Situating the Contributions of LessWrong to the Philosophy of Language

S. Reason 2021

Introduction

In 2020, I offered bounties to anyone able identify clear precedents, in mainstream philosophy, for ideas advanced by Eliezer Yudkowsky or Scott Alexander as their own. This was in service of a larger project of better understanding the intellectual contribution of LessWrong, against a backdrop of criticism that LessWrong had "badly reinvented" much of analytic philosophy.

LessWrong was founded in the late 2000s by Yudkowsky, bringing with him audiences built up from *Overcoming Bias*, the Singularity Institute, and various transhumanist and futurist mailing lists. It was arguably one of, if not the, most influential online watering holes in the 2010s for philosophy—giving birth to the rationalist diaspora, popularizing effective altruism, and mainstreaming AI safety concerns. We can call the larger para-academic intellectual community, of which LessWrong is one prominent example, "Game B," in contrast with academia's "Game A."

Somewhat predictably, for those familiar with Bourdieu's work on fields and social capital, Game A thinkers have been largely skeptical of, if not downright antagonistic toward, Game B intellectual production. Distrust toward blogosphere production is frequently dismissed because it lacks both formal peer review and credential-based barrier to entry; a common objection is that "anyone can write a blog," and this makes it a fundamentally unserious and untrustworthy medium. At the same time, many Game B thinkers run the gamut from feeling alienated by, or deeply dismissive of, Game A as overly bureaucratic, motivated by symbolic capital, and/or philosophically incoherent. LessWrong rationalists and academic philosophers are two "tribes" made up of generally thought, intelligence & reflective people, focused on answering many of the same questions, who as-of writing remain deeply suspicious of one another's intellectual production. To an outside observer, it is unclear whether or which of these mutual dismissals is warranted. Has LessWrong "reinvented the wheel"? Is academia in as bad of shape as rationalists claim?

Rationalists point to various institutional & social structures that they believe subsidize incoherent beliefs among academic philosophers, and lead to selection (including self-selection) problems whereby promising undergraduates who notice the field's incoherences are uninterested in or unable to enter graduate-level philosophy. Simultaneously, advising professors are motivated to select and groom advisees who will affirm their own intellectual projects, thus ensuring their work's survival. Rationalists tend to believe that certain incoherent linguistic stances, most prominently an implicit Platonist-formalist approach to words, has lead many academic philosophers to spend careers working on "non-problems." Still, these attacks have been led mostly by a small handful of (albeit highly influential) LessWrong contributors such as Yudkowsky (e.g. "Against Modal Logics" 2008) and Luke Muehlhauser (e.g. "Philosophy Is A Diseased Discipline" 2011, "LessWrong and Mainstream Philosophy" 2011). Other members have pushed back on such criticisms, noting a long tradition of philosophers , from nominalists to pragmatists to would-be conceptual engineers, whose work on language anticipates Yudkowsky and Muehlhauser's perspective.[1]

Criticisms of LessWrong from academic philosophy have not typically been advanced by mainstream, high-profile philosophers, who, broadly speaking, have significantly awareness of LessWrong than rationalists do them. Indeed, the only remarks known to this author, by a high-profile academic philosopher regarding LessWrong, are mostly positive, coming from Dave Chalmers in a Reddit AMA. Chalmers specifically highlights the advantages of LessWrong's alienation from mainstream academic thought:

> As a professional philosopher who's interested in some of the issues discussed in this forum, I think it's perfectly healthy for people here to mostly ignore professional philosophy, for reasons given here. But I'm

[1] See comments by poke, RobinHanson, michael_webster2, Kenny, Tyrell_McAllister, Vladimir_M, Alicorn, Pjeby on the aforementioned posts.

> interested in the reverse direction: if good ideas are being had here, I'd like professional philosophy to benefit from them. [. . .] (The two main contributions that I'm aware of are ideas about friendly AI and timeless/updateless decision theory. I'm sure there are more, though. Incidentally I've tried to get very smart colleagues in decision theory to take the TDT/UDT material seriously, but the lack of a really clear statement of these ideas seems to get in the way.)[2]

Instead, criticisms typically come from highly online philosophy graduate students, who will e.g. make a claim that Yudkowsky's ideas do not constitute a meaningful contribution to philosophical discourse. "Strong" versions of the critique claim Yudkowsky's entire intellectual corpus is an unwitting reinvention, or merely confused sophistry. More moderate or even-handed versions of the critique concede that Yudkowsky has made contributions to decision theory (in the form of "timeless" decision theory, or TDT) and perhaps to the nascent philosophy of AI safety, but that his philosophy of language, as laid out in "A Human's Guide To Words," is wholly reinvention. Examples of claims in this vein:

- "The only original thing in LW is the decision theory stuff and even that is actually Kant."[3]
- "Alright, I've read a bit more into Less Wrong, and I believe I finally have acquired a fair assessment of it: It's the number 1 site for repackaging old concepts in Computer Science lingo & passing it off as new. And hubris. Also Eliezer Yudkowsky is a pseudointellectual hack."[4]
- "Eliezer Yudkowsky is a pseudointellectual and the [S]equences are extremely poorly written, poorly argued, and are basically poorly necromanced mid 20th century analytic philosophy."[5]

Better situating LessWrong within analytic philosophy, and philosophy of language more broadly, has important ramifications for emerging debates on the value of academic institutions, autodidacticism, online intellectual culture, and the standard graduate philosophy program approach (where one is first steeped in a history of the discourse before attempting to make progress on its unresolved questions). For

[2] Since Yudkowsky's contribution to decision-theory is well-documented, I have chosen here to focus on his, and the larger board's contributions to philosophy of language.

[3] @PeliGrietzer, Twitter. https://twitter.com/peligrietzer/status/1163166149607079937.

[4] @StartlinglyOkay, Twitter (since deleted). https://twitter.com/StartlinglyOkay/status/976195475241078784.

[5] @Aphercotropist, Twitter. https://twitter.com/aphercotropist/status/1249083120810246144.

example, when is it "cheaper" to reinvent rather than search out, and to what extent is the answer a function of a field's signal-to-noise ratio or general accessibility? In what ways might there be advantages to "starting blind," similar to how we think of, in hillclimbing fitness landscapes, the relative advantage of those climbing the "foot" of a new hill to surpass a discourse's current local maximum? Dave Chalmers, in the earlier-cited AMA, wrote that "One way that philosophy makes progress is when people work in relative isolation, figuring out the consequences of assumptions rather than arguing about them. The isolation usually leads to mistakes and reinventions, but it also leads to new ideas. Premature engagement can minimize all three."

Unfortunately, the initial ambition of this paper—to investigate the precedents in mainstream philosophy for Yudkowsky's work of philosophy of language, *A Human's Guide to Words* (*AHGtW*), has been thwarted on account of several complications. First, the enormous size of the task: as a relative outsider to academic philosophy myself, a thorough survey of its history would take years of labor. Second, the simple fact that many of the inspirations for *AHGtW* are openly cited by Yudkowsky himself—Peirce on "making beliefs pay rent," James on pragmatism, Alfred Korzybski and S. I. Hayakawa's General Semantics.[6] In other words, the low-hanging fruit, for this question of precedent, is freely available in Yudkowsky's own writings.

Third, the schools of philosophy which most influenced Yudkowsky are arguably of relatively marginal popularity in academic philosophy, thus it is possible for both camps' critiques to be true (that the field is deeply confused, and also that Yudkowsky's work is unoriginal). Fourth, answering the question of whether Yudkowsky's extension of these thinkers' work constitutes a "genuine" advance, or merely re-iterates what was "already implicit" there, is largely subjective. Those critical of Yudkowsky would point out that philosophers as popular and widely read as Wittenstein (e.g. in his concept of "family resemblance") anticipate arguments in *AHGtW*. Those who would wish to give Yudkowsky credit would point out that, despite pragmatist (or Wittgensteinian) arguments being made since the late 19th C, philosophy of language was nonetheless dominated throughout the 20th C by an approach known as conceptual analysis, which appears borderline

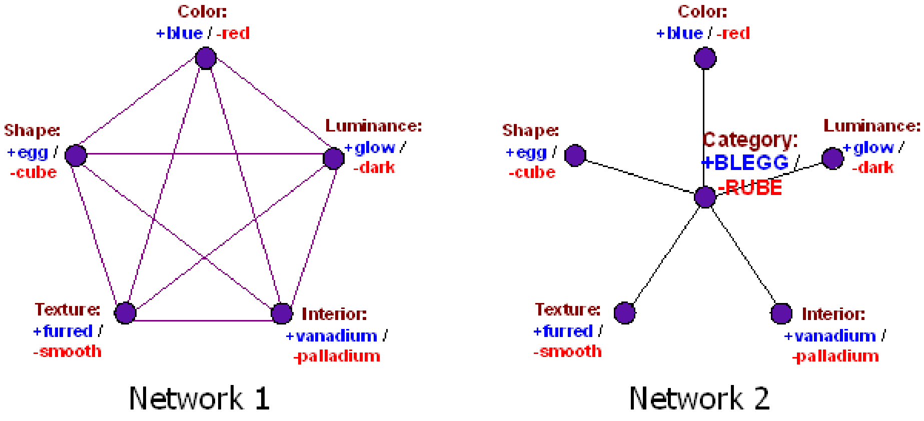
[6] Many critics of Yudkowsky have called his ideas a reinvention of logical positivism, which I believe reveals a deep misreading (or, perhaps, a shallow reading by his critics). It is precisely a logical positivist view of words, definitions, and properties that Yudkowsky argues so thoroughly against in *AHGtW*.

