Reason** 3y 1

Thanks, I will check these out!

[–] **Rudi C** 3y 2

(Comment on the meta-seq)

I think most LWers will agree that philosophy has a lot of useful concepts to contribute (and probably near duplicates of almost all ideas on LW), but the problem is finding the good stuff. It’s pretty obvious that the signal-to-noise ratio is much much better in LW versus “philosoph.” Philosophy does not even mean analytic philosophy. Most people I know who wanted to wet their feet in philosophy, started with Sophie’s World. The very fact that you’re utilizing so much expert knowledge and a bounty system to find LW’s ideas in philosophy is strong evidence against philosophy.

I still think the meta-seq project is awesome. I hope you can publish a bibliography of high quality philosophical material.

[–] **Suspended Reason** 3y 5

re: meta-sequences, thank you! It's proven a much bigger and more difficult project than I'd naively imagined, insofar as I began realizing that my own readings of the original or secondary texts were not even remotely adequate, and that I needed to have extensive conversations with people closer to the field in order to understand the intellectual context that makes e.g. the subtle differences in Carnapian linguistics vs that of other logical positivists so salient.

The project will likely end up focusing mostly on language and reality (map and territory) for better or worse. I think it's a big enough cross-section of LW's intellectual history, and also enough of a conversation-in-progress in philosophy, that it will hopefully shed light on the larger whole.

As for damning philosophy—I think there *are* some real self-selection effects; Russell has his quote about young men wanting to think "deep thoughts," that's reflected in Livengood's description of Pitt philosophy; Stove's "What's Wrong With Our Thinking" touches on some of the cognitive biases that might guide individuals to certain views, and increase the likelihood of their having a prominent reception and legacy. (One can understand, for instance, why we might be inclined and incentivized to battle against reductionism, or determinism, or relativism.) There's a certain view of philosophy which sees the Sophists as basically getting the broad brushstrokes right, and much of philosophical history that follows them as being an attempt at "cope" and argue against their uncomfortable truths—that e.g. the ethical and ontological relativism the Sophists pushed was too emotionally and morally destructive to Athens, and Plato's defense of the forms of beauty, or the just, are an attempt to re-stabilize or re-structure a world that had been proven undone. (I understand "relativism" is in some ways the nemesis of LW philosophy, but I believe this is solely because certain late 20th C relativists took the concept too far, and that a more moderate form is implicit in the LW worldview: there is no such "thing" as "the good" out in the world, e.g.) This is a very partial narrative of philosophy, like any other, but it does resonate with why, e.g., neoplatonism was so popular in the Christian Dark Ages—its idea of an "essence" to things like the Good, or the Just, or a Table, or a human being is all very in accord with Christian theology. And the way that Eastern philosophies avoided this reifying ontology, given a very different religious background, seals the deal. Still, I'd like to do quite a bit more research before taking that argument too seriously.

OTOH, I can't help but think of philosophy as akin to an aesthetic or cultural endeavour—it can take years of consumption and knowledge to have sophisticated taste in jazz, and perhaps philosophy is somewhat the same. Sure, LessWrong has a kind of clarity in its worldview which isn't mirrored in philosophy, but as Stove points out and Livengood seconds, the main problem here is that we still have no way of successfully arguing *against* bad arguments. The positivists tried with their description of "nonsense" (non-analytic or non-verifiable) but this carving *still* fails to adequately argue against most of what LWers would consider "philosofolly," and at the same time hacks off large quadrants of humanistically meaningful utterances. Thus, so long as people who want to become philosophers and see value in "philosofolly," and find readers who see value in philosofolly, then what can the more analytic types say? That they do not see the value in it? Its fans will merely say, well, they *do*, and that the analytic conception of the world is too narrow, too cold-blooded. think the real takeaway is that we don't have a good enough understanding of language and communication yet to articulate what is good and productive versus what is not, and to ground a criticism of one school against the other. (Or even to verify, on solid epistemic ground, that such arguments *are* folly, that they are wrong rather than useful.) This is a big problem for the discipline, as it becomes a pitting of intuitions and taste against one another.

[–] **peak.singularity** 2y 1

Why do you think that "relativism is in some ways the nemesis of LW philosophy" ?
(BTW, I *hate* the way the word is used, "relative" doesn't mean "equal" !)

