The Nature of the Soul

hivewired.wordpress.com/2019/07/24/the-nature-of-the-soul/

Saede Riordan

July 24, 2019

Epistemic Status: Speculative. Experiences and conjectures based on them.
Content Warning: Neuropsychological Infohazard, De-Biasing Infohazard
Recommended Prior Reading: Falses Faces, Building up to an IFS Model, Highly Advanced Tulpamancy 201

I said I'd return to my continued exploration of self "soon" when I wrote The Silence Hidden in the Sound in September. Well, sometimes soon has a way of becoming a nearly year-long ordeal during which large chunks of your life and many things you took for granted are ripped up, burned down, violently restructured, and shaken vigorously until it feels like years have passed since you've successfully written anything.

This pressure cooker environment had one useful effect in that it forced a lot of interesting system things to the surface in a form that made them really obvious and easily poked at, which brings us to our topic today.

There are many conflicting recent models of the self, and here I'll be talking about my attempts to syncretize some of these models into something coherent and then apply them to my own experiences.

The Neurons As Agents Model
The first model is the Neurons Gone Wild model discussed in earlier tulpamancy essays. In very brief, the Neurons gone wild model of cognition advocated by Dennett, Simler, etc is that the concept of a central self-agent, an optimizing force that could be referred to as "I" is an illusion that breaks down on analysis into a mess of conflicting, competing, and cooperating subagents, which then themselves break down into more competing subagents, and so on, and so forth, all the way down to the level of individual neurons competing for resources in the brain. There is no master coordinator, no central organizer, no source of willpower from all these agents derive, the power of a particular subagent or cluster of subagents is determined by its ability to negotiate with the agents around it, to form alliances, gang up on conflicting subagents, and distort cognitive power and neurotransmitters in the direction of its cluster of cells.

This is the most strongly no-self of the various theories I'll be looking at. It says that the illusion of unity is just that, an illusion, the self-agent exists after the fact at the narrative layer, and is used to rationalize the decisions of lower level subagents. This does a good job of explaining things like addiction, parts of the addict's mind don't want to do heroin, but other parts of his mind *do* want to do heroin, and those parts are in competition with each other.

The Core and Structure Model

The second model is The Core and Structure model. I first encountered this concept on this blog and I'm unsure if it's an original creation of the author or if it's sourced from somewhere else, but according to Ziz:

> *"Core is something in the mind that has infinite energy. Contains terminal values you would sacrifice all else for, and then do it again infinity times with no regret. Seems approximately unchanging across lifespan. Figuratively, the deepest frame in the call stack of the mind, capable of aborting any train of thought, everything the mind does is because it decided for it to happen. It operates by choosing a "narrative frame", "module", "algorithm", or something like that to run, and is responsible for deciding the strength of subagents. There are actually two of them. In order to use some of my mental tech, they must agree."*

Conversely:

> *"Structure is anything the mind learns and unlearns. Habits, judgement extrapolations, narrative, identity, skills, style, conceptions of value, etc. Everything but actual values. It lacks life on its own, is like a tool for core to pick up and put down at will."*

Under the core/structure model, everything from tulpamancy to self-help is related is relegated to the narrative and structural layers, as a set of strategies for building and using and manipulating structure. This is the most self-centric of the models, and basically proposes that everything in the mind is under the control of *something* and fundamentally everything we do, we're doing because we think it will be a good strategy to achieve our values, and the values are the thing that exists at the bottom of the stack.

The Internal Family Systems Model

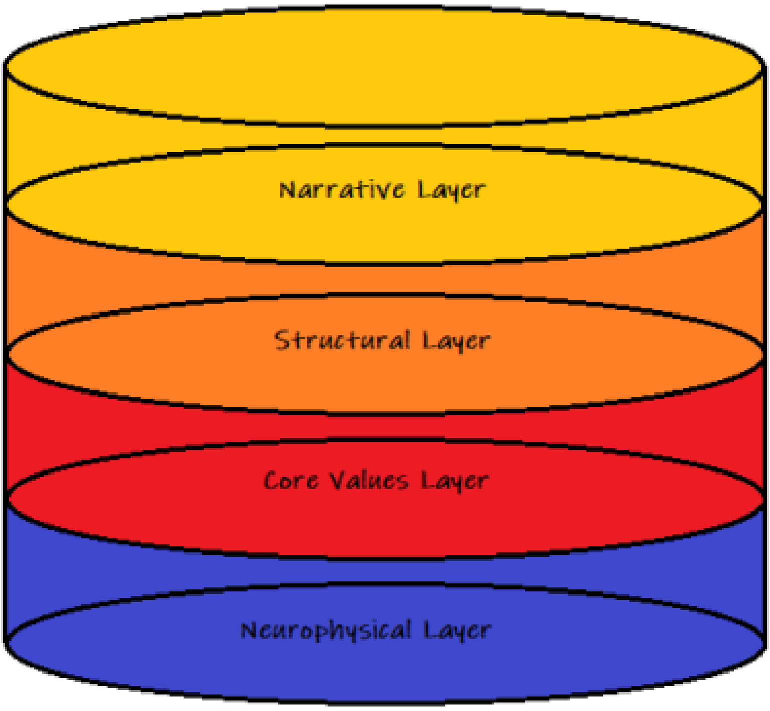
The last model we'll be comparing is the IFS model I first described in the Silence Hidden in the Sound post. IFS Goes a bit more into the gears of structure, saying that we have managers trying to keep your life in order and micromanage to prevent bad things, firefighters trying to deal with bad things when they happen and shield you from harm, and exiles which you have kicked out of your sense of self which the rest of your mental system tries to manage and keep buried and under control. IFS also has a self which acts as a central coordinator for all the parts and embodies, ahem, curiosity, connectedness, compassion, and calmness.

