
NE
E
SE
S
SW
W
NW
N

Multiagent Models of Mind

1. Sequence introduction: non-agent and multiagent models of mind
2. Book Summary: Consciousness and the Brain
3. Building up to an Internal Family Systems model
4. Subagents, introspective awareness, and blending
5. Subagents, akrasia, and coherence in humans
6. Integrating disagreeing subagents
7. Subagents, neural Turing machines, thought selection, and blindspots
8. Subagents, trauma and rationality
9. System 2 as working-memory augmented System 1 reasoning
10. Book summary: Unlocking the Emotional Brain
11. A mechanistic model of meditation
12. A non-mystical explanation of insight meditation and the three characteristics of existence: introduction and preamble
13. A non-mystical explanation of "no-self" (three characteristics series)
14. Craving, suffering, and predictive processing (three characteristics series)
15. From self to craving (three characteristics series)
16. On the construction of the self
17. Three characteristics: impermanence
18. Beliefs as emotional strategies
19. My current take on Internal Family Systems “parts”

Sequence introduction: non-agent and multiagent models of mind

A typical paradigm by which people tend to think of themselves and others is as *consequentialist agents*: entities who can be usefully modeled as having beliefs and goals, who are then acting according to their beliefs to achieve their goals.

This is often a useful model, but it doesn't quite capture reality. It's a bit of a fake framework. Or in computer science terms, you might call it a leaky abstraction.

An abstraction in the computer science sense is a simplification which tries to hide the underlying details of a thing, letting you think in terms of the simplification rather than the details. To the extent that the abstraction actually succeeds in hiding the details, this makes things a lot simpler. But sometimes the abstraction inevitably leaks, as the simplification fails to predict some of the actual behavior that emerges from the details; in that situation you need to actually know the underlying details, and be able to think in terms of them.

Agent-ness being a leaky abstraction is not exactly a novel concept for Less Wrong; it has been touched upon several times, such as in Scott Alexander's Blue-Minimizing Robot Sequence. At the same time, I do not think that it has been quite fully internalized yet, and that many foundational posts on LW go wrong due to being premised on the assumption of humans being agents. In fact, I would go as far as to claim that this is the biggest flaw of the original Sequences: they were attempting to explain many failures of rationality as being due to cognitive biases, when in retrospect it looks like understanding cognitive biases doesn't actually make you substantially more effective. But if you are implicitly modeling humans as goal-directed agents, then cognitive biases is the most natural place for irrationality to emerge from, so it makes sense to focus the most on there.

Just knowing that an abstraction leaks isn't enough to improve your thinking, however. To do better, you need to know about the actual underlying details to get a better model. In this sequence, I will aim to elaborate on various tools for thinking about minds which look at humans in more granular detail than the classical agent model does. Hopefully, this will help us better get past the old paradigm.

My model of what I think our subagents looks like draws upon a number of different sources, including neuroscience, psychotherapy and meditation, so in the process of sketching out my model I will be covering a number of them in turn. To give you a rough idea of what I'm trying to do, here's a summary of some upcoming content.

Published posts:

(*Note: this list may not always be fully up to date; see the sequence index for actively maintained version)*

Book summary: Consciousness and the Brain. One of the fundamental building blocks of much of consciousness research, is that of Global Workspace Theory (GWT). This could be described as a component of a multiagent model, focusing on the way in which different agents exchange information between one another. One elaboration of

GWT, which focuses on how it might be implemented in the brain, is the Global Neuronal Workspace (GNW) model in neuroscience. Consciousness in the Brain is a 2014 book that summarizes some of the research and basic ideas behind GNW, so summarizing the main content of that book looks like a good place to start our discussion and for getting a neuroscientific grounding before we get more speculative.

Building up to an IFS model. One theoretical approach for modeling humans as being composed of interacting parts is that of Internal Family Systems. In my experience and that of several other people in the rationalist community, it's very effective for this purpose. However, having its origins in therapy, its theoretical model may seem rather unscientific and woo-y. This personally put me off the theory for a long time, as I thought that it sounded fake, and gave me a strong sense of "my mind isn't split into parts like that".

In this post, I construct a mechanistic sketch of how a mind might work, drawing on the kinds of mechanisms that have already been demonstrated in contemporary machine learning, and then end up with a model that pretty closely resembles the IFS one.

Subagents, introspective awareness, and blending. In this post, I extend the model of mind that I've been building up in previous posts to explain some things about change blindness, not knowing whether you are conscious, forgetting most of your thoughts, and mistaking your thoughts and emotions as objective facts, while also connecting it with the theory in the meditation book The Mind Illuminated.

Subagents, akrasia, and coherence in humans. We can roughly describe coherence as the property that, if you become aware that there exists a more optimal strategy for achieving your goals than the one that you are currently executing, then you will switch to that better strategy. For a subagent theory of mind, we would like to have some explanation of when exactly the subagents manage to be collectively coherent (that is, change their behavior to some better one), and what are the situations in which they fail to do so.

My conclusion is that we are capable of changing our behaviors on occasions when the mind-system as a whole puts sufficiently high probability on the new behavior being better, when the new behavior is not being blocked by a particular highly weighted subagent (such as an IFS-style protector) that puts high probability on it being bad, and when we have enough slack in our lives for any new behaviors to be evaluated in the first place. Akrasia is subagent disagreement about what to do.

Integrating disagreeing subagents. In the previous post, I suggested that akrasia involves subagent disagreement - or in other words, different parts of the brain having differing ideas on what the best course of action is. The existence of such conflicts raises the question, how does one resolve them?

In this post I discuss various techniques which could be interpreted as ways of resolving subagents disagreements, as well as some of the reasons for why this doesn't always happen.

Subagents, neural Turing machines, thought selection, and blindspots. In my summary of *Consciousness and the Brain*, I briefly mentioned that one of the functions of consciousness is to carry out artificial serial operations; or in other words, implement a production system (equivalent to a Turing machine) in the brain.

While I did not go into very much detail about this model in the post, I've used it in

later articles. For instance, in *Building up to an Internal Family Systems model*, I used a toy model where different subagents cast votes to modify the contents of consciousness. One may conceptualize this as equivalent to the production system model, where different subagents implement different production rules which compete to modify the contents of consciousness.

In this post, I flesh out the model a bit more, as well as applying it to a few other examples, such as emotion suppression, internal conflict, and blind spots.

Subagents, trauma, and rationality. This post interprets the appearance of subagents as emerging from *unintegrated memory networks*, and argues that the presence of these is a matter of degree. There's a continuous progression of fragmented (dissociated) memory networks giving arise to increasingly worse symptoms as the degree of fragmentation grows. The continuum goes from everyday procrastination and akrasia on the "normal" end, to disrupted and dysfunctional beliefs on the middle, and conditions like clinical PTSD, borderline personality disorder, and dissociative identity disorder on the severely traumatized end.

I also argue that emotional work and exploring one's past traumas in order to heal them, is necessary for effective instrumental and epistemic rationality.

