Toward A Working Theory of Mind

This is the first in a series of essays that are meant to lay out the models I've developed over the course of the last seven years of my practice. "My practice" is a coaching and guided self development technique, and much of its efficacy relies on the model of mind I use to generate introspective queries. Working with specific assumptions about the basic capacities of mind allows me to note when someone seems to be operating at less than full capacity, and to launch a targeted investigation into why, so that we can work in a structured way toward reclaiming capacity and spurring growth. This is especially helpful in cases where the person feels they're at a natural ceiling, as it gives us kind of second opinion to work with instead of accepting felt limitations as absolute.

While I find the model extremely functional as a framework, I don't believe it to be at all the last word on the nature or structure of the mind. True accuracy in this domain would be a tall order. I like to tell people that the domain of psychology is all observer effect—the mind is an extremely pliable thing, and this fact combined with the fact that it must be its own observer makes it especially inclined to appear to behave in whatever way one happens to believe it will.

Nevertheless, it's possible to refine theory via application, even in psychology. Applied models of mind can easily generate interface structures that temporarily confirm an incorrect model, but a practice that fails to produce stable and significant experiential and behavioral change must ultimately have its underpinnings revisited. The model that I use now has been developed and clarified by years of application. Its ontological claims are still deeply philosophical in nature—in the limit, they are as much the domain of theology as they are of psychology—but insofar as I have been able to test and develop them practically, I have done so. This is not meant to be a theological document. The framework is merely meant to provide a reference against which one can check one's psychological position.

As far as I am concerned, the truest metric in life of a person's mental health and internal alignment is the degree to which they are successfully manifesting their desired state of the world. I have found that all purely psychological models of mental health and spiritual advancement are vulnerable to a kind of abuse that I have become very wary of: it is extremely tempting, in a world as difficult as the one we live in, to aim for dissociation and call it enlightenment. Dissociation can come in many different aesthetic skins—irony, positive thinking, yogic bliss, universal love, stoicism—but it is the same de-correlation of psychological state from reality in each case. It can be a very attractive option when it seems that the only other option is the despair and volatility that arise from attachment to the world—in fact, it can be difficult to see how anything other than some form of dissociation could constitute mental health.

I believe in the possibility of healthy investment in the world. It is what I want for myself and for everyone I work with. The framework that I am about to lay forth, however, offers a purely psychological model of what is possible for a human mind, and makes no claims about what counts as a good state of the world. This makes it vulnerable to the kind of abuse I mentioned above. If you are inclined to take psychological or ethical or spiritual models as your primary prescription—if you are the kind of person who seeks gurus or ideologies, secular or otherwise, in order to attain a moral state irrespective of any material state—please go away. My work is not an invitation to dissociate. It is an invitation to terraform your world.

Without further ado,

1. The Fundamental Nature and Ontology of the Mind

Let's start off with something you'll have heard many times before: that your experience of the world is self-created. You can know that you are experiencing sensations and interpreting them as a tree or a house or a chicken, but you have no way of knowing what the things you're seeing and hearing and touching actually are, or indeed whether the sights or sounds or feelings actually correspond to anything outside of you at all. The experience of a chicken is an imposed interpretation on a field of sense data.

Introductory cognitive science classes like to demonstrate this fact by drawing attention to the ways that interpretations can change or break—for example, simple optical illusions like the Necker Cube provide very mild examples of the way a set of sensations can be multiply interpretable. The examination of processing disorders like aphasia can tell us interesting things about how our conceptual constructions are composed by showing us what happens when different parts of a person's ability to perceive or interpret sensations break down. You may have encountered a TED talk at some point by Jill Bolte Taylor, a neuroscientist who experienced a stroke and described the experience of losing her ability to interpret her sensations. For Taylor, this was something of a religious experience. This makes sense—many of the forms of enlightenment that meditation traditions pursue include (or just are) the realization that the interpretations are imposed on the sensations, rather than the world actually objectively being the way the person interprets it. "Maya"—the Hindu concept of Illusion—is the interpretation assumed to be reality. The basic ideas of the constructed nature of experience and the unknowability of what's beyond it are ancient. All good philosophers note the fundamental inability to know that there is anything outside of their experience as their base epistemic position.

The framework I use in my practice makes distinctions between a few fundamental aspects of mind that provide me with a basis for modeling what is constructed in experience and how. Let's run through them:

The *awareness* is just experience. You could theoretically have awareness with no sensations, which might be similar to what people imagine the experience of nothing to be—though this would not actually be the experience of nothing, because the awareness would experience itself, and the awareness itself is not nothing. You can't actually experience nothing. Awareness is not constructed by the mind; it is there whether the mind is constructing interpretations or not.

The *sensations* are experienced by the awareness, but it's important to make a distinction between the sensations and the interpretations of the sensations—for example, "hand" is not a set of sensations, it's an interpretation imposed upon some colors and temperatures and proprioceptive data. The colors and temperatures and proprioceptive data are the sensations. You could imagine an awareness that experienced those colors and temperatures and proprioceptive data but did not interpret them as a hand, or as anything. In some sense, the experience of completely uninterpreted sensations might also be similar to some idea of the experience of nothing—you might think of it as "pure noise," except that even "noise" is a concept, and so the evaluation of the uninterpreted sensations as uninterpreted would not be part of that experience.