incoherent or nonsensical from a pragmatist perspective. Furthermore, where Wittgenstein in some sense renounces philosophy, on account of its linguistic problems, Yudkowsky lays out a number of strategies for overcoming such problems, including his concept of "tabooing your words"—a nearly identical version of which was advanced by David Chalmers, several years after Yudkowsky's post, in his paper "Verbal Disputes" (2011). It is this author's opinion that, if critics of Yudkowsky wish to undermine his contribution on the basis of its anticipation by pragmatists, then they ought to apply the same judgment to many leaders of their own discourse, who have, arguably done little to get past the point famously posed by William James in "What Is Pragmatism?" Those who read Chalmers's "Verbal Disputes" (2011), or the many takedowns of conceptual analysis launched since the 1980s, will find little that is not "merely" extrapolation on James's 1904 position. What seems clear is that such big ideas as these are rarely invented wholesale—rather, it is the reiterating, re-framing, and perpetual re-application of these ideas that constitutes much of philosophy's real work. In this light, the skepticism of LessWrong's philosophical contributions, especially coming primarily from younger, more professionally precarious members of academic philosophers—in sharp contrast with the charity extended by high-profile members like Chalmers—begins to look more like a squabble over symbolic capital than a critique of the "ideas alone."

Instead, the goal of this paper will be to begin bridging the two discourses—first, by summarizing the arguments of Yudkowsky in *AHGtW* so that they might be better understood by academic philosophers; second, by laying out for LessWrong rationalists a better understanding of how mainstream academic philosophy has evolved on the questions raised in *AHGtW*. Hopefully, this will lay the foundation for future research, and even an eventual, satisfying answer to the issues of precedent that this paper hoped—but fails—to adequately address.

Contents Overview

I cannot be comprehensive in my reading of contemporary philosophy, as this is a lifetime project. Instead, I've focused on a subdomain which I already have some familiarity with, philosophy of language and the history of analytic phil. I believe this is an appropriate choice given it reflects what one founding LessWrong user, Luke Muehlhauser, describes as the pertinent intervention of LessWrong-style philosophy into mainstream thought—namely, that it is part of a movement "to massively reform philosophy in light of recent cognitive science."[7] Yudkowsky's view on language, and the views of other LessWrong writers, is founded on a cognitive-scientific perspective, building from machine learning frameworks, Kahneman & Tversky's work in Judgment & Decision-Making, as well as theories of probabilistic cognition, to speculate about language. For instance, in his *AHGtW* chapter "Neural Categories," Yudkowsky uses his understanding of neural networks in order to hypothesize about what broad approaches to object-classification would be more or less efficient for human minds to use:



Whereas Network 1 (left), a classification system based on "necessary and sufficient conditions," requires O(N^2) connections, Network 2 only requires O(N).

This paper consists of three parts: first, a synopsis of Yudkowsky's view of linguistic meaning as advocated in *A Human's Guide To Words*. Second, an interview with Associate Professor Jonathan Livengood of Urbana-Champagne, in which he helps situate LessWrong's perspective within the existing analytic tradition. Third is an exploration of the problems of conceptual analysis in academic philosophy. The practice of conceptual analysis, and its companion ideology of "necessary and

[7] "LessWrong Rationality and Mainstream Philosophy," 2011.

sufficient conditions," are the targets of frequent scorn by LessWrong members, who see the approach as a scientifically unfounded and philosophically incoherent framework for understanding language. As we will see, it is less that contemporary philosophers would disagree with this assessment, and more that the field has failed to adequately update its methodology and approach in light of conceptual analysis's failures.