Against "System 1" and "System 2". The terms System 1 and System 2 were originally coined by the psychologist Keith Stanovich and then popularized by Daniel Kahneman in his book *Thinking, Fast and Slow.* Stanovich noted that a number of fields within psychology had been developing various kinds of theories distinguishing between fast/intuitive on the one hand and slow/deliberative thinking on the other. Often these fields were not aware of each other. The S1/S2 model was offered as a general version of these specific theories, highlighting features of the two modes of thought that tended to appear in all the theories.

Since then, academics have continued to discuss the models. Among other developments, *Stanovich and other authors have discontinued the use of the System 1/System 2 terminology as misleading*, choosing to instead talk about Type 1 and Type 2 processing. In this post, I will build on some of that discussion to argue that Type 2 processing is a *particular way of chaining together the outputs of various subagents using working memory*. Some of the processes involved in this chaining are themselves implemented by particular kinds of subagents.

Book summary: Unlocking the Emotional Brain. Written by the psychotherapists Bruce Ecker, Robin Ticic and Laurel Hulley, *Unlocking the Emotional Brain* claims to offer a neuroscience-grounded, comprehensive model of how effective therapy works. In so doing, it also happens to formulate its theory in terms of belief updating, helping explain how the brain models the world and what kinds of techniques allow us to actually change our minds. Its discussion and models are closely connected to the models about internal conflict and belief revision that are discussed in previous posts, particularly "integrating disagreeing subagents".

A mechanistic model of meditation. Meditation has been claimed to have all kinds of transformative effects on the psyche, such as improving concentration ability, healing trauma, cleaning up delusions, allowing one to track their subconscious strategies, and making one's nervous system more efficient. However, an explanation for why and how exactly this would happen has typically been lacking. This makes people reasonably skeptical of such claims.

In this post, I want to offer an explanation for one kind of a mechanism: meditation increasing the degree of a person's introspective awareness, and thus leading to increasing psychological unity as internal conflicts are detected and resolved.

A non-mystical explanation of insight meditation and the three characteristics of existence: introduction and preamble. Insight meditation, enlightenment, what's that all about?

The sequence of posts starting from this one is my personal attempt at answering that question. It seeks to:

- Explain what kinds of implicit assumptions build up our default understanding of reality and how those assumptions are subtly flawed.
- Point out aspects from our experience whose repeated observation will update those assumptions, and explain how this may cause psychological change in someone who meditates.
- Explain how the so-called "three characteristics of existence" of Buddhism - impermanence, no-self and unsatisfactoriness - are all interrelated and connected with each other in a way that is connected to the previously discussed topics in the sequence.

Farther out (sketched out but not as extensively planned/written yet)

The game theory of rationality and cooperation in a multiagent world. Multi-agent models have a natural connection to Elephant in the Brain -style dynamics: our brains doing things for purposes of which we are unaware. Furthermore, there can be strong incentives to continue systematic self-deception and *not* integrate conflicting beliefs. For instance, if a mind has subagents which think that specific beliefs are dangerous to hold or express, then they will work to suppress subagents holding that belief from coming into conscious awareness.

"Dangerous beliefs" might be ones that touch upon political topics, but they might also be ones of a more personal nature. For instance, someone may have an identity as being "good at X", and then want to rationalize away any contradictory evidence - including evidence suggesting that they were wrong on a topic related to X. Or it might be something even more subtle.

These are a few examples of how rationality work has to happen on two levels at once: to debug some beliefs (individual level), people need to be in a community where holding various kinds of beliefs is *actually safe* (social level). But in order for the community to be safe for holding those beliefs (social level), people within the community also need to work on themselves so as to deal with their own subagents that would cause them to attack people with the "wrong" beliefs (individual level). This kind of work also seems to be necessary for fixing "politics being the mind-killer" and collaborating on issues such as existential risk across sharp value differences; but the need to carry out the work on many levels at once makes it challenging, especially since the current environment incentivizes many (sub)agents to sabotage any attempt at this.

(This topic area is also related to that stuff Valentine has been saying about Omega.)

This sequence is part of research done for, and supported by, the Foundational Research Institute.

Book Summary: Consciousness and the Brain

One of the fundamental building blocks of much of consciousness research, is that of Global Workspace Theory (GWT). One elaboration of GWT, which focuses on how it might be implemented in the brain, is the Global Neuronal Workspace (GNW) model in neuroscience. Consciousness and the Brain is a 2014 book that summarizes some of the research and basic ideas behind GNW. It was written by Stanislas Dehaene, a French cognitive neuroscientist with a long background in both consciousness research and other related topics.

The book and its replicability

Given that this is a book on psychology and neuroscience that was written before the replication crisis, an obligatory question before we get to the meat of it is: how reliable are any of the claims in this book? After all, if we think that this is based on research which is probably not going to replicate, then we shouldn't even bother reading the book.

I think that the book's conclusions are at least reasonably reliable in their broad strokes, if not necessarily all the particular details. That is, some of the details in the cited experiments may be off, but I expect most of them to at least be pointing in the right direction. Here are my reasons:

First, scientists in a field usually have an informal hunch of how reliable the different results are. Even before the replication crisis hit, I had heard private comments from friends working in social psychology, who were saying that everything in the field was built on shaky foundations and how they didn't trust even their own findings much. In contrast, when I asked a friend who works with some people doing consciousness research, he reported back that they generally felt that GWT/GNW-style theories have a reasonably firm basis. This isn't terribly conclusive but at least it's a bit of evidence.

Second, for some experiments the book explicitly mentions that they have been replicated. That said, some of the reported experiments seemed to be one-off ones, and I did not yet investigate the details of the claimed replications.

Third, this is a work of cognitive neuroscience. Cognitive neuroscience is generally considered a subfield of cognitive psychology, and cognitive psychology is the part of psychology whose results have so far replicated the best. One recent study tested nine key findings from cognitive psychology, and found that they all replicated. The 2015 "Estimating the reproducibility of Psychological Science" study, managed to replicate 50% of recent results in cognitive psychology, as opposed to 25% of results in social psychology. (If 50% sounds low, remember that we should expect some true results to also fail a single replication, so a 50% replication rate doesn't imply that 50% of the results would be false. Also, a field with a 90% replication rate would probably be too conservative in choosing which experiments to try.) Cognitive psychology replicating pretty well is probably because it deals with phenomena which are much easier to rigorously define and test than social psychology does, so in that regard it's closer to physics than it is to social psychology.

On several occasions, the book reports something like "people did an experiment X, but then someone questioned whether the results of that experiment really supported the hypothesis in question or not, so an experiment X+Y was done that repeated X but also tested Y, to help distinguish between two possible interpretations of X". The general vibe that I get from the book is that different people have different intuitions about how consciousness works, and when someone reports a result that contradicts the intuitions of other researchers, those other researchers are going to propose an alternative interpretation that saves their original intuition. Then people keep doing more experiments until at least one of the intuitions is conclusively disproven - replicating the original experiments in the process.