The question of whether the sensations are constructions of the mind is the same question as "does any of this exist or am I hallucinating all of it?" This is an impossible question to answer with certainty. I happen to believe that sensations do in fact correspond to realities that are (at least partially) independent of the observer. I think it may be the case that *all* sensations correspond to some reality that is partially independent of the observer, even the imaginations—imaginations may just be very heavy conceptualizations of relatively minimal sets of sensations. That is, the difference between the visual perception of an elephant and the imagination of an elephant may just be that the concept "elephant" is being used to interpret light hitting your eyes after reflecting off of an elephant in one case, and being imposed on a very minimal set of random visual and/or proprioceptive sensations in the other case. In the first case, the sensations being interpreted as "elephant" correspond (potentially) to an actual elephant, and in the second case they are merely part of the default sensation landscape of the body, and it is the body's state that they correspond to.

The *concepts* are that with which you interpret the sensations. In this document I'm going to talk constantly about concepts, interpretations, models, beliefs, and constructs. These are, roughly speaking, all the same stuff. A concept is like "hand" or "tree" or "house" or "government"; a model is a complex, dynamic, structured interpretation of a dynamic data set, e.g. a model of economics, or a model of your mother. These will be contextually interchangeable, but sloppily speaking my typical usage will be something like—a concept is analogous to a word, a belief is analogous to a sentence and a model is analogous to a book. An interpretation is the use of a particular concept or set of concepts to parse sensations—just as we can theoretically think about about sensations separately from concepts, so too can we think about concepts separately from sensations.

Actually, as a pedantic aside, I'm not sure we can literally do that—I work with the assumption that the use conditions for a concept are its application to a set of sensations, and so in order to think about "a concept distinct from sensations," you must be applying that concept to some set of sensations. The sensations can be extremely minimal—some people who think the thought "a concept distinct from sensations" might generate a visual representation of a conceptual framework in empty space, while others might merely refer to points in proprioceptive space, likely without even noticing they're doing so. It can certainly seem that nothing is happening except the summoning of the concept, but I currently believe that, even in the least visual, least audial, least verbal thoughts, there are sensations involved in summoning and manipulating the concepts.

While it may not be possible to actually think "concept" without applying the concept of a concept to some set of sensations, it is possible to apply the concept "concept absent sensations" to some set of sensations! More importantly, I do believe there is a sense in which we have the concepts even when we're not applying them to sensations. Which is to say, you're not reinventing the concept of a rabbit every time you encounter the set of sensations that are best interpreted as "rabbit." It was there, in some form, ready to be applied.

The conceptual content is the construction of the mind, and the way that sensations under-determine their interpretation is what things like the Necker Cube illusion are meant to demonstrate. Most people vastly underestimate the space of possible interpretations of the world around them. Let's try something: pick a nearby object, like a tree or a pillow. Try imagining that it's a rabbit. Don't squint your eyes or imagine that it looks different; just apply the concept "rabbit" to it. Now, try interpreting it as your mom. Now, interpret it as trying to kill you. Now interpret it as secretly dancing. Isn't it odd that you can do this? Yet you can. Normally, we would think that a person who believed a pillow was their mother or a tree was trying to kill them was insane. I believe that most of us are this insane—that we are misinterpreting the world around us to a comparable degree on a constant basis. But we are only likely to notice our misinterpretations when they diverge wildly from the interpretations of those around us, or when we are so surprised or make an error so obvious that we are forced to re-examine our models.

Let's call the application of conceptual content to sensations the *engagement* of that content. To engage the concept of a rabbit—that is, to think "rabbit"—one applies the concept "rabbit" to a set of sensations (perhaps a visual image, perhaps light bouncing off a rabbit and hitting your eyes, perhaps the slight physical memory of something warm and fuzzy, or perhaps to a toothbrush if we're having some trouble that day or happen to be tripping balls).

If one does this consciously, this is *attention*. One does not have to notice that one is applying concepts for the concepts to be applied. You might drive from home to work, deeply absorbed in some line of thought about a presentations you're going to give that afternoon, and arrive at your destination with detailed anticipations about how your presentation will go but no memory of the drive. Nevertheless, you were engaging many complex models outside of your attention—your model of the route, of traffic laws and driving hazards, and of how to operate the car, not to mention your models of how to coordinate your body—not only how to sit and move the gear gear shift and the steering wheel but how to breathe and digest and regulate your body temperature and heartbeat. Which models of how to regulate your body you were engaging probably shifted as you thought about standing up in front of everyone and talking. The models you're engaging outside of your attention are still models, many of them in formats far away from verbal representation, but, I believe, fundamentally no different from your models of the content of your presentation.

Attention is a critical idea because attention is what allows you to intentionally change a model. Concepts are formed, shaped and changed by their application. Learning is the formation and changing of concepts. Thus, if you want to improve a concept or a model, you have to obtain its application conditions—that is, you have to use it. If it is to change, it has to be given the opportunity to take new inputs. And if you want to direct the change, you have to figure out a way to get the model in attention. A model's application conditions are not always easy to obtain; figuring out how to get something into attention is a big part of my practice. We'll come back to this idea in more detail later.