This last section traces the history of conceptual analysis (CA) and examines recent paradigm shifts in analytic, away from CA and into what is often called "conceptual engineering." I find that recent developments in analytic philosophy generally affirm the view of linguistic meaning espoused at LessWrong, but also chronologically preempt LessWrong's framework. The two cultures appear to have stumbled on somewhat similar understanding of language, but via very different paths and timelines. Though similar arguments were advanced by the American pragmatist philosophers half-a-century earlier, analytic philosophy gained its first real awareness of the problems with CA through the later Wittgenstein in his 1953 *Philosophical Investigations*—but it has been slow to update on the philosopher's arguments. Despite several interventions (notably from cognitive-scientific fronts, such as Eleanor Rosch's work on prototype theory in the 1970s), CA is still considered by many to be the default mode of analytic philosophy; indeed, the *Stanford Encyclopedia of Philosophy* entry argues that "many" view CA as the "essence" of philosophy generally. LessWrong philosophy, on the other hand, having built its foundation on cognitive science instead of analytic phil, had no dominant methodology to overturn; in the wake of *AHGtW*, it has incubated a number of cultural technologies, including discursive norms like "tabooing [one's] words," which reflect modern understandings of language.

A Human's Guide to Words

Forthcoming; work in progress.

Situating LessWrong in the analytic tradition: An interview with Jonathan Livengood

Jon Livengood is an associate professor of philosophy at Urbana-Champaign who spent time on LessWrong during its first iteration in the late 2000s and early 2010s. At the time he was a graduate student at the University of Pittsburgh, in the midst of writing a dissertation on causal inference under John Norton, Peter Spirtes, and Edouard Machery.

One of the central criticisms of mainstream philosophy at LessWrong has always been aimed at its tendency (sometimes called "conceptual analysis") to reify cognitive concepts or linguistic terms—to perceive them, in other words, as having a simple, one-to-one correspondence with regularities or features of the world (see "Taboo Your Words," "Concepts Don't Work That Way," "LessWrong Rationality and Mainstream Philosophy"). Livengood and I discuss the state of conceptual analysis in philosophy departments, and its recent replacement by conceptual engineering. We also discuss some of the problems of academic philosophy, continuities between LessWrong and analytic thought, and the status of insights like Bayesianism, verificationism, the pragmatist motto "making beliefs pay rent," and Korzybski's "map and territory" distinction.

A glossary as context for our conversation:

- *Conceptual analysis*: a method of philosophy in which a concept is assumed to have necessary and sufficient criteria which can be described simply and robustly; for instance, there might be a set of criteria which elegantly compress and describe all native-speaker utterances of a concept like "truth," "knowledge," "beauty," "memory," etc. Typically, a philosophical opponent will rebut a proposed set of criteria by offering counterexamples: cases in which a use-case of a concept does not meet the proposed criteria (or in which a non-member of the conceptual *does* meet them). Michael Bishop's "The Possibility of Conceptual Clarity in Philosophy" is an excellent if skeptical introduction.
- *Conceptual engineering*: a recently proposed shift in philosophical method, which abandons the idea of concepts as having "necessary and sufficient"

criteria, and instead of analyzing concepts, attempts to rigorize or redefine them so they can be made more useful for a philosophical problem at hand.

On conceptual analysis & the history of philosophy

Livengood: [Before we start, since we're discussing LessWrong versus more traditional philosophy...] It's not clear to me that there's any unique thing we could think of as philosophy, full-stop, or "philosophical discourse today." I think a better picture is there are a bunch of overlapping activities and pursuits; sometimes they have goals that are nearby, and a lot of behavioral practice can live happily in any of those circumstances, but the ends people have in mind are a little different. We can have a lot of shared discourse in philosophical spaces; we all go to the same conferences and there isn't much disconnect, but when you try to get into what exactly people are trying to *do* with these projects, it can come pretty far apart.

Reason: Well, perhaps one angle here—I've heard it argued that conceptual analysis is the foundational, inseparable, aprioristic mode of philosophizing that goes back to antiquity and forms a throughline from philosophy's past to present. And though it's not always stated, the implication is that by turning a leaf from conceptual analysis to conceptual engineering, you've fundamentally changed the nature of the field: what it thinks it's up to in terms of lexicography, how it understands definitions, its place in offering linguistic prescriptions versus descriptions, the factorings of concepts and how people use them, and a larger transition from armchair philosophizing to the kind of experimental, empirical work you're doing with causation. Does that sound like a resonant narrative, or how off am I?

Livengood: I think that's a popular narrative. There's a fair amount of nuance that gets trampled, but it's not a naive or amateurish view, there are philosophers I really like, such as Stephen Stich, who would give more or less this account of the development of Western philosophy. And you can definitely see elements of it in the Platonic dialogues: Socrates shows up in the marketplace, and someone runs into him, and they say something off the cuff like, "So-and-so was really courageous yesterday," or Euthyphro says, "I'm doing the pious thing by prosecuting my father for murder." And Socrates will go, "*Oh*. So you must know what *courage*, or *piety* is. Tell me about that." The structure usually looks like the other person giving a cluster-type definition, "Piety is when you do these sorts of thing—going to sacrifices, doing what the gods require, visiting the temple on a regular basis." And then Socrates says, "no, I don't want a list."