The analysis of the general reliability of cognitive psychology is somewhat complicated by the fact that these findings are not pure cognitive psychology, but rather cognitive *neuroscience*. Neuroscience is somewhat more removed from just reporting objective findings, since the statistical models used for analyzing the findings can be flawed. I've seen various claims about the problems with statistical tools in neuroscience, but I haven't really dug enough into the field to say to what extent those are a genuine problem.

As suggestive evidence, a lecturer who teaches a "How reliable is cognitive neuroscience?" course reports that before taking a recent iteration of the course, the majority of students

answered the question “If you read about a finding that has been demonstrated across multiple papers in multiple journals by multiple authors, how likely do you think that finding is to be reliable?” as “Extremely likely” and some “Moderately likely”. After taking the course, “Moderately likely” became the most common response with a little under half of the responses, followed by “Slightly likely” by around a quarter of responses and “Extremely likely” with a little over 10% of the responses. Based on this, we might conclude that cognitive neuroscience is moderately reliable, at least as judged by MSc students who’ve just spent time reading and discussing lots of papers critical of cognitive neuroscience.

One thing that’s worth noting is that many of the experiments, including many of the ones this book is reporting on, include two components: a behavioral component and a neuroimaging component. If the statistical models used for interpreting the brain imaging results were flawed, you might get an incorrect impression of what was happening in the brain, but the behavioral results would still be valid. If you’re maximally skeptical of neuroscience, you could choose to throw all of the “inside the brain” results from the book away, and only look at the behavioral results. That seems too conservative to me, but it’s an option. Several of the experiments in the book also use either EEG or single-unit recordings rather than neuroimaging ones; these are much older and simpler techniques than brain imaging is, so are easier to analyze reliably.

So overall, I would expect that the broad strokes of what’s claimed in the book are reasonably correct, even if some of the details might be off.

Defining consciousness

Given that consciousness is a term loaded with many different interpretations, Dehaene reasonably starts out by explaining what he means by consciousness. He distinguishes between three different terms:

- **Vigilance:** whether we “are conscious” in the sense of being awake vs. asleep.
- **Attention:** having focused our mental resources on a specific piece of information.
- **Conscious access:** some of the information we were focusing on, entering our awareness and becoming reportable to others.

For instance, we might be awake (that is, *vigilant*) and staring hard at a computer screen, waiting for some image to be displayed. When that image does get displayed, our *attention* will be on it. But it might flash too quickly for us to report what it looked like, or even for us to realize that it was on the screen in the first place. If so, we don’t have *conscious access* to the thing that we just saw. Whereas if it had been shown for a longer time, we would have conscious access to it.

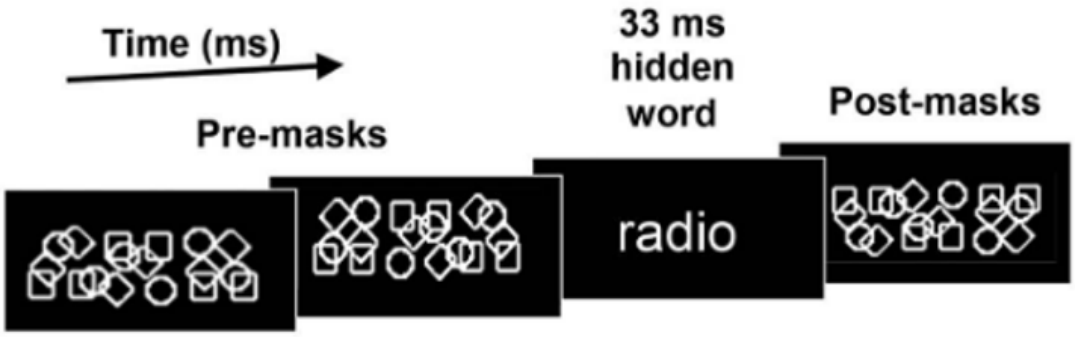
Dehaene says that when he’s talking about consciousness, he’s talking about conscious access, and also that he doesn’t particularly care to debate philosophy and whether this is really *the* consciousness. Rather, since we have a clearly-defined thing which we can investigate using scientific methods, we should just do that, and then think about philosophy once we better understand the empirical side of things.

It seems correct to say that studying conscious access is going to tell us many interesting things, even if it doesn’t solve literally *all* the philosophical questions about consciousness. In the rest of this article, I’ll just follow his conventions and use “consciousness” as a synonym for “conscious access”.

Unconscious processing of meaning

A key type of experiment in Dehaene’s work is *subliminal masking*. Test subjects are told to stare at a screen and report what they see. A computer program shows various geometric

shapes (masks) on the screen. Then at some point, the masks are for a very brief duration replaced with something more meaningful, such as the word “radio”. If the word "radio" is sandwiched between mask shapes, showing it for a sufficiently brief time makes it invisible. The subjects don't even register a brief flicker, as they might if the screen had been totally blank before the word appeared.



By varying the duration for which the word is shown, researchers can control whether or not the subjects see it. Around 40 milliseconds, it is invisible to everyone. Once the duration reaches a certain threshold, which varies somewhat by person but is around 50 milliseconds, the word will be seen around half of the time. When people report not seeing a word, they also fail to name it when asked some time after the trial.

However, even when a masked target doesn’t make it into consciousness, some part of the brain still sees it. It seems as if the visual subsystem started processing the visual stimulus and parsing it in terms of its meaning, but the results of those computations then never made it all the way to consciousness.

One line of evidence for this are subliminal priming experiments, not to be confused with the controversial “social priming” effects in social psychology; unlike those effects, these kinds of priming experiments are well-defined and have been replicated many times. An example of a subliminal priming experiment involves first flashing a hidden word (a prime) so quickly that the participants don’t see it, then following it by a visible word (the target). For instance, people may be primed using the word “radio”, then shown the target word “house”. They are then asked to classify the target word, by e.g. pressing one button if the target word referred to a living thing and another button if it referred to an object.

Subliminal repetition priming refers to the finding that, if the prime and target are the same word and separated by less than a second, then the person will be quicker to classify the target and less likely to make a mistake.

There are indications that when this happens, the brain has parsed some of the prime’s semantic meaning and matched it against the target's meaning. For example, priming works even when the prime is in lower case (radio) and the target is in upper case (RADIO). This might not seem surprising, but look at the difference between e.g. “a” and “A”. These are rather distinct shapes, which we’ve only learned to associate with each other due to cultural convention. Furthermore, while the prime of “range” speeds up the processing of “RANGE”, using “anger” as a prime for “RANGE” has no effect, despite “range” and “anger” having the same letters in a different order. The priming effect comes from the meaning of the prime, rather than just its visual appearance.

The parsing of meaning is not limited to words. If a chess master is shown a simplified chess position for 20 milliseconds, masked so as to make it invisible, they are faster to classify a visible chess position as a check if the hidden position was also a check, and vice versa.