In addition to the conceptual structures themselves, it seems likely that there are constraints on or features of concept formation that aren't properly categorized as concepts—meaning, they guide or are part of the conceptual formation machinery, but are not produced by and cannot be edited by that machinery. Time and space are likely "built in" in this way, rather than being the products of parsed sensations. It's possible that there are other "built in" features of the conceiver—things that would be necessary features of all concepts or features of the structure that concepts are built in. I don't believe that I know what all of these are, but I do believe that one of them—one that is absolutely critical to understand—is Purpose.

Purpose/Goal/Telos/Good

Concepts are formed for a purpose. This is something that I think modern cognitive science understands only locally and shallowly, and that many of the meditation traditions that track the constructed nature of experience miss altogether. Cog sci will take a concept like "snake," note that it's constructed somehow from visual primitives that pick out a wavy line in a visual field that takes straight lines as default, and make a guess that it's important to prioritize the recognition of snakes for evolutionary reasons. While that last bit is something of a leap, this is a functional explanation—i.e., the idea is that the concept is formed for the purpose of survival—and so to my mind it has a leg up on a bunch of the nirvana/moksha-seeking traditions, which will say something closer to "the snake is an illusion; let go of your attachment to the snake." This is bullshit; don't be a California Buddhist. The snake is an illusion—that is, you are constructing your experience of it—but you have conceptual machinery for a reason.

But the reason is not fundamentally survival. If you have to explain all human behavior, survival by itself is pretty unsatisfactory as a motivation. You can sort of do it, at a stretch, and people often try, explaining altruistic behavior as motivated by a species-level survival drive, homosexuality as a glitch or as an obscure form of pro-sociality that somehow contributes to species survival, suicide the same, the purchase of expensive fashionable clothes as an expression of the drive to propagate one's own genes forward in time. You'd have to do something a little fancier, I think, to explain behaviors like those of terrorists who want to wipe out the human species altogether as expressions of a survival drive, but people have definitely tried.

You could take other things besides survival to be the fundamental driver. For many years, I believed that the sole fundamental motivation of mind was to connect to other minds. If you do this, you can do similar antics as one does when trying to cash everything out in survival. I used to be especially focused on explaining violent and manipulative behavior as contorted connection drive—sometimes correctly and sometimes less correctly, as it turned out. That's its own story, but suffice it to say I've updated away from that position.

Whatever you choose as a theoretical base motivation, you're going to have to jump through some hoops explaining the entire space of thought and behavior and preference according to that motivation, because the space is in fact complex and convoluted and full of apparent contradiction. The alternatives to positing a unified fundamental motivation are to posit a fundamentally fragmentary set of drives, or arbitrary motivation. People will often try to explain human motivation as mimetic and/or fully stimulus response, where we can class the action of a fully mimetic mind as a kind of stimulus response, but I hope it will be self-evident why fully environmentally determined motivation and fully imitative motivation are incoherent ideas. We can also discard arbitrary motivation, because clearly there exists some pattern to thought and behavior and preference. As for the option of fundamentally fragmentary drives—for example, the idea that there is a sex drive and a death drive and both exist as incommensurable primitives in the mind, always in competition—all I can say is that I have never found two apparently separate behaviors or preferences to in fact be based in fundamentally separate or incommensurable drives. If you take someone who on the one hand wants to live and on the other hand wants to die, I predict—and I expect to be right—that the "part" of them that wants to live is focused on one set of information, and the "part" of them that wants to die is focused on a a different set of information. Those sets of information are compartmentalized, which results in functionally separate and conflicting impulses, but they can be de-compartmentalized—and once that happens, some more basic motivation will flow through the combined information landscape and output a coherent and non-conflicting set of preferences.

What is the more basic motivation? I don't know. It is the goal—it is the telos, the will, the fundamental drive; it is value; it is Goodness. Good is only ever coherently definable in terms of the desire of an agent; of the Good, all I can say is that there is that which the agent fundamentally desires, and it is in essence a unified whole. And, since we are discussing not a model of an individual mind but a model of the invariant features of mind in general—a model that is meant to stand independent of individuation or embodiment—the claim is not merely that the individual human has a fundamentally unified or unifiable will, but that at the most fundamental layer, the Goal or the Good is invariant across minds. This is an extremely strong claim. It is not, however, a particularly unusual one—many spiritual traditions make similar claims, usually articulated more fuzzily, along the lines of "we are all one." But "we are all one" is an extremely strong claim if taken seriously, and can lead people to do really, really stupid and dangerous things. An astounding amount of complexity and divergence can arise from a single algorithm that processes varied inputs, and we will explore that complexity and divergence in depth as we go. For now, just know that the claim of the universality of telos is not an invitation to treat—or to value—everyone the same.