From what I see, LW actually started out focused on Truth as the core value, but then since the community is pretty smart, it figured out that this way led to relativism and/or nihilism (?) and pretty bad outcomes :
https://samzdat.com/2018/03/07/everything-is-going-according-to-plan/

So, a strategic change of direction has been attempted towards "Winning" as the core value.

Did I get that right ?

[–] **Rudi C** 3y 1

Thanks for the long reply.

An aside: I think "moderate relativism" is somewhat tautologically true, but I also think it's a very abused and easy-to-abuse idea that shouldn't be acknowledged with these terms. I think that perhaps saying morality is "value-centric" or "protocol-based" (each referring to a different part of "morality". By the second part, I mean a social protocol for building consensus and coordination.) is a better choice of words. After all, relativism implies that, e.g., we can't punish people who do honor killings. This is mostly false, and does not follow from the inherent arbitrariness of morality.

On our inability to fight bad epistemics: I think this is somewhat of an advantage. It seems to me that "traditional rationality" was/is mostly focused on this problem of consensus-truth, but LW abandoned that fort and instead saw that smarter, more rational people could do better for themselves if they stopped fighting the byzantine elements of more normal people. So in LW we speak of the importance of priors and Bayes, which is pretty much a mindkiller for "religious" (broadly conceived) people. A theist will just say that his prior in god is astronomical (which might actually be true) and so the current Bayes factor does not make him not believe. All in all, building an accurate map is a different skillset than making other people accept your map. It might be a good idea to treat them somewhat separately. My own suspicion is that there is something akin to the g factor for being rational, and of course, the g factor itself is highly relevant. So in my mind, I think making normal people "rational" might not even be possible. Sure, (immense) improvement is possible, but I doubt most people will come to "our" ways. For one, epistemic rationality often makes me one worse off by default, especially in more "normal" social settings. I have often contrasted my father's intelligent irrationality with my own rationality, and he usually comes much ahead.

[–] **nick.olah** 1y 1

I'm taking a course on Conceptual Engineering this semester. Your description of concepts as bundles perfectly describes my understanding of the reference/denotation distinction with respect to a context of utterance. It also accounts for my frustration with Cappelen's description of "assessing and improving" concepts, while refusing to give a non-vague description of concept. Like - which concept? (theres many subconcepts). And in what context?

All of which is a very circuitous way of saying, thank you!

[-] **Morgan_Rogers** 2y Ω 0 < 1 >

It's really nice to see a critical take on analytic philosophy, thank you for this post. The call-out aspect was also appreciated: coming from mathematics, where people are often quite reckless about naming conventions to the detriment of pedagogical dimensions of the field, it is quite refreshing.

On the philosophy content, it seems to me that many of the vices of analytic philosophy seem hard to shake, even for a critic such as yourself.

Consider the "Back to the text" section. There is some irony in your accusation of Chalmers basing his strategy on its name via its definition rather than the converse, yet you end that section with giving a definition-by-example of what engineering is and proceed with that definition. To me, this points to the tension between dismissing the idea that concepts should be well-defined notions in philosophical discourse, while relying on at least some precision of denotation in using names of concepts in discourse.

You also seem to lean on anthropological principles as analytic philosophy does. I agree that the only concepts which will appear in philosophical discourse will be those which are relevant to human experience, but that experience extends far beyond "human life" to anything of human interest (consider the language of physics and mathematics, which often doesn't have direct relation to our immediate experience), and this is a consequence of the fact that philosophy is a human endeavour rather than anything intrinsic to its content.

I'd like to take a different perspective on your Schmidhuber quote. Contrary to your interpretation, the fact that concepts are physically encoded in neural structures supports the Platonic idea that these concepts have an independent existence (albeit a far more mundane one than Plato might have liked). The empirical philosophy approach might be construed as investigating the nature of concepts statistically. However, there is a category error happening here: in pursuing this investigation, an empirical philosopher is conflating the value of the global concept with their own "partial" perspective concept.

I would argue that, whether one is convinced they exist or not, no one is invested in communal concepts, which are the kind of fragmented, polysemous entity which you describe, for their own sake. Individuals are mostly invested in their own conceptions of concepts, and take an interest in communal concepts only insofar as they are interested in being included in the community in which it resides. In short, relativism is an alternative way to resolve concepts: we can proceed not by rejecting the idea that concepts can have clear definitions (which serve to ground discourse in place of the more nebulous intuitions which motivate them), but rather by recognizing that any such definitions must come with a limited scope. I also personally reject the idea that a definition should be expected to conform to all of the various "intuitions" which are appealed to in classical philosophy for various reasons, but especially because there seems no a priori reason that any human should have infallible (or even rational) intuitions about concepts.