I have reported the above results as saying that the brain does unconscious processing about the meaning of what it sees, but that interpretation has been controversial. After all, something like word processing or identifying a position in check when you have extensive chess experience, is extremely overlearned and could represent an isolated special case rather than showing that the brain processes meaning more generally. The book goes into more detail about the history of this debate and differing interpretations that were proposed; I won't summarize that history in detail, but will just discuss a selection of some experiments which also showed unconscious processing of meaning.

In arithmetic priming experiments, people are first shown a masked single-digit number and then a visible one. They are asked to say whether the target number is larger or smaller than 5. When the number used as a prime is congruent with the target (e.g. smaller than 5 when the target number is also smaller than 5), people respond more quickly than if the two are incongruent. Follow-up work has shown that the effect replicates even if the numbers used as primes are shown in writing ("four") and the target ones as digits ("4"). The priming even works when the prime is an invisible *visual* number and the target a conscious *spoken* number.

Further experiments have shown that the priming effect is the strongest if the prime is the same number as the target number (4 preceding 4). The effect then decreases the more distant the prime is from the target number: 3 preceding 4 shows less of a priming effect, but it still has more of a priming effect than 2 preceding 4 does, and so on. Thus, the brain has done something like extracting an abstract representation of the magnitude of the prime, and used that to influence the processing of the 'target's magnitude.

Numbers could also be argued to be a special case for which we have specialized processing, but later experiments have also shown congruity effects for words in general. For example, when people are shown the word "piano" and asked to indicate whether it is an object or an animal, priming them with a word from a congruent category ("chair") facilitates the correct response, while an incongruent prime ("cat") hinders it.

Some epilepsy patients have had electrodes inserted into their skull for treatment purposes. Some of them have also agreed to have those electrodes used for this kind of research. When they are shown invisible "scary" words such as *danger, rape,* or *poison*, electrodes implanted near the amygdala - the part of the brain involved in fear processing - register an increased level of activation, which is absent for neutral words such as *fridge* or *sonata*.

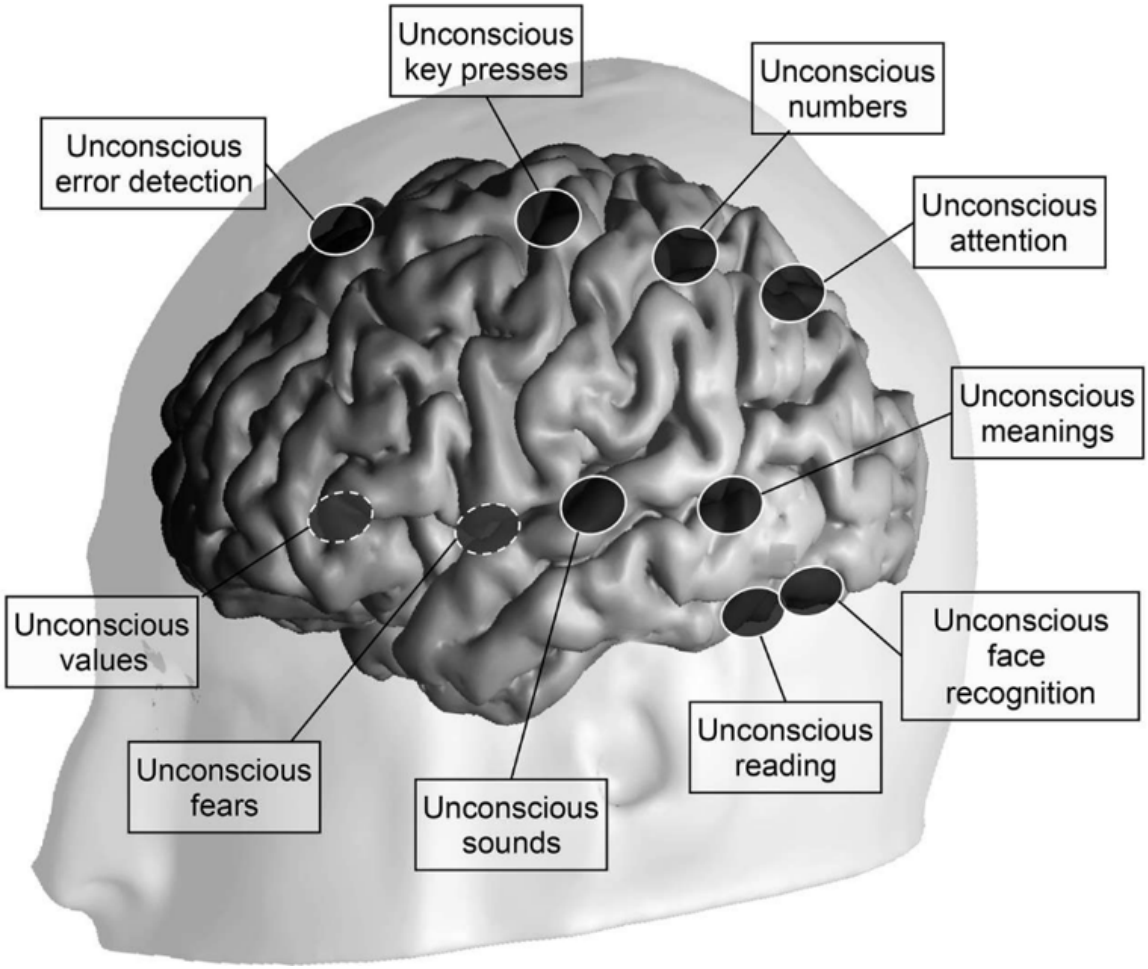
In one study, subjects were shown a "signal", and then had to guess whether to press a button or not press it. As soon as they pressed it, they were told whether they had guessed correctly (earning money) or incorrectly (losing money). Unknown to them, each signal was preceded by a masked shape, which indicated the correct response: one kind of a shape indicated that pressing the button would earn them money, another shape indicated that not pressing the button would earn them money, and a third one meant that either choice had an equal chance of being correct. Even though the subjects were never aware of seeing the shape, once enough trials had passed, they started getting many more results correct than chance alone would indicate. An unconscious value system had associated the shapes with different actions, and was using the subliminal primes for choosing the right action.

Unconscious processing can also weigh the average value of a number of variables. In one type of experiment, subjects are choosing cards from four different decks. Each deck has cards that cause the subject to either earn or lose reward money, with each deck having a different distribution of cards. Two of the decks are "bad", causing the subjects to lose money on net, and two of them are "good", causing them to gain money on net. By the end of the experiment, subjects have consciously figured out which one is which, and can easily report this. However, measurements of skin conductance indicate that even before they have consciously figured out the good and bad decks, there comes a point when they've pulled enough cards that being about to draw a card from a bad deck causes their hands to

sweat. A subconscious process has already started generating a prediction of which decks are bad, and is producing a subliminal gut feeling.