So, we have a model of mind—awareness containing sensations, sensations interpreted by concepts, concepts composed within a structured framework and in service of a fundamental Goal. I believe that the ability to make and think through the distinctions laid out above is extremely important, for reasons that will hopefully become clear later in this document. To enable ourselves to make those distinctions, we've defined the elements of our ontology, and we've clarified our understanding of the elements by imagining them separately from each other —awareness without sensations or concepts, sensations without concepts, and concepts without sensations. In reality, I don't know which if any such separations ever occur. The idea of sensations or concepts as mental phenomena without awareness is incoherent (in the way that the idea of a thought without a mind is incoherent), so we can rule that one out. Some meditators speak of achieving states of "pure awareness," where they are not interpreting their sensations and/or not experiencing sensations, and have no will or desire but "just are." I believe it's possible to turn one's attention away from sensations, and I believe it's possible to dissolve and/or inhibit the application of many, many of one's conceptual constructs. I don't believe, however, that it's actually possible to turn off the will, and I also believe that while all of what people are typically cognizant of as their conceptual content is probably dissolvable, the generator of conceptual content is not. So my actual model is something more like, there exists the "fabric of mind," which is the awareness-will, which irrevocably contains the parser/conceptualizer/creative apparatus, and through which sensations pass and are responded to by that apparatus in accordance with the will.

* * * * *

Before we continue any further, I want to note something really important: if the model I've stated is correct, then your sole purpose in this world is to navigate your way toward what you want—sole purpose not in a moral sense, but in a determinist one, in that if the model above is correct, any idea of morality on which you might have that mission or any other was generated by the function structuring your model of the world in order to pursue your Goal, and thus is superseded by that function.

It is not possible to do other than pursue what you want, but there is a truth about what you actually want, and a truth about how to get it, and it's quite possible to be ignorant or wrong on both counts.

We'll come back to that. Onward.

2. Mind in Environment

Think of the model articulated above as a kind of base. These are claims about invariant features of mind; they tell us nothing about any specific aspect of individual or cultural psychology. You won't get from this model to Oedipal complexes, bipolar disorders, anxiety or depression, childhood trauma, or any of the other structures people typically associate with psychology.

Those things are all the result of the mind interacting with an environment. The model of mind outlined so far is so minimal that the description does not even assume it is implemented in a brain; however, if we put that mind into the brain and body of a cat, or a human girl, or an 18th century French aristocrat, it now has an environment—not merely the environment that the body of the cat or the girl or the French aristocrat inhabits, but the body itself. [1]

The body is in some sense the most fundamental part of the mind's environment, in that it generates—or at least populates—the sensorium. The eyes take in visual sensations, the ears take in auditory sensations, the skin takes in touch sensations, etc. In addition to whatever information is being taken in by the perceptive organs, there is a ton of sensory information corresponding to internal states of the body—the feelings of blood flow, digestion, hormone release, muscular activation and coordination, immune response, etc.

Imagine the mind we described—the awareness-will with its creative conceptual apparatus—now flooded with the experience of all of these sensations, and ready to conceptualize them. How is it to do this?

Part of the answer is "pattern recognition." When people talk about pattern recognition, they mean that something is recurring in nature, and it is possible to notice the recurrence. If we are imagining a newborn infant mind conceiving its environment (which, remember, includes its body) we might pick out the combination of the mother's face plus some warm snuggly feelings as a stably recurring phenomenon.

Pattern recognition can't be the whole story, because pattern recognition *ex nihilo* is impossible. Computer programs designed to recognize patterns in large data sets must start with some initial set of concepts, some initial dimension or set of dimensions along which it is possible to organize data. If nothing else, it must start with the concept of pattern, some idea of "same" or "match" and some concept like "dimension" (way of matching). Our pattern recognition machinery—the conceiver described in the first section—might have enough "built in" constraints or features of the type we discussed in the first section that conceptual possibility space is fairly determined: put two minds in the world and they will roughly conceptualize it into categories like air, water, trees, animals, parents, etc., absent any other constraints and regardless of whether they can communicate with each other.

However, with a conceptual apparatus as minimal as the one I posited earlier—time, space, and Purpose, plus this idea of similarity or recognition—that kind of over-determination seems pretty unlikely. The "Purpose" idea could load a lot of constraints into the parser, but I don't believe it does. I believe the fundamental drive is so minimally (or broadly) defined that it

[1] For now, in order to play with the idea of the mind described above interacting with the body, we are going to speak dualistically. Mind-body dualism is the belief that the mind and the body are separate phenomena, with mental phenomena being non-physical and the body and the material world being a separate sphere. I don't actually believe mind-body dualism, but for the moment it will be cleaner to speak of the mind and body as though they were separate, in order to posit an order of conceptual operations that would be too hard to think about if we started out behaving as though there were no distinction.

doesn’t even have an idea as complex as biological survival. Any function or goal of the body that pertains specifically to the body must on this model be construed as a feature of the mind’s environment. Good as experienced and interpreted through the body is an order more complex than the Good that inheres “fundamentally” in the mind.

So it is not just the way that the body generates the sensorium that creates the mind’s environment, but also the body’s functions—things like digestion, reproduction, sleep. These provide constraints on the way that the Good can be pursued, and thus constraints on conceptual formation.

The interface between the body’s functions and the mind’s experience is the phenomenology. You experience things pleasure and pain, satisfaction and dissatisfaction, attached to physical experiences like hunger, tiredness, and arousal as the body goes through different states. The mind as described in section 1, minimal parser guided by goal, would not form the concept “food” absent a body; the mind parsing the landscape of sense data that includes simultaneously the smells and tastes and textures of edible substances and feelings of hunger and digestion, co-occurring in a regular way with the experiences of pleasure and pain—that is, tagged as salient—is much more constrained and will reliably construct “food” concepts.