Reason: He wants the essence.

Livengood: Right, *give me the account*. And the other person realizes what Socrates wants is a definition, so they give an attempt at a definition. Socrates gives a counter-example, so they patch the definition; Socrates gives another counter-example and they patch the definition; and eventually everyone gets tired and leaves. That's the structure of a dialogue, especially the early ones.

There's something really nice about that format, and something that looks very similar to even contemporary work. One of the corners of the literature I know fairly well, the causation literature, a lot of it looks like that. Take David Lewis in the 1970s offering a counterfactual account of causation with a simple core idea: causation is like counterfactual dependence of a certain sort, or some pair of counterfactual dependence claims. And then people point out problems with that account, so he offers patches—in 1986, in 2000, another posthumously. There's a series of counterexamples and revisions to try to capture the counterexamples, and this process repeats and repeats. You wonder if the dialogue's gonna end in the same was as the Platonic dialogues: effectively people get bored with it, and move on, or if there's something like a satisfactory theoretical resolution.

There's an interesting, difficult, subtle kind of question about what the aims of that procedure really are; you'd asked when you wrote me, you used the word "lexicography" in your setup. I don't think for the most part philosophers have been trying to do, or thought of themselves as doing, lexicography. It seems to me that philosophers up until the 20th century, really, were doing one of two things. The boring older thing is doing metaphysics, where the target is supposed to be a thing "out there" in the world, and it's not so much that the project is figuring out how we use language, but about getting at whatever the thing is "out there." Think about this the same way you think about scientific things, Newton and the apocryphal apple. You say: "That thing we just saw, let's call that gravity; there are objects, and when they're unsupported, they fall." What's the right account of that? We know what we're talking about, we fixed our reference, but now we want to give an account.

It seems to me like historically, philosophers were aiming at the same type of things. You should think of Socrates as saying something like, "We've seen examples of what we might call courage, or piety—there's a *thing*, out there in the world" and here I think he's making a mistake, there's this abstract object "justice" or "piety" or "courage," and that thing I want to give an account of in the same way I give an account of gravity, or matter, or space.

Reason: The mistake being that he reifies a cognitive cluster space of "the good" or "the pious" as matching onto a discernible structure in the world, as opposed to being a garbage heap humans have found useful to call "pious" historically. Do you think philosophy that falls into that style of thought identifies and corrects its mistakes before Wittgenstein, or is Wittgenstein rightfully treated as a big deal in part for noticing it?

Livengood: Wittgenstein is tricky in a few different ways, and the 20th century on this is... contentious. There are two related things that happened where, the history is not so obvious yet, and so there are still live debates about how to think about it. There's this movement of analytic philosophy, you'll see Frege get included, Russell and Moore typically, Wittgenstein and maybe Carnap; sometimes the Ordinary Language group will get picked up like Austin; but there's this core British group that's tough to distinguish from realists.

Reason: They're rebelling from British idealism.

Livengood: And there's this focus on figuring out the meaning of terms; this is a big part of Russell's writing, for example; and there's a lot of concern with the logical structure of speech. Then there's a related phenomenon—sometimes it's smooshed together, sometimes they're separated—this idea of philosophical analysis, and this related idea of the linguistic turn. A number of people think that sometime in the 20th century there's a shift; often they're thinking of Carnap, who is very explicit about the difference between a material kind of discourse, which is how I've described Socrates—giving this account of a thing in the world, like piety—and another mode, Carnap's formal mode, which is, treating this term that shows up in our language, "piety," now with quotation marks. I'm talking about a linguistic object. And of course there's a possible further shift to paying attention to our *concepts*, which are supposed to be attached in some way to a linguistic term.

Reason: I guess one contention I'd advance is, to me, 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. I don't know, what do you think of that?

Livengood: There's a lot right about what you said, and the kinds of challenges you see in the middle of the 20th century are serious problems for this whole collection of approaches, but I think it's important to see that this kind of move, especially from Carnap, which was prefigured a bit by what Russell was doing, was an important advance because it didn't necessary reify the target of the inquiry. In some cases you might want to say, "Gravity, that's something we can responsibly talk about as existing in the world," but for other things, we might just want to talk about what our language is doing. It might just be transactional—what kind of inferences we're going to make, what linguistic acts we're gonna trade back and forth; it might not be tracking anything out in the world. So there's been a pretty serious advance from the picture you're getting from Socrates up through the 20th century, to when people start focusing on the language, and thinking of linguistic acts or the structure of the language as themselves the targets of the investigation.