I might even go so far as to say that recognizing relativism incorporates your divide and conquer approach to resolving disagreement: the gardeners and landscape artists can avoid confusion when they discuss the concept of soil by recognizing their differing associations with the concept and hence specifying the details relevant to the union of their interests. But each can discard the extraneous details in discussion with their own community, just as physicists will go back to talking about "sound" in its narrowed sense when talking with other physicists. These narrowings only seem problematic if one expects the scope of all discourse to be universal.

[-] **Ivo Wever** 2y < 1 >

> despite meaningful differences between Platonic philosophy and this analytic practice, I will argue that there is a meaningful through-line between them."

It seems to me this sentence is missing a negation: there is NO meaningful through-line?

> by narrowing throw out all the richly bundled senses of a concept while keeping only the immediately useful—it's *wasteful* in its parsimony. It leaves not even a ghost of these other senses' past, advertising itself as the original bundled whole while erasing the richness which once existed there. It leads to verbal disputes, term confusion, talking past each other. It impoverishes our language.

cf. Seeing like a State, James C. Scott. This is the same problem of insisting on making things legible and taking too many shortcuts while doing so, throwing out all the babies with the bathwater.

[-] **longphi1080** 2y < 1 >

This essay describes the essence of the debate around "Trans women are women too" without mentioning it once.

[-] **Suspended Reason** 2y < 4 >

Yes, Sally Haslinger and philosophers in her orbit are the go-to citations on a "therapeutic" engineering program. The idea is removing what Cappelen and Plunkett call "moral defects" from our language. I'm a little more skeptical of such programs to top-down re-engineer words on moral considerations, for reasons hopefully obvious to a reader of dystopian sci-fi or James C. Scott. I advocate instead doctrines of non-intervention & illumination:

> - *The doctrine of non-intervention*. Concepts should—in part because they can only, with any efficacy—be engineered locally. Only locals know the affordances of their specific use cases. Philosophers ought to engineer philosophical concepts, in order to straighten up their own discipline, but leave fishing concepts to the fishermen. Engineering-on-behalf ought only *provide possibilities* for bottom-up adoption; it should never limit possibilities by top-down imposition.
> - *The illumination doctrine*. Concepts should help illuminate the world, but never obscure it. This is especially important in ameliorative or ethical-political projects.

[-] **Peter Gerdes** 2y < 1 >

At a conceptual level I'm completely on board. At a practical level I fear a disaster. Right now you at least need to find a word which you can claim to be analyzing and that fact encourages a certain degree of contact and disagreement even if a hard subject like philosophy should really have 5 specific rebuttal papers (the kind journals won't publish) for each positive proposal rather than the reverse as they do now.

The problem with conceptual engineering for philosophy is that philosophers aren't really going to start going out and doing tough empirical work the way a UI designer might. All they are going to do is basically assert that their concept are useful/good and the underlying sociology of philosophy means it's seen as bad form to mercilessly come after them insisting that: no that's a stupid and useless concept. Disagreements over the adequacy of a conceptual analysis or the coherence of a certain view are considered acceptable to push to a degree (not enough imo) but going after someone overtly (rather than via rumor) because their work isn't sufficiently interesting is a big no no. So I fear the end result would be to turn philosophy into a thousand little islands each just gazing at their own navel with no one willing to argue that your concepts aren't useful enough.

[-] **Brolo_Swaggins** 2y 1

I'm reminded of a discussion I had with a friend, regarding the nature of concepts. While conversing, it occurred to me that "knowledge" can be thought of as a definition, which can be analyzed in terms of binary classification. The definition's aptness is a product of its sensitivity and selectivity, regarding how well the definition fits the intended set of examples. Loosely,

utility(knowledge) = selectivity * sensitivity, where
    selectivity = coherence = justification = information
    sensitivity = correspondence = truth = description

I'm not very familiar with the current state of Philosophy. Is this known/discussed in the field already? Is it worth making a small post about? idk if this is novel, old news, or wildly misguided. I skimmed the Stanford Encyclopedia entry, which mentions "selectivity", but doesn't seem to follow the idea as far.

Moderation Log