A similar unconscious averaging of several variables can also be shown using the subliminal priming paradigm. Subjects are shown five arrows one at a time, some of which point left and some of which point right. They are then asked for the direction that the majority of the arrows were pointing to. When the arrows are made invisible by subliminal masking, subjects who are forced to guess feel like they are just making random guesses, but are in fact responding much more accurately than by chance alone.

There more examples in the book, but these should be enough to convey the general idea: that many different sensory inputs are automatically registered and processed in the brain, even if they are never shown for long enough to make it all the way to consciousness. Unconsciously processed stimuli can even cause movement commands to be generated in the motor cortex and sent to the muscles, though not necessarily at an intensity which would sufficient to cause actual action.



What about consciousness, then?

So given everything that our brain does automatically and without conscious awareness, what’s up with consciousness? What is it, and what does it do?

Some clues can be found from investigating the neural difference between conscious and unconscious stimuli. Remember that masking experiments show a threshold effect in whether a stimulus is seen or not: if a stimulus which is preceded by a mask is shown for 40 milliseconds, then it's invisible, but around 50 milliseconds it starts to become visible. In experiments where the duration of the stimulus is carefully varied, there is an all-or-nothing effect: subjects do not report seeing more and more of the stimulus as the duration is gradually increased. Rather they either see it in its entirety, or they see nothing at all.

A key finding, replicated across different sensory modalities and different methods for measuring brain activation (fMRI, EEG, and MEG) is that a stimulus becoming conscious involves an effect where, once the strength of a stimulus exceeds a certain threshold, the neural signal associated with that stimulus is massively boosted and spreads to regions in the brain which it wouldn't have reached otherwise. Exceeding the key threshold causes the neural signal generated by the sensory regions to be amplified, with the result that the associated signal could be spread more widely, rather than fading away before it ever reached all the regions.

Dehaene writes, when discussing an experiment where this was measured using visually flashed words as the stimulus:

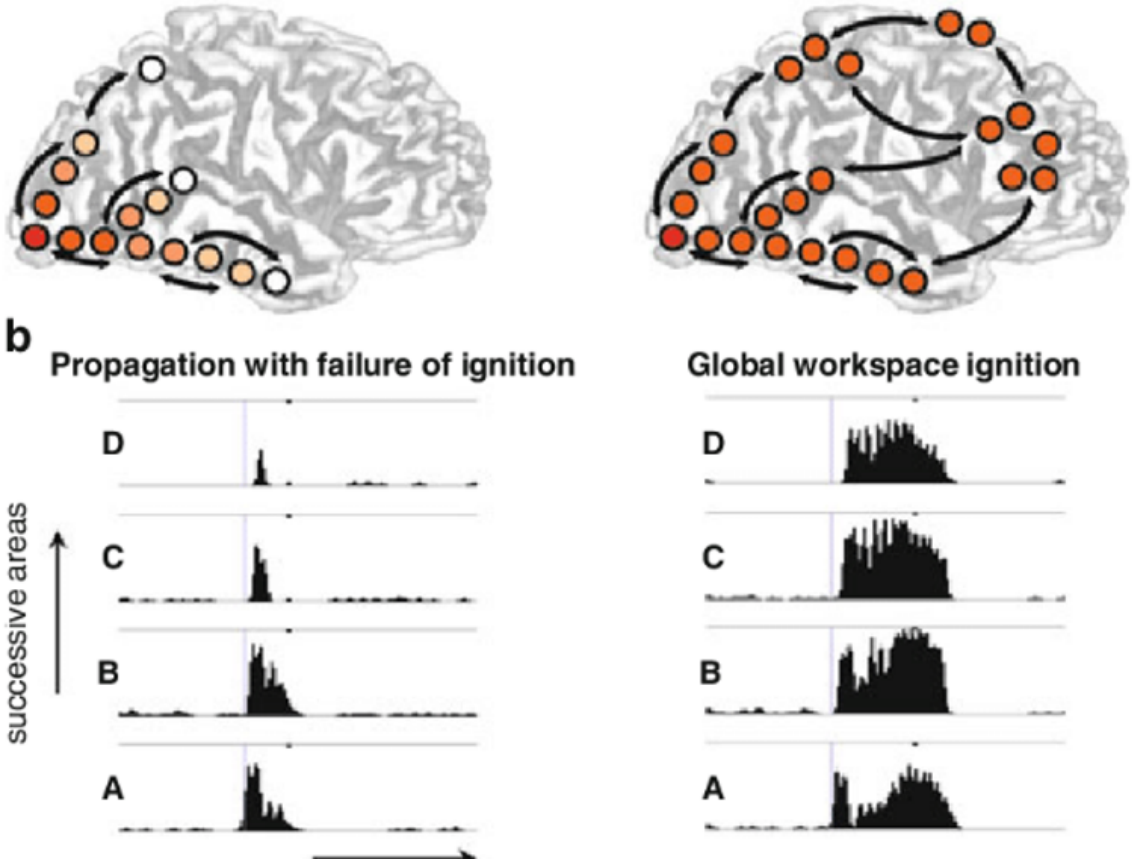
> By measuring the amplitude of this activity, we discovered that the amplification factor, which distinguishes conscious from unconscious processing, varies across the successive regions of the visual input pathway. At the first cortical stage, the primary visual cortex, the activation evoked by an unseen flashed word is strong enough to be easily detectable. However, as it progresses forward into the cortex, masking makes it lose strength. Subliminal perception can thus be compared to a surf wave that looms large on the horizon but merely licks your feet when it reaches the shore. By comparison, conscious perception is a tsunami—or perhaps an avalanche is a better metaphor, because conscious activation seems to pick up strength as it progresses, much as a minuscule snowball gathers snow and ultimately triggers a landslide.
>
> To bring this point home, in my experiments I flashed words for only 43 milliseconds, thereby injecting minimal evidence into the retina. Nevertheless, activation progressed forward and, on conscious trials, ceaselessly amplified itself until it caused a major activation in many regions. Distant brain regions also became tightly correlated: the incoming wave peaked and receded simultaneously in all areas, suggesting that they exchanged messages that reinforced one another until they turned into an unstoppable avalanche. Synchrony was much stronger for conscious than for unconscious targets, suggesting that correlated activity is an important factor in conscious perception.
>
> These simple experiments thus yielded a first signature of consciousness: an amplification of sensory brain activity, progressively gathering strength and invading multiple regions of the parietal and prefrontal lobes. This signature pattern has often been replicated, even in modalities outside vision. For instance, imagine that you are sitting in a noisy fMRI machine. From time to time, through earphones, you hear a brief pulse of additional sound. Unknown to you, the sound level of these pulses is carefully set so that you detect only half of them. This is an ideal way to compare conscious and unconscious perception, this time in the auditory modality. And the result is equally clear: unconscious sounds activate only the cortex surrounding the primary auditory area, and again, on conscious trials, an avalanche of brain activity amplifies this early sensory activation and breaks into the inferior parietal and prefrontal areas

Dehaene goes into a considerable amount of detail about the different neuronal signatures which have been found to correlate with consciousness, and the experimental paradigms which have been used to test whether or not those signatures are mere correlates rather than parts of the causal mechanism. I won't review all of that discussion here, but will summarize some of his conclusions.