Reason: It's hard to understand the history backwards; much of what past philosophers got right now seems obvious, while everything non-obvious is wrong.

Livengood: I think that's right; one of the things that's fun about doing history of philosophy is seeing how very smart people can be deeply confused about things. They have an *idea* but it's vague and mashed-up, and today you'd say, "You're running together six different things, you have to pull apart and distinguish them." It's a thing that happens a lot, reading the history.

Livengood's experience with LessWrong

Reason: I want to ask how you think of the historic state of philosophy, or what it would be like to project a historical view on the present, but I want to ask about LessWrong, so let's jump back and forth. How'd you get exposed to the community? What was your experience?

Livengood: I started reading in the 2000s, I don't remember exactly which pieces. Much of it was just self-reinforcing; for the most part, stuff that happened on LessWrong [then] seemed indistinguishable to me from high-level amateur, low-level professional discourse in philosophy? Smart graduate students, people who had really decent ideas but lacked the professional language to express it. That's the way the LessWrong community struck me at the time; I was a graduate student just starting, and it felt like, "Yeah! I'm having a conversation with other people doing the same kind of thing I'm doing." There's sometimes an impression that the people on LessWrong were doing something wildly out of step from what philosophers would ordinarily think of themselves as doing, and that was not my impression.

Reason: Both naysayers and advocates for LessWrong or Yudkowsky do often emphasize the gap like you say, and I think unless you're very knowledgeable about the field, you hear a lot of bad arguments coming out of philosophy, both historically and still today. (Sturgeon's Law.) And most philosophers worth their chops in these fields are aware of these historical arguments being flawed; they're maybe more generous, and probably see these (today obvious) ideas as highly non-obvious in their times.

Livengood: Again, the thing I said earlier, that there isn't "such a thing, fullstop" as philosophy—LessWrong [at that time] seemed fruitfully engaged in similar kinds of questions, concerns, and problems to at least some parts of contemporary academic philosophy, and parts of contemporary philosophy I like and think are non-trivial. It's not a ghettoized, small corner of philosophy; there are robust projects that are shared by a number of departments across the world that do things this way.

I would agree LessWrong does things differently, there's a house style, but it's not like the collection of theses they defend or are pursuing or developing are so far out of the mainstream that academics wouldn't recognize it as philosophy, or as being reasonable approaches to philosophy.

Romantic vs. professionalized philosophy

Reason: Well, that's why I reached out in the first place; you'd left a comment on Luke Muehlhauser's "Train Philosophers With Pearl and Kahneman, not Plato and

Kant" gesturing to this effect—that at least in your graduate program, at Pittsburgh, cognitive science was very paid-attention-to.

Livengood: The Pittsburgh scene is a little peculiar; just background-wise, at the University of Pittsburgh there are two departments which at the time were on the same floor. There's an enormous, 42-story cathedral of learning at Pittsburgh, lovely neo-Gothic, built in the 30s, and these two departments were right across the hall: there was the philosophy department, and there was the History and Philosophy of Science (HPS) department. My PhD is from the latter.

Those departments are very different in the way they think about what philosophy is doing, the way they train their graduate students, the way their courses are conducted, their faculty. Maybe the best way to describe that difference is there are two divergent attitudes of how philosophy should go, what I'd describe as the professionalized view and the romantic view. The HPS side tended to be more professionalized; you find an interesting problem, chip away at it, advance the field a bit, and at the end of a long career, you and the people you're working in conversation with will have learned something, you'll have advanced human knowledge. This is the way things have to go: most of us are not geniuses, we're just ordinary people chipping away at a problem.

And then there's the romantic view that says look, the people we read and engage with—Aristotle, Descartes, Kant, Wittgenstein—are these super-geniuses who thought thoughts nobody else had ever thought before, who shook the foundations of human knowledge and turned things upside down. This is the aim: to become one of those people.

And the difference in graduate training in the two programs is, HPS you come in, write some papers, get out in 6-8 years, get a job, everybody does that. The Pitt Philosophy program you come, think some things, try to think the deep thoughts; the very best people go on to an awesome career, the rest of you, well, we're happy to burn through a hundred grad students to find a diamond.

My sympathies are, as you might expect, entirely with the professionalized view.

Analytic communities with similarity to LessWrong's outlook

Reason: Have you read Clark Glymour's manifesto?