You might have noticed a messy implication of this ghost-in-the-shell kind of model: the idea is that the mind has its own goal, and the body has its own goals or at least functions, and these hook up somehow, such that the embodied mind knows that the body’s functions are relevant to or necessary for its goal. Where and how does this hook-up occur? I have no idea as to the details, but the analogy of a car seems apt to me: a car is constructed to be relatively intuitive to the driver that occupies it. Some learning is necessary, different cars are all slightly different, and if a driver has the skill he to do so can customize the car’s internal workings (you’ve probably heard of people training themselves to accomplish strange and usually impossible physical feats—Wim Hof and David Blaine come to mind). However, to avoid responding with the unsatisfying shrug that generally accompanies the mind-body problem, we would have to leave our dualist working simplification, and that is a different topic. So we’re just going to have to bear with our rather implausible ghost in the shell for now.

Let’s go back to our infant learning to conceive of his mother and the warm snuggly feelings that accompany her presence. We’ll call this concept *Mom + Warm Snuggliness.* This is actually a terrible articulation of the idea I’m trying to convey, because linguistically it preserves mom and warm and snuggliness as separate ideas. This is not what’s happening conceptually for our infant. Instead, try to imagine that the infant is forming a concept that tags the entire swath of sensations that accompany his mother’s presence: the visual sensations that correspond to the mother’s face, the temperature sensations that correspond to being held, the proprioceptive and textural and affective sensations that correspond to the release of oxytocin (contentedness, sleepiness), the olfactory sensations corresponding to her body and physical state and the baby's own physical state, and more. This concept isn't nearly as clear as the abstraction indicated by the word "mother"—it's a single conceptual mashup of a person (though perhaps without the concept of person), the feeling of love and connection, a set of physical states and feelings and actions, etc.

I believe this is how early concept formation works. Further distinctions get drawn later, as the different elements of the experience occur and are noticed separately from each other. However, the basic concepts that underpin our more intricate and abstract conceptual structures are deep conflations of the external and the internal. There is no particular reason to start out with the idea of "external" and "internal;" one simply starts with the sensory field, and must parse the sensations in whatever groupings they seem to occur. By the time we are using concepts with verbal correspondents like "mother," we are working many layers of abstraction and distinction beyond our earliest conceptual foundations. Not only is the normal idea of "mother" a far more specific and distinct occurrence in the sensorium than entirety of the set of sensations that correspond to her presence, the set of sensations that are in fact specific to "mother" are not in the sensorium at all times—that is, her continued existence when she is not sensorily present is an inference. "Mother" as a distinct entity is therefore an abstraction from experience—far more abstract than the earlier concept I tagged *Mom + Warm Snuggliness*.

If this is how concepts are formed, then our most basic and visceral understandings of the world are very different from our verbally communicated concepts, containing what we would consider many strange conflations and disjunctions. If, for example, the base concept that contains (some of) the feelings that later get abstracted into the concept that gets verbally tagged "love" is a concept like *Mom + Warm Snuggliness*, then really, some semblance of an Oedipal complex might be nigh unavoidable—not because of perversion but merely because of the conditions of the concept's construction.

On the other hand, not all concepts are built in direct reference to clusters of sensory phenomena; some concepts are constructed in reference to something that other people appear to be referring to. When a toddler is told by his mother to be polite, "polite" doesn't refer to a concrete set of sensations the way that *Mom + Warm Snuggliness* does; "polite" refers to something mysterious that his mother seems to want. Politeness may well come to be associated with a physical state—whatever physical state garners or coincides with positive feedback from his mother—but the initial referent of the concept was not a naturally arising correspondence between the physical state and/or affect and the world, but rather the apparent desire of another entity. That is, the concept formation was not directly guided by the evaluative apparatus of the child—the one guided but his own Goal—but was indirectly guided by the evaluative apparatus of someone else. This would not occur if the child's evaluative apparatus did not value the other person's evaluation—or, perhaps more likely, their happiness or approval—but nevertheless it is a very different order of operations.

Thus, as far as we are concerned with how the mind structures itself in response to its environment, we must consider other people and their evaluative frameworks to be critical features of the environment, potentially vastly more weighty than the physical features of the environment in terms of the mind's architecture. What is likely to be more salient to a child's evaluative apparatus—the color and texture of the carpet, or his mother's pleasure and displeasure? For most children, certainly the latter. The parents' attention directs the child's attention, and later the child's friends and teachers direct the child's attention—which means that whatever is directing their attention is directing his attention.

Since we are talking about the ways that psychological structure is determined by the body and the ways it is determined by the social environment—essentially questions of nature or nurture—we should briefly touch on the question of genetic determination. Let's just explore a couple different ideas of how this might work. You could imagine several different mechanisms, and the fact that a whole host of personal tendencies is extremely likely to be "inherited" from the parents via the attentional direction just described above makes it difficult to tell whether, when we observe what appear to be familial features of a person's psychology, we're working with information directly encoded in the body or information transmitted at high fidelity in early life (where early life includes the prenatal environment). On this model, remember, information encoded in the body is still classed as a feature of the mind's environment, so we're essentially asking what are the different parts of a mind's environment that information can come from. If, for example, certain conceptual structures are encouraged or inhibited by chemical tendencies of the body (e.g., testosterone or serotonin production), then what we have is a mind parsing an environment in which phenomenological warmth or relaxation or excitement or whatever else is produced by the chemical tendency co-occurs more or less frequently than average with externally generated sense data, and must be simultaneously explained. That is, mind's experience of the external world is always being correlated with its experience of the body, and the experience of the body can be genetically influenced, though obviously not fully genetically determined.