Consciousness involves a neural signal activating self-reinforcing loops of activity, which causes wide brain regions to synchronize to process that signal.

Consider what happens when someone in the audience of a performance starts clapping their hands, soon causing the whole audience to burst into applause. As one person starts clapping, other people hear it and start clapping in turn; this becomes a self-reinforcing effect where your clapping causes other people to clap, and you are more likely to continue clapping if other people are also still clapping. In a similar way, the threshold effect of conscious activation seems to involve some neurons sending a signal, causing other neurons to activate and join in on broadcasting that signal. The activation threshold is a point where enough neurons have sufficient mutual excitation to create a self-sustaining avalanche of excitation, spreading throughout the brain.

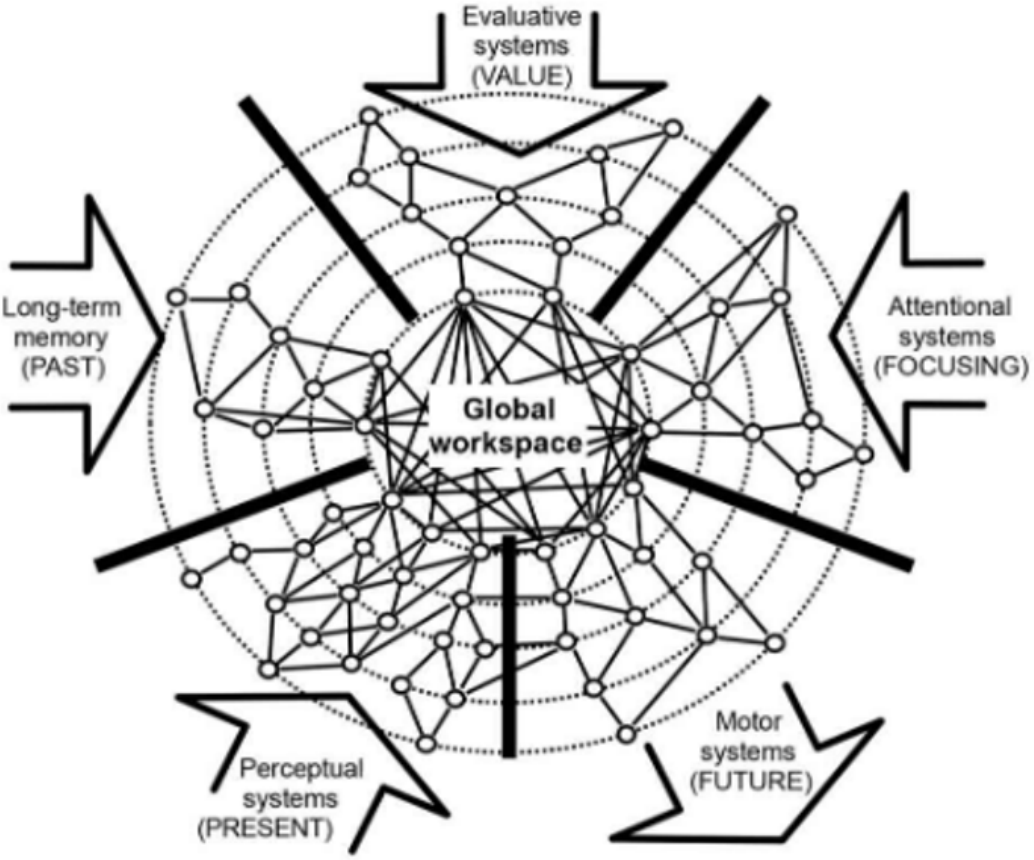
The spread of activation is further facilitated by a “brain web” of long-distance neurons, which link multiple areas of the brain together into a densely interconnected network. Not all areas of the brain are organized in this way: for instance, sensory regions are mostly connected to their immediate neighbors, with visual area V1 being primarily only connected to visual area V2, and V2 mostly to V1 and V3, and so on. But higher areas of the cortex are much more joined together, in a network where area A projecting activity to area B usually means that area B also projects activity back to area A. They also involve triangular connections, where region A might project into regions B and C, which then both also project to each other and back to A. This long-distance network joins not only areas of the cortex, but is also connected to regions such as the thalamus (associated with e.g. attention and vigilance), the basal ganglia (involved in decision-making and action), and the hippocampus (involved in episodic memory).



A stimulus becoming conscious involves the signal associated with it achieving enough strength to activate some of the associative areas that are connected with this network, which Dehaene calls the “global neuronal workspace” (GNW). As those areas start broadcasting the signal associated with the stimulus, other areas in the network receive it and start broadcasting it in turn, creating the self-sustaining loop. As this happens, many different regions will end up processing the signal at the same time, synchronizing their processing around the contents of that signal. Dehaene suggests that the things that we are conscious of at any given moment, are exactly the things which are being processed in our GNW at that moment.

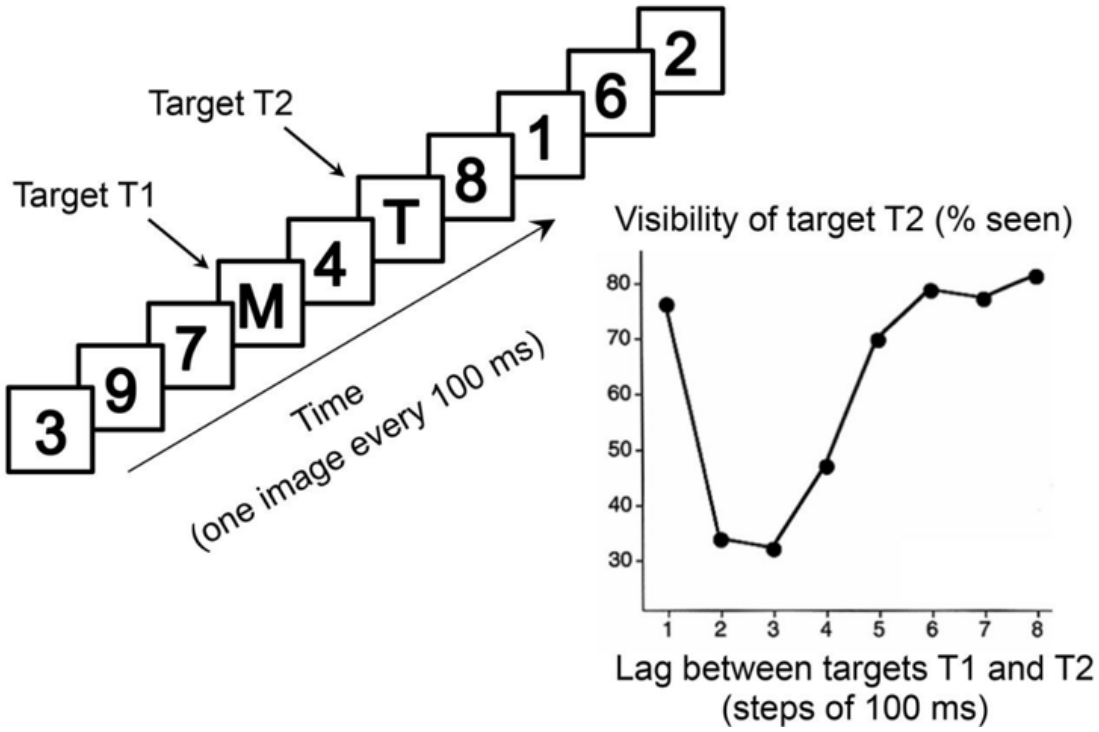
Dehaene describes this as “a decentralized organization without a single physical meeting site” where “an elitist board of executives, distributed in distant territories, stays in sync by exchanging a plethora of messages”. While he mostly reviews evidence gathered from investigating sensory inputs, his model holds that besides sensory areas, many other regions - such as the ones associated with memory and attention - also feed into and manipulate the contents of the network. Once a stimulus enters the GNW, networks regulating top-down attention can amplify and “help keep alive” stimuli which seems especially important to focus on, and memory networks can commit the stimulus into memory, insert into the network earlier memories which were triggered by the sight of the stimulus, or both.

