Mindbuilding Notes, Part 2 – The Deep Core Network

With about 5000 concepts in the database, I reached the edges of what I consider the "deep core network"; the network of concepts that are *strictly* defined in terms of one another, not counting things like examples or, generally, connotations. Long ago, I had sort of naïvely viewed this core as an endpoint in set-theory: "**imaginable_concepts*" or some such top-shelf ontostatus. I'll discuss that attempted endpoint more later. For the moment, the point is that the DCN is better viewed as the 200-odd concepts downstream of **imaginable_concepts,* including itself. In fact, it is where the hierarchical or hydrological metaphor falls apart, unless we view it as some kind of roiling vortex.

Although I still have about 800 loose ends in the database, plus whatever Zoë has in Pittsburgh, I have switched over to fleshing out the DCN. My rationale for this was that some of the structural choices I'm making in the DCN are already cropping up in "outer" regions, and I'll handle them more consistently if I do it all at once. So this was a pretty epic task, and as of the moment I'm typing this, I've just completed a first draft.

In a way that oddly resembles mathematics work, this project feels like it occupies a middle ground between the rote effort of "discovery" and the creative inspiration of "invention". For instance, let's say I want to define the concept **homeostasis*. There are many different plausible ways to do that, starting with a dictionary definition like:

the tendency toward a relatively stable equilibrium between interdependent elements, esp. as maintained by physiological processes.

or

the tendency of a system, especially the physiological system of higher animals, to maintain internal stability, owing to the coordinated response of its parts to any situation or stimulus that would tend to disturb its normal condition or function.

or

the maintenance of metabolic equilibrium within an animal by a tendency to compensate for disrupting changes

These are all legitimate definitions, and roughly synonymous even though they are worded differently. But they don't help Sphinx much. In the first place, they are overly long, aiming for human-readability rather than the convenience of a machine. Much worse, dictionaries present multiple definitions for polysemes, but they do not specify which of those definitions is referenced when they use a polyseme in the context of another definition. For instance, the first definition above uses the word "elements". In the same dictionary that definition comes from, I can find twelve definitions of "elements", seven of which could plausibly be the referent the definition of homeostasis. These ambiguities build up across definitional chains, particularly where narrow definitions become metonymic for broader definitions of the same word (as is certainly the case with "elements"). The ultimate result is that we get significant slippage in meaning. For instance, in the first definition above, homeostasis is typical of "physiological processes". In the second definition, it is typical of the much narrower "physiological system of higher animals". In the third definition, homeostasis is associated with *any* animals, but strictly so: there is no suggestion of homeostasis outside of animals.

So much for dictionary definitions. I am also trying to avoid the route of clever-and-radically-counterintuitive definitions. It is all very well for Wittgenstein to say "the world is the totality of facts, not things", and I appreciate the similarities between his project and mine. I could use Plato's definition of **good*, and radical feminist definition of **consenting*, and so forth. But I am trying to define concepts in a simple descriptivist fashion; I'm not comparison-shopping for Ultimate Truth or even good politics. There will be time enough to *discuss* the topics later.

So...I went with the following definition of **homeostasis*:

*[activities] *tending_towards[object] *equilibria [typically] *biological_processes*

I'm pretty happy with that definition, but it's important to note that it wasn't the only possibility. A completely different version might be:

*[object] *stasis [of_or_pertaining_to] *dynamic_systems*

While these two definitions don't share any immediate downstream concepts(!), it is fairly likely that they would do so within two or three generations. That is to say, my definition of **equilibria* would probably come back around to mention **dynamic_systems*, or vice versa. So while I can make minor changes, I don't have much real creative freedom. If I define homeostasis as:

*[object] *american_television_shows [emphasizing] *vampires_as_metaphors*

then I'm simply wrong. So while I am designing the DCN from scratch, there is also a constant sense that I am revealing something like an architecture that already exists in our shared ontology. I want to be clear that I don't mean the DCN exists as some *eidolon,* or even in actual human cognition. It wouldn't surprise me if some of the features in the DCN, or in any AI effort, do mirror human cognitive architectures, because that is after all what the object of the game is. But human brains are far, far more complex, and they're also quirky in ways we wouldn't necessarily want to emulate. My own emotions, for instance, are profoundly affected by weather, lighting, and my digestive system; it's hard to see why a designer would want to emulate those architectures.