I have worked with people who appear to access genetic or ancestral memory in the course of introspection—I have seen, for example, one woman appear to recover a memory of what she believed was her grandmother's village getting bombed in a war that happened before she was born, and I myself have found what appear to be highly specific and fairly ancient cultural models of deportment influencing my behavior. I believe it's possible to genetically encode fairly complex situationally contingent "plans"—things like, for example, ketosis—and one could imagine an aggressive ketosis contingency encoded into genetic plans as a result of an ancestor experiencing famine. However, I think that for a modern person's ketosis contingency to be psychologically associated with the experiential details of the potato famine of 1845, there have to be some non-physiologically encoded attentional things going on. In any case, I have never encountered particularly strong evidence of this kind of thing drastically shaping someone's psychological structure independent of familial enculturation.

One final complicating factor in the nature versus nurture question is that, in addition to the kind of attentional direction we saw in the politeness example, I believe children also learn physiological regulation from their parents, especially their mothers, especially pre- and post-natally. So a physiological tendency that strongly influences the child's interpretive landscape might be genetically encoded, but might also be learned from physical contact with the parent.

For many psychological "family traits," I think that a wide variety of positions in this nature-and-nurture space are equally plausible causal candidates. I have no ideological attachment to one over the other. Many people who prefer nature-type explanations do so because they are concerned about responsibility being unfairly allocated to those who can't help their genetic inheritance, or because they want to prove that some group is terminally better or worse than some other group. Conversely, many people who prefer nurture-type explanations prefer them because they want to believe in the power of good parenting or education or therapy to equalize or uplift everyone. To me, nurture does not indicate ease of change—trauma can be insidious in its structuring of the psychology, and those who cling to their fears may as well have been born with them written into their bones. On the other hand, I have no real reason to believe that genetic inheritance is indelibly fixed. Encoding is encoding, even at the physiological level, and if it can be accessed, it can be informed. To me, tractability is less a question of nature versus nurture, and more a question of how early a structure was built or implemented (since this determines how much has subsequently been built on top of it, dependent upon it or patterned on it), how much a person wants to and believes they can change, and how much of the environment around them can be brought to incentivize change by making it evidently the "fit" thing to do (i.e., conducive to the achievement of the Goal in that environment).

Having addressed to the extent we can for now the question of nature versus nurture and, broadly speaking, collapsed the two ideas into the broader idea of "environment," let's go back again to that idea. Our idea of the mind's "environment" is unusually broad, including the body, the surrounding physical environment, and the other people it interacts with—everything that produces sense data. Yet we've still neglected a critical feature of the mind's environment: its own conceptual structure. The parsing of the environment affects the environment via the actions it prescribes for the person, and thereby also affects what the mind will encounter in its environment. As the mind builds up conceptual structure, the application of that conceptual structure causes it to recognize things in the environment—i.e., see what it has seen before. And, as the mind pursues its Goal via the conceptual structure it builds, seek what it has sought before. Thus the conceptualization of the sense data becomes self-reinforcing. This is why the early conceptual structure has an outsize effect on the state of the mind: it is the "parent" of later conceptual structure. As the mind conceptualizes its sense data, it creates its own environment, making it familiar and navigable, but also frequently entrapping itself in the process, as it builds more and more structure that is templated upon early error.

We now have the beginning of a basis for a model of structured psychology—the specific sets of sense data provided by the mind's environment that it then conceptualizes,

3. Agents and the Development of Internal Conflict

Earlier, I made a big deal of asserting that the mind's fundamental drive is unitary. But it seems obvious that within individual people there are different, often conflicting drives. How would it come to be that in a mind with a unified goal, we would experience both the desire to stay in a relationship and to leave it, or the desire to diet and the desire to eat, or the desire to make money and the desire to stop working?

The answer lies in the distinction between the Goal and the subgoals that the mind develops in the process of conceptualizing the environment. Internal conflict arises from the fact that distinct subgoals do not necessarily know about each other or know how to share information. They may not know that they are ultimately working toward the same end, or agree about how to get there. In the same way that the reasons behind a foreign cultural practice can be opaque to an outsider, so can one function of the mind be opaque to another function. In some cases, this opacity can even be intentional, either because one part of the mind does not want to understand another, or because it does not want to be understood itself. In other cases, the problem is merely one of failure to communicate.

To get a better understanding of this idea of subgoals, let's go back to our baby with its *Mom + Warm Snuggliness* concept. Actually, this is not just a concept—it's a subgoal: something that the baby has conceptualized because it has recognized in the experience a pattern relevant to its Goal. The baby now has a conceptual structure directing its will that not only recognizes *Mom + Warm Snuggliness* but seeks or anticipates it, and responds with action or impulse when it encounters it.