In the experiments on subliminal processing, an unconscious prime may affect the processing of a conscious stimulus that comes very soon afterwards, but since its activation soon fades out, it can’t be committed to memory or verbally reported on afterwards. A stimulus becoming conscious and being maintained in the GNW, both keeps its signal alive for longer, and also allows it better access to memory networks which may store it in order for it to be re-broadcast into the GNW later.

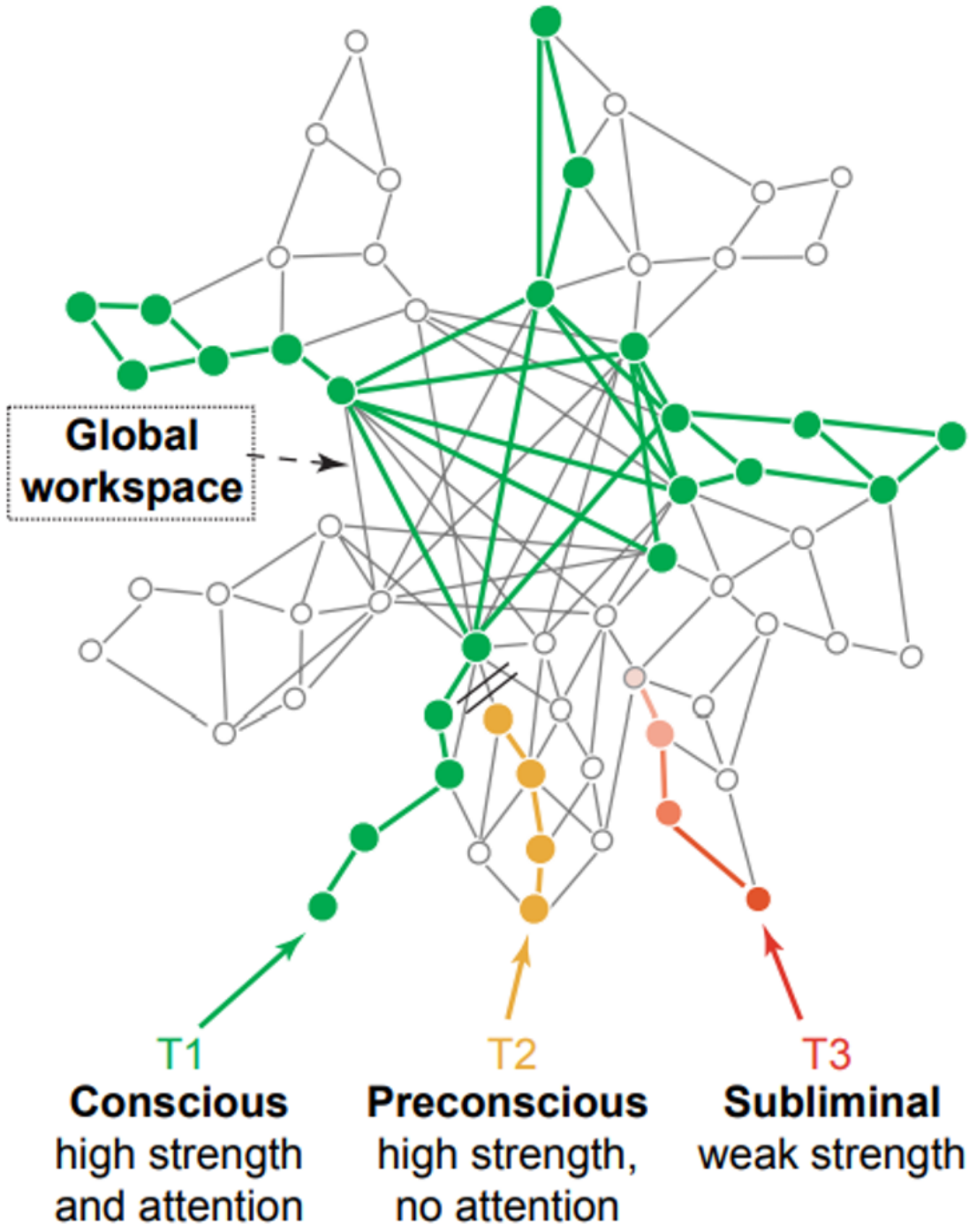


The global workspace can only be processing a single item at a time.

Various experiments show the existence of an “attentional blink”: if your attention is strongly focused on one thing, it takes some time to disengage from it and reorient your attention to something else. For instance, in one experiment people are shown a stream of symbols. Most of the symbols are digits, but some are letters. People are told to remember the letters. While the first letter is easy to remember, if two letters are shown in rapid succession, the subjects might not even realize that two of them were present - and they might be surprised to learn that this was the case. The act of attending to the first letter enough to memorize it, creates a “blink of the mind” which prevents the second letter from ever being noticed.



Dehaene’s explanation for this is that the GNW can only be processing a single item at once. The first letter is seen, processed by the early visual centers, then reaches sufficient strength to make it into the workspace. This causes the workspace neurons to synchronize their processing around the first letter and try to keep the signal active for long enough for it to be memorized - and while they are still doing so, the second letter shows up. It is also processed by the visual regions and makes it to the associative region, but the attention networks are still reinforcing the signal associated with the original letter and keeping it active in the workspace. The new letter can’t muster enough activation in time to get its signal broadcast into the workspace, so by the time the activation generated by the first letter starts to fade, the signal from the second letter has also faded out. As a result, the second signal never makes it to the workspace where it could leave a conscious memory trace of having been observed.