My initial strategy was to build an actual network diagram. I love that format, but neural networks are usually much too large to actually be schematized in a useful way. I have a whole pet peeve about people presenting "maps of the internet" or whatnot, and no doubt the map encodes five zillion and twelve data points, but it always just looks like a kid scribbling with the intent to turn the paper black. What are we really supposed to learn from that? On the other hand, 200-or-so is just at the upper limit of what I might be able to usefully diagram, so that's a little grail I was aiming for. As you'll soon see, I haven't attained it.

Pretty soon, my diagrams were unreadably complex, so I moved to Plan B: printing out lists of the leading edges, and taking them with me biking. I would ride up a hill, thinking about how to define **tending_towards* or **equilibria*, and then I would stop at the top of the hill and write down my notes, etc. And then I would go home and cut some stone and and take a bath and make dinner. My own crazy little triathlon, and I certainly enjoyed it.

Given that I have very little control over what is in the DCN, it has been interesting to see what shows up, and what's absent. Most of it has to do with ontology, parts vs. wholes, pattern-recognition, dynamics, semiotics, causality, and intellection: I expected that. There's a surprisingly (to me) large amount of math, mainly in terms of concepts that have far-reaching analogies, like "gaining" or "decreasing". Four numeric concepts appear: *0, *1, *2, and *2_or_more. There's a good deal of temporal orientation and spatial metaphor, and some very basic concepts in biology (living things) and physics (matter). There are only a few concepts that don't have a concise English equivalent: the ones that come to mind are **gen_texts* (for the broadest, most pomo version of "text") and **applicability_to_new_concepts*.

What isn't there, of note, is the Sphinx's self-concept (although the idea of "self" certainly occurs), the notion of humans or computer programs, the notion of memories, or any emotions. This last point especially surprised me, given my firm belief that emotions are central to cognition. Again, these omissions are not intentional on my part; like Laplace I could say "I had no need for those assumptions..." in the mechanics of recursive definition. They do, in retrospect, make a bit of sense. It is certainly possible for someone to have a coherent worldview without memories, and I think a great many Taoists and Buddhists would be comfortable with the idea that we can have a worldview without a self-concept. Emotions are, I'm still convinced, fundamental, but I guess I'm comfortable with the idea that they aren't as fundamental as ontology. If someone seems to have a solid grasp on reality, but is emotionally blank, we might consider that lamentable, but we still have a great many social roles for them. Indeed, there is a very well-endowed error considering such persons *gifted.* On the other hand, if someone seems emotionally typical but has no grasp on reality, we consider them psychotic, and we have no role for them in society.

I'm not quite finished. There are a handful of minor loose ends, and I suppose its not impossible that desire will get looped in on the outskirts of "choosing". But I doubt that. Some of the conjunctive links in the DCN are not themselves defined, though I have avoided any "magic conjunctions" that don't appear elsewhere in the database. Initially, the idea that *[of_or_pertaining_to]* was itself undefined bothered me a great deal. I've gotten over that, although I have big plans to go back through and weed out any unnecessary conjunction types. There are two lines of reasoning for my recent peace of mind. The first, though it is perhaps a *tu quoque* argument, is that human beings can't define words like "of", either. (I am reminded of Clinton's marvelous sophistry: "that depends on what your definition of 'is' is...")

More generally, I am becoming ever-more comfortable with the idea of recursive definitions. I remember that when I was a child, I had a joke dictionary of some sort that contained the following bit of cleverness:

circular reasoning	*see "reasoning, circular"*
[much later]	
reasoning, circular	*see "circular reasoning"*

Many years later, I had a little triumph in convincing a whole room full of lawyers than we could not use the word "harassment" in a policy definition of "harassment". That impulse is central to Searle's complaint in *The Chinese Room,* but I'm not so sure I feel it anymore. Consider a version of the joke above:

*yin	=	*[the_opposite_of]*	_*yang_
*yang	=	*[the_opposite_of]*	_*yin_

It is tempting to say that this is a Chinese room; there's no real understanding here. But if someone then poses the question "Day is to night as yin is to ____?" we are now equipped to answer the question correctly, which clearly is within the scope of what it means to understand things. In effect, the circular definition of _*yin_ and _*yang_ has provided us with a function definitional of "opposites". There are other ways to define yin and yang, of course, but there are concepts like "home" or "art" or "consciousness" that are difficult to define non-recursively; at best you just create a slightly wider circle for your circular reasoning. And then the same argument applies:

**kingdom*	=	*[supercategory_of]*	**phylyum*		
**phylum*	=	*[supercategory_of]*	**class*	*[subcategory_of]*	**kingdom*
**class*	=	*[supercategory_of]*	**order*	*[subcategory_of]*	**phylum*
**order*	=	*[supercategory_of]*	**genus*	*[subcategory_of]*	**class*
**genus*	=	*[supercategory_of]*	**species*	*[subcategory_of]*	**order*
**species*	=	*[subcategory_of]*	**genus*		

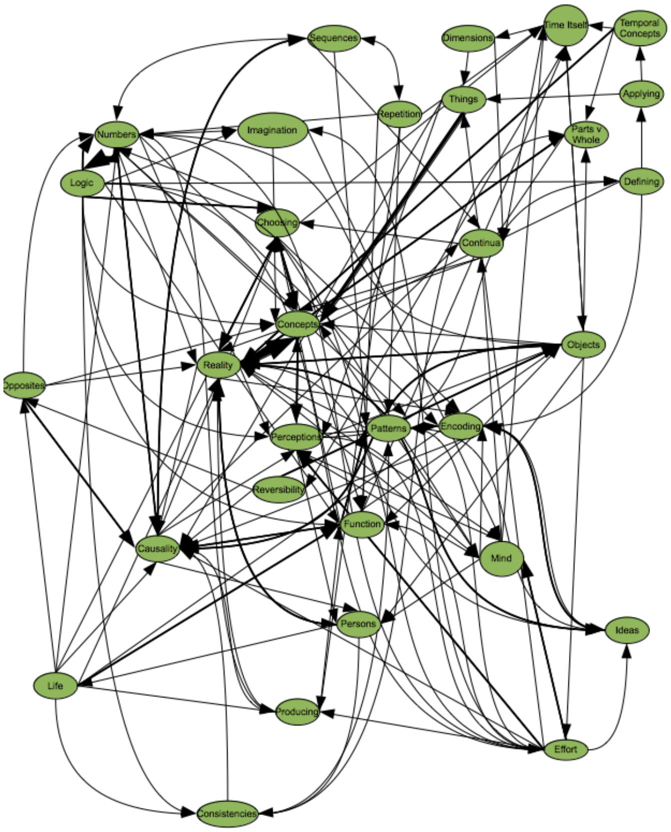
or

*north_america	=	[north_of]	*central_america		
*central_america	=	[north_of]	*south_america	[south_of]	*north_america
*south_america	=	[south_of]	*central_america		

From one point of view this is all meaningless recursion. And yet we suddenly we can take a phrase like “there is only one species of marsupial in North America” and begin to make a whole series of correct inferences about it.

Again, I'm looking at a DCN of about 200 concepts, defined by circular reasoning, or at least maelström reasoning. There is a qualitative difference, though. With yin and yang, or a cladistic hierarchy, we have the cognitive tools to step back and visualize the whole structure in terms of other metaphors. We have, after all, physical analogies for the ontological structures we are talking about, and we understand them very well: rings, spirals, tree trunks, grids, etc. By the time we are talking about an asymmetrical, shifting network of several hundred concepts, all those bets are off. Even the words we use, like “network” or “web” or “vortex” imply qualities of symmetry and centrality that are misleading.

A compressed version of the DCN looks like this, with each blob enclosing 1-18 internally linked concepts:



Sort of like a hairball from a cat who just ate an epistemology text. And that's the compressed version; the whole thing covers my sun room floor in little cardboard cut-outs. I took a picture, but you can't even make out the individual nodes. Now, I could probably get some fancy 3D diagramming software and spend a million hours cleaning it up, and make a GIFset showing the path of a single definitional sequence bouncing around in there endlessly. But when I'm done, I wouldn't be any better off, because I still don't have any mechanical reference for this sort of “network”.

And that, I think, is the hat trick of consciousness. There is a sort of naïve version of consciousness theory that laments how we are mapping the labyrinth, and looks for some tiny wellspring of magic in the middle of it, as-yet undiscovered. A version of the critique of AI takes the same view, since it is clear enough that there is no magic well inside a hard drive. But these both miss the point: the *simplest possible mapping* of the labyrinth is still much too complicated for us to really understand.