Livengood: Yes, so that's the other element in the mix. There are these two Pitt departments, both quite good, the Philosophy program at the time was top five in the world, and HPS program has been for a long time *the place* to do philosophy of science. And then across the street is Carnegie Mellon, which, their philosophy department is basically Glymour's construction. Whoever the president or provost was recruited Clark out of Pitt to establish a philosophy department, and Glymour's like, great, I can build a philosophy department from scratch, the way I'd want to run a philosophy department. It's a peculiar place. The way I've heard it described is that CMU's philosophy department is what you get when you treat philosophy as a kind of engineering**.** I think that's not inaccurate. I happen to think that's beautiful, a really good look for philosophy.

Reason: What would you call the CMU, HPS, maybe LSE, you can throw LessWrong in there it sounds like—

Livengood: I would include also Irvine, University of Minnesota, Indiana University sometimes has had this vibe. It's not quite positivist, but it's in that neighborhood—science-friendly, professionalized, trying to make progress, caring about mathematics and empiricism.

Reason: It's the kind of people who would've been positivists in the 50s.

Livengood: If Carnap were alive today he'd be in this camp. Whether he'd have the views he had back then, well, he probably wouldn't; we learn things, we hope that these things change minds.

Reason: I've heard this approach is also popular in Europe.

Livengood: Yeah, the LMU at Munich has the same kind of character. European programs are trickier because much of it is tied to local funding regimes, but there do seem to be more of these mathematically, empirically informed projects.

Reason: A popular metaphor at LessWrong is Korzybski's "map and the territory," though it may have gotten there via Hayakawa. Is it a good metaphor, or do its reductions actually set you back, as some detractors claim?

Livengood: I think I'm mostly a fan of the Korzybski metaphor. It's serviceable. I think it has some limitations where the map *is* the territory, which can happen when the map-making makes the thing. Here I'm thinking of pretty mundane cases, like how something being *money* depends on how we treat it, and also more controversial cases, like the construction of gender and race or the status of mathematical objects. Or do you think that misses the point of the metaphor?

Reason: Bayes, underrated, overrated?

Livengood: Hm... a bit of both. Bayesian approaches in philosophy of science and epistemology today are pretty standard. Bayesian analysis of scientific reasoning is a project that's probably overrated, at least in philosophy. Bayes in undergraduate education generally is probably underrated; I teach a 100-level intro to logic course, and I tell the students, if you take a Stats 100 class, you'll see frequentist approaches to probability, and frequentist statistical inference techniques, so I'm going to give you something different, give you a Bayesian take on it. So far I haven't yet have a student saying, well, this is obviously the way people think about probability, this is boring and I've seen it in my other classes.

Reason: We're obviously familiar with the idea of scientific progress. Ethics get described surprisingly similarly, where there's a kind of drift; whether that drift happens "on its own," in an inevitable ratchet, or whether people have to work to make it happen, is unclear; but this is the way changing norms around race, sexuality, animal rights get talked about typically. Do you feel like the shift that departments like HPS or CMU are leading, the transition from conceptual analysis, will win out or become dominant? How do you see the field a hundred years out?

Livengood: Predictions that far out are tricky. It's not obvious to me we'll have anything that look like contemporary universities in a hundred years. You asked over email about technological developments and philosophical progress, and there are lots of positive impacts there. Increases in massive online instruction, I'm not sure how that will shake out.

Academic vs. public philosophy

Reason: Last year you wrote, "I don't think philosophers are especially well-equipped in virtue of their training to help out in the current crisis. We're more like high-trained sports fencers when a general melee is breaking out. We've trained to participate in a game that has specific restricted rules, that are implicit and often hard to fathom; if we go out into the world and try to fix it playing by our usual rules, the result will be predictably bad." This seems right to me, but the question becomes, who is filling this role? We don't have literal swordfights, so it's not a big deal if human capital is channeled into play-fencing. We do have these figurative swordfights though, so the question becomes, who is filling this role in public discourse?

Livengood: I thought your list was pretty good. [*I had emailed along Tyler Cowen's comments that for better or worse, amateurs in philosophy are currently running the public-facing discipline: Silicon Valley stoicism, Nicholas Nassim Taleb, LessWrong-style rationalism and post-rationalism, ex-New Atheists like Sam Harris, psychologists like Jordan Peterson.*] It gets filled in a variety of a way, some by professional or near-professional philosophers by way of podcasts, but much of it in larger circuits are indeed filled by people like Sam Harris, Jordan Peterson, and then even less interesting people like Ben Shapiro.

Reason: Zizek seems like one of the few entries from a more traditional philosophy tradition.