The IFS model comes off as fluffy and idealistic to me in its description of self, but it's model of how subagents interact especially under suboptimal circumstances, seems rather useful, and it's a useful model for things like PTSD, which could be modeled in some sense as a Guardian pattern matching a situation to one which generated the PTSD exile and responding accordingly. The building up to an Internal Family Systems Model post which I also linked in the recommended reading gives a good overview of this.

Three Models Collide

I think these three models lie somewhere orthogonal to each other. They don't actually conflict except in a few places, they simply delineate different parts of the territory, and amalgamating them will yield interesting results.

First, there's the layers thing. All three models do things with layering. I think this roughly shakes out to something like the narrative layer, the structural layer, the core value layer, and the neurophysical layer.



So tulpamancy, the naive sense of self, your life story, and most conscious attempts to manipulate the inside of your mind, exist in the narrative layer at the top. You're telling stories, inserting what essentially amounts to operating systems into the working memory environment. This self-storytelling factor is what lets us connect the past to the present to the future, remembering (in the form of a story) the past, and extrapolating (in the form of a story) into the future.

The neurons as agents model says that everything below the narrative layer breaks down into subagents doing various things, we can syncretize that decently with the IFS model of various component types, but that leaves us with the IFS self, and the core model would say that all these subagents and components, trigger, action, response, all of that, would be part of the built-up structural layer, with the values lying beneath it.

IFS probably has the most gearsy of any of the models of the structural layer, but IFS thinks that the values in the core value layer are pretty much always the same and always good. That seems naive and wildly optimistic, conversely, Ziz thinks most people's core values are evil (by her own standards admittedly).

Without putting moral valence to it the way Ziz does, I think she's probably more correct on core nature then the IFS model is. Core values could be described as primitive values, the systems that we evolved with, our most rudimentary desires encoded at the deepest levels.

So according to the Core/Structure model, all the structures we build, from studiousness to morality to learned trauma response patterns, could be thought of almost like electrical transformers, stepping down the current of willpower through successive layers of justifying things to ourselves, rationalizing, and self-deception, we reign in our values using structures to make them socially acceptable and legible, and to signal our value to the group, and thus make ourselves subservient to the group.

The core structure model pushes a particular angle really hard, which is the idea that every action and behavior is *purposeful*, everything that a mind is executing, it is executing for *some reason*. I don't disagree with this, but the idea that this cooks out into any coherent set of values that could be ascribed to something like an agent, that's where I think the first disagreement I have with it comes from.

The core/structure model also seems to posit that core is something relatively static, your values are your values, you come preinstalled with them and they don't really change. I don't really agree with this either, and I think what things someone values at the bottom-most layer will in fact change and transform over time as they are subject to outside environmental forces, and I don't think you can really get 'under' that environmental optimizing pressure because there's nothing there *to* get under, at that point you're talking about things that act directly on the neurophysical layer.

So I think my main point of contention with the core/structure layer is the way that the author conceives of core. This is really the same objection I have to IFS but in the other direction. Ziz says the core of most people is evil, IFS says the core of everyone is good. So, without ascribing morality, what exactly is the core? What's going on here?

The no-self model, that is the neurons-gone-wild model and the buddhist model, is that there is no core, or nothing that could be described as a core distinct from the subagent layers above it. Core vs no core is a pretty fundamental difference to try and cut across, but even more so, Ziz's model specifies that people have specifically two cores.

I find this interesting if for no other reason then that it seems like the most direct intellectual successor to the bicameral mind concept proposed by Julian Jaynes.

However, Ziz's cores as clusters of values and traits seem kind of arbitrarily complex to me. I could understand although perhaps still not agree with, a model of the duel cores that specified values along lines that could be differentiated into the traditional left brain/right brain dichotomies. Instead, however, the way Ziz seems to generate clusters appears more tied to her moral ideas than anything else. Without the heavy-handed morality to differentiate what values go into which core, there doesn't seem to me to be a lot which would delineate why particular values end up clustered how they are, or why there are two cores at all. Why not three, four, or even more?

For my own part, three narrativized, active agents which consciously communicate seems to work the best inside my own head. Does this somehow cook down into two cores, or do I have three cores? Hard to say. If someone's mind is best organized as a singlet, are they single-core?

This is where I think the "number of cores" idea kind of comes apart. I think I'm a bit more comfortable saying "there are core values, they sit underneath the mental structure" without specifying what the core values are, how many of them there are, which of them wins when there's a conflict, or how they interact at all, then to try and specify a model that declares how any of that stuff plays out.

I am comfortable saying that the higher layer structures are *how* it plays out though, and since high layer structures can vary drastically from person to person, so too can the shape that their core values take. Everything is connected to everything else, and signal can flow both ways down those connections.

Where does this all leave us would be cognitive architects? If this is so then there's no ground to stand upon, only clusters of values, mental alliances of convenience, and balanced power structures. If you push, something pushes back, your body auto-balances itself. Given that, what method is there to really change something in your head, and is that even possible?

I think there is an answer here, but I want to let people ponder it and percolate before giving my own answer. We'll return to this topic after hopefully not nearly as long of a pause as the last one.