In order to be able to seek and anticipate *Mom + Warm Snuggliness,* it must be that the conceptual structure (model) pertaining to the subgoal is more complex than just the subgoal itself. It must contain not just the concept of the subgoal but also conceptual structure representing the ways that other things relate to it, e.g. the conditions under which it is more or less likely to occur (perhaps time of day), or the conditions under which it is more or less important to attain (perhaps hunger versus satiation, or wakefulness versus sleepiness).

In this particular example, the action or impulse that occurs in response to the anticipation of *Mom + Warm Snuggliness* is fairly subtle; perhaps the baby relaxes. But you could imagine a different structure where the person had an extensive model of playground status dynamics, which contained among other things a subgoal like "pursue justice" and concept like "bully," such that upon recognizing a bully (i.e., applying the concept to a situation where it seemed to fit), the person would punch the bully in the face.

I think of this entire structure as a *function*. The function is the will being channeled through a dynamic conceptual structure (a model) toward a conceptualized aspect of the Goal (a subgoal). The function takes inputs in the form of sense data, parses them in accordance with the conceptual structure, and produces outputs in form of impulse, action, emotion, and/or thought. Outputs may also include amendments to the conceptual structure. I believe that individual functions of the mind come into being where Goal-relevance is perceived in the environment, and conceptual structure is built to pursue that aspect of, or apparently necessary step toward, the Goal. That aspect or step is the subgoal; the function is that which pursues the subgoal. (From here out, I will drop the term subgoal—a function will have a goal and the mind will have its Goal.)

Frequently, it makes sense to refer to such functions as agents, or parts of the person. To some of the people reading this, the term "part" will sound familiar from their work with Internal Family Systems, or IFS, which is the only therapy style I've encountered that explicitly ontologizes the mind as a multi-agent system, although many others do so implicitly/occasionally/to varying degrees. (If you're interested in the IFS ontology, you can read more here). To me, the term "part" usually represents the personification of a function; I like to use it when I'm interacting directly with a "part" of someone or otherwise describing it as though it is a person. In other cases, I like to use the term "function" because it emphasizes the idea that the individual function is part of a more general system, rather than being a fundamentally separate entity. And I like to use the term "agent" when I want to capture the idea of intentionality in a way that the term "function" doesn't, but want to remain impersonal-sounding so as to seem abstract and cool.

Anyway. Now that we have a clear idea of agents/functions/parts, we can look at the conditions under which they come into conflict. Let's start with a "simple" case, by which I mean a case in which there is no intentional obfuscation, merely a failure of communication.

Earlier, I made the analogy between a conceptual model and a book. This is actually not a great analogy for several reasons. For one thing, a book is static, while a model evolves as it is applied, and even begets "offspring" as new situations are encountered that can be parsed with elements of existing models. For another thing, a book is meant to be read. A model is meant to be applied, but is not necessarily formatted to be understood or communicated. Thus, a person can end up in a position where they seem not to be able to "access" information they know they have. Imagine someone trying to eat healthy. She "knows" she should eat vegetables, but she craves ice cream instead. So we have two functions at work here—one that wants vegetables and one that wants ice cream. Let's think about the differences between these functions' models.

In some sense, the two models contain very similar information—at least, they purport to model the same domain. The function that wants the ice cream probably has a model of the caloric and chemical content of both ice cream and vegetables, formatted as an understanding of how they taste and how it will feel to eat them—their felt effects on the person's energy levels and digestion, and the emotional states associated with those. The function that prescribes vegetables probably also has a model of the the caloric and chemical content of both ice cream and vegetables, formatted more abstractly, perhaps recalled as visual/audial imaginations—perhaps a memory of a conversation the person had with her mom or a health article she once read. (The latter function's model may not be very precise—it might be as simple as "ice cream is fatty and sugary and vegetables have nutrients and vitamins." But even in the case where the person is counting calories and tracking specific nutrients, it's worth noting that the abstract model of the caloric and chemical contents of food is almost definitionally nowhere near as detailed as the model generating the craving.)

So, two models of the caloric and chemical contents of food. To expect to be able to use the ice cream-craving function's model to write down the chemical composition or number of calories of the ice cream, however, is obviously ridiculous (at least without training). In that sense, the information in the model of ice cream that generates the craving is not at all the same information as the information in the explicit model (the one that wants the person to eat vegetables). Both models contain descriptions of ice cream, but the formatting of the information is so different that the two can only be said to be the same information in a very abstract sense. Simply put, they're not.

However, the two functions are attempting to model the world and produce action in the same domain, and thus are coming into conflict. The vegetable-prescribing function can't "talk to" the ice cream-seeking function, because when it says "that ice cream doesn't have any nutrients," the ice-cream seeking function not only has no idea what it's talking about—it may have no idea that it's talking at all. So the person tells herself, "I should eat vegetables," and this affects the status of her ice-cream craving not one iota.

You could imagine her integrating the two models, building a translation between them by creating a highly detailed mapping between her sensory experience of food and her explicit understanding of nutrition. (You could even imagine her getting good enough to be able to approximately guess a food's caloric and nutritional content from taste and digestive feel.) In the world where both parts shared similar goals or recognized the validity of each other's goals, this would resolve the conflict—the person might crave ice cream, then be able to propose the idea of vegetables to herself via visceral imagination, at which point the craving might shift. Or perhaps she might get a feeling in response to the visceral imagination of eating vegetables that she could understand as "no, I'm not really hungry, I just want to feel comfort right now." At that point she might decide to that comfort was a good idea, and eat the ice cream without feeling guilty about it. Or, she might be able to counter-offer with something healthier but equally comforting. If the relationship between the two parts is good enough that both recognize the Goal-relevance of each other's goals, and her counter-offer is good, the counter-offer will not only be acceptable but may present as viscerally superior to the part that was previously craving ice cream, since it will value attaining both comfort and health.