Livengood: Yeah, there are a few outliers. Peter Singer has had a fair amount of popular public impact. There are other with marginal public influence, but who are clearly important, such as Martha Nussbaum or Dan Dennett. They matter, even if they're not nearly as visible as people like Zizek, or Chomsky, or Singer. I don't know how many public-facing philosophers we need in a society of this size; it does seem like, given that I'm not especially impressed by people like Harris and Peterson and Shapiro, we could use more public-facing philosophy—but there's also a question of why it is the market has taken up those individuals, whether there are just market-type demands that are satisfied by the ideas they're producing that wouldn't take up public bandwidth the way more mainline philosophical production would.

On making beliefs pay rent

Reason: I can't let you go before asking about Peirce, who you've written quite a bit about. One of the views of his that surfaces on LessWrong is a demand that beliefs pay rent. Now, I know people make a lot of the differences between pragmatism and positivism, and certainly Russell hated the pragmatists, but there seems to be a kernel or core, maybe you could call it weak verificationism, where if one person believes one thing, and another believes another thing, then there should be some observable difference that matters, something that ought to tell us who is right our wrong. That if there's nothing in the world that can distinguish between our arguments, maybe we're not in disagreement at all. Verificationism proper comes under a lot of flack these days; maybe you can suggest a better handle for the rough, generic version I'm describing; but I'm curious, is verificationism a good idea that's needed a lot of qualification over the 20th C, or is it a bad idea that got us off on the wrong foot?

Livengood: I think it's a great idea that's mostly right. It's similar to what we were talking about with primary and secondary sources: the bulk of its value lies in pretty simple statements, even though those statements aren't quite right. They have counterexamples, or haven't had enough detail built into them, but you get the gist. It's still an open question as to whether an adequate account of the verification criterion can be made to work, but I'm not sure it really matters with respect to the practical service the idea performs. Something like Peirce's pragmatic maxim, or various Positivist views, or the verificationism Quine goes in for—all of those are quite salutary attitudes to have. Broadly good, broadly healthy, and they inspire broadly good practices in our intellectual lives.

Now, when you start trying to narrow it down to a dogmatic thesis, then I'm not so sure a verificationist account of meaning is going to quite work. There are some obvious failures; A.J. Ayers' account doesn't work, it's pretty easy to kill it, and Church gives devastating counterexamples.

Reason: If we cast Ayers as a conceptual engineer, isn't he just telling us what a meaningful sentence is?

Livengood: Yes! This is more or less the Carnapian route. Carnap's accounts have not been knocked over in the way Ayers has been.

Reason: Well, I'll just ask a couple minutes more of your time: One paper I've gotten a lot out of is Michael Bishop's 1992, "The Possibility of Conceptual Clarity in Philosophy." He talks about a "counterexample" style of philosophizing that's broader than conceptual analysis, where the philosopher sits in the figurative armchair, proposes a definition, and another armchair-occupant posits a counterexample which pokes a hole in the original proposal. Much like a Socratic dialogue. Given this has been the standard method for both proposing and rejecting proposals, it seems that, if we grant prototype theory and reject classical accounts of concept—if we believe concepts are fuzzy and polysemous; that there will always be edge-cases to a conceptual carving, and there's no way to losslessly compress into a few simple criteria the high entropy use-in-the-world by millions of decentralized speakers over time—if we grant this about concepts, should we let the classically analytic rulings from the 20th C about what is "meaningful" or "true" or "knowledge" stand? Ought we revisit those debates to see if they might be useful factorings, even if they aren't necessary and sufficient?

Livengood: Yes. The best example I can give is work by Joseph Halpern, a computer scientist at Cornell. He's got a couple really interesting books, one on knowledge one on causation, and big parts of what he's doing are informed by the long history of conceptual analysis. He'll go through the puzzles, show a formalization, but then does a further thing, which philosophers need to take very seriously and should do more often. He says, look, I have this core idea, but to deploy it I need to know the problem domain. The shape of the problem domain may put additional constraints on the mathematical, precise version of the concept. I might need to tweak the core idea in a way that makes it look unusual, relative to ordinary language, so that it can excel in the problem domain. And you can see how he's making use of this long history of case-based, conceptual analysis-friendly approach, and also the pragmatist twist: that you need to be thinking relative to a problem, you need to have a constraint which you can optimize for, and this tells you what it means to have a right or wrong answer to a question. It's not so much free-form fitting of intuitions, built from ordinary language, but the solving of a specific problem.

Conceptual Engineering: An Introduction for Rationalists

Republished from an original 2020 post on the LessWrong board titled "Conceptual Engineering: The Revolution in Philosophy You've Never Heard Of." It is in part an attack on Chalmers's approach to conceptual engineering, which tellingly errs by unwittingly perpetuating a conceptual-analytic frame, despite ostensibly trying to break free of it.

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) 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.

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.