Such successful communication between parts, however, depends on their willingness to understand or attend to each other's information, and/or value the achievement of their goals. You could easily imagine a person with a pair of functions similar to the ones we just described, except where the vegetable-prescribing function strictly cared only about following an explicit set of nutritional rules and didn't care about the other function's nutritional models or desire for comfort. This version of the vegetable-prescribing function instead would actively ignore and do its best to override all of the other function's signals. Obviously, in this scenario, the communication channel we imagined above would never get built.

Imagine yet another variation: the vegetable-prescribing function is willing to make some allowances for ice cream cravings, but only where they're properly justified. Imagine that it only gives the green light to eat ice cream if it's been a really, truly hard day. Now imagine that the ice cream-craving function is experiencing a certain type of hunger and wants the sugar and fat that the ice cream contains, but has learned that it's only going to get them if it can make the case that things are going sufficiently poorly. Now, when the person is debating with herself about whether to eat the ice cream, she finds herself imagining all the worst aspects of the day, dwelling on and amplifying the difficult parts, spending a lot of attention catastrophizing. This is happening because the available rationale for ice cream is determined by the constraints she has set on what counts as acceptable; the ice cream-craving function has an incentive to obfuscate its models, instead fabricating information that will pass a filter. Thus the possibility of accurate introspection is compromised.

This kind of thing happens all the time. People who believe they shouldn't ask for care if they're healthy and able to care for themselves will often find themselves feeling mysteriously sick. People who believe they can only be loved if they're sufficiently moral will fabricate altruistic justifications for their behavior that obscure their real reasons. A lot of internal fragmentation—that is, the state of different parts of the mind being unable to share information, or more simply put, the inability of a person to know themselves—is created and stabilized by parts of the mind inadvertently setting up adversarial incentives for others, where both are vying to produce action in a shared domain.

Setting up good communications between functions, however, can be easier said than done. A function with a directly adversarial relationship to another function is often configured that way because it came into being for the express purpose of fighting or compensating for the first function. Take our ice cream-craving function and our vegetable-prescribing function. It's possible that the vegetable-prescribing function developed when the person read a book on nutrition and decided it would be a good idea to try to eat more healthily. I should mention, by the way, that calling these guys the "ice cream-craving function" and the "vegetable-prescribing function" is a silly simplification, and I'm not at all implying that ice cream and vegetables are their only or their primary goals. In the scenario we're describing now, where the vegetable-prescribing function arises from an encounter with a nutrition book, it's much more likely that the person already had an elaborate self-management and information-gathering function, and the plan to encourage herself to eat vegetables was a simple addendum developed by a function that knew how to seek and integrate a wide range of information into its plans. It's also possible, however, that the vegetable-seeking function arose specifically in response to some terrible failure of the ice cream-seeking function. If the person's cravings were pretty out of whack and generating some externalities that it didn't seem to be able to take into account, and she found herself eating until she felt sick or gaining weight she didn't want to gain, the second function might have come into being specifically to rein in the first.

If this is the case, we can say that the ice cream-craving function is *prior to* the vegetable-prescribing function. The vegetable-prescribing function has the goal of being healthy, but is largely concerned with correcting the ice cream-craving function. If we see this, we know something about the responsive evolution of the person's psychological structure. Psychological structure develops over time in response to perceived error or incompleteness in existing structure, when the mind observes its actions and forecasts its own failure to achieve its Goal. In my experience, common functional growth over time looks much like the branches of a tree—tapering from extremes into nuance as the person's different functions get increasingly refined in their mutual compensation patterns. In some cases, people with deeper than average understandings of their own inner workings are able to more frequently understand and amend the original function generating an error rather than develop a separate compensating function, and so experience more integration and less "branching." Dysfunctional development, on the other hand, is more "bipolar," with the different parts of the mind each refusing to accept the realities of the other and instead implementing more and more extreme compensations and repressions on both sides, until functional sustained action becomes virtually impossible.

At this point, we have a working model of psychological structure, sufficiently detailed to allow a huge amount of navigation, correction and development. We have a base ontology of mind that lets us understand the distinction between reality and interpretation, and maintain an understanding of the commensurability of goals; we have the understanding of the interpretive apparatus building conceptual structure in response to its environment; and we have the understanding of that same interpretive apparatus building further conceptual structure in response to its projections of its own behavior in its environment. This lets us understand the general pattern of psychological genesis, and, with good introspective training, also lets us trace the genesis of a specific psychological structure, which lets us understand its subjective environment and primary concerns, its checks and balances, and thus positions us to be able to communicate with it in a targeted manner and help it develop positively.

Please use these models gently—don't over-impose them where they don't actually seem to fit, and don't expect to be able to immediately solve all your problems with them. Building relationships takes time and care, and applying these models well means building strong relationships, with anyone you are trying to understand but especially with yourself.

Bon Voyage!