When two simultaneous events happen, it doesn't *always* mean that awareness of the other one is suppressed. If there isn't too much distraction - due to "internal noise, distracting thoughts, or other incoming stimuli" - the signal of the second event may survive for long enough in an unconscious buffer, making it to the GNW after the first event has been processed. The use of a post-stimulus masking shape in the subliminal masking experiments helps erase the contents of this buffer, by providing a new stimulus that overwrites the old one. In these cases, people's judgment of the timing of the events is systematically wrong: rather than experiencing the events to have happened simultaneously, they believe the second event to have happened at the time when the event entered their consciousness.

As an interesting aside, as a result of these effects, the content of our consciousness is always slightly delayed relative to when an event actually happened - a stimulus getting into the GNW takes at least one-third of a second, and may take substantially longer if we are distracted. The brain contains a number of mechanisms for compensating the delay in GNW access, such as prediction mechanisms which anticipate how familiar events should happen before they've actually happened.

Disrupting or stimulating the GNW, has the effects that this theory would predict.

One of the lines of argument by which Dehaene defends the claim that GNW activity is genuinely the same thing as conscious activity, and not a more correlate, is that artificially interfering with GNW activity has the kinds of effects that we might expect.

To do this, we can use Transcranial Magentic Stimulation (TMS) to create magnetic fields which stimulate electric activity in the brain, or if electrodes have been placed in a person's brain, those can be used to stimulate the brain directly.

In one experiment, TMS was used to stimulate the visual cortex of test subjects, in a way that created a hallucination of light. By varying the intensity of the stimulation, the researchers could control whether or not the subjects noticed anything. On trials when the subjects reported becoming conscious of a hallucination, an avalanche wave associated with consciousness popped up, reaching consciousness faster than normal. In Dehaene's interpretation, the magnetic pulse bypassed the normal initial processing stages for vision and instead created a neuronal activation directly at a higher cortical area, speeding up conscious access by about 0.1 seconds.

Experiments have also used TMS to successfully erase awareness of a stimulus. One experiment described in the book uses a dual TMS setup. First, a subject is zapped with a magnetic pulse that causes them to see a bit of (non-existent) movement. After it has been confirmed that subjects report becoming conscious of movement when they are zapped with the first pulse, they are then subjected to a trial where they are first zapped with the same pulse, then immediately thereafter with another pulse that's aimed to disrupt the signal from getting access to the GNW. When this is done, subjects report no longer being aware of having seen any movement.

The functions of consciousness

So what exactly *is* the function of consciousness? Dehaene offers four different functions.

Conscious sampling of unconscious statistics and integration of complicated information

Suppose that you a Bayesian decision theorist trying to choose between two options, A and B. For each two options, you've computed a probability distribution about the possible outcomes that may result if you choose either A or B. In order to actually make your choice, you need to collapse your probability distributions into a point estimate of the expected value of choosing A versus B, to know which one is actually better.

In Dehaene's account, consciousness does something like this. We have a number of unconscious systems which are constantly doing Bayesian statistics and constructing probability distributions about how to e.g. interpret visually ambiguous stimuli, weighing multiple hypotheses at the same time. In order for decision-making to actually be carried

out, the system has to choose one of the interpretations, and act based on the assumption of that interpretation being correct. The hypothesis that the unconscious process selects as correct, is then what gets fed into consciousness. For example, when I look at the cup of tea in front of me, I don't see a vast jumble of shifting hypotheses of what this visual information might represent: rather, I just see what I think is a cup of tea, which is what a subconscious process has chosen as the most likely interpretation.

Dehaene offers the analogy of the US President being briefed by the FBI. The FBI is a vast organization, with thousands of employees: they are constantly shifting through enormous amounts of data, and forming hypotheses about topics which have national security relevance. But it would be useless for the FBI to present to the President every single report collected by every single field agent, as well as every analysis compiled by every single analyst in response. Rather, the FBI needs to internally settle on some overall summary of what they believe is going on, and then present that to the President, who can then act based on the information. Similarly, Dehaene suggests that consciousness is a place where different brain systems can exchange summaries of their models, and to integrate conflicting evidence in order to arrive to an overall conclusion.

Dehaene discusses a few experiments which lend support this interpretation, though here the discussion seems somewhat more speculative than in other parts of the book. One of his pieces of evidence is of recordings of neuronal circuits which integrate many parts of a visual scene into an overall image, resolving local ambiguities by using information from other parts of the image. Under anesthesia, neuronal recordings show that this integration process is disrupted; consciousness "is needed for neurons to exchange signals in both bottom-up and top-down directions until they agree with each other". Another experiment shows that if people are shown an artificial stimulus which has been deliberately crafted to be ambiguous, people's conscious impression of the correct interpretation keeps shifting: first it's one interpretation, then the other. By varying the parameters of the stimulus, researchers can control roughly how often people see each interpretation. If Bayesian statistics would suggest that interpretation A was 30% likely and interpretation B 70% likely, say, then people's impression of the image will keep shifting so that they will see interpretation A roughly 30% of the time and interpretation B roughly 70% of the time.

> What we see, at any time, tends to be the most likely interpretation, but other possibilities occasionally pop up and stay in our conscious vision for a time duration that is proportional to their statistical likelihood. Our unconscious perception works out the probabilities - and then our consciousness samples from them at random.

In Dehaene's account, consciousness is involved in higher-level integration of the meaning of concepts. For instance, our understanding of a painting such as the Mona Lisa is composed of many different things. Personally, if I think about the Mona Lisa, I see a mental image of the painting itself, I get an association with the country of Italy, I remember having first learned about the painting in a Donald Duck story, and I also remember my friend telling me about the time she saw the original painting itself. These are different pieces of information, stored in different formats in different regions of the brain, and the kind of global neuronal integration carried out by the GNW allows all of these different interpretations to come together, with every system participating in constructing an overall coherent, synchronous interpretation.

All of this sounds sensible enough. At the same, after all the previous discussion about unconscious decision-making and unconscious integration of information, this leaves me feeling somewhat unsatisfied. If it has been shown that e.g. unconsciously processed cues are enough to guide our decision-making, then how do we square that with the claim that consciousness is necessary for settling on a single interpretation that would allow us to take actions?

My interpretation is that even though unconscious processing and decision-making happens, its effect is relatively weak. If you prime people with a masked stimulus, then that influences

their decision-making so as to give them better performance - but it doesn't give them *perfect* performance. In the experiment where masked cues predicted the right action and unconscious learning associated each cue with the relevant action, the subjects only ended up with an average of 63% correct actions.

Looking at the cited paper itself, the authors themselves note that if the cues had been visible, it would only have taken a couple of trials for the subjects to learn the optimal behaviors. In the actual experiment, their performance slowly improved until it reached a plateau around 60 trials. Thus, even though unconscious learning and decision-making happens, conscious learning and decision-making can be significantly more effective.

Second, while I don't see Dehaene mentioning it, I've always liked the PRISM theory of consciousness, which suggests that one of the functions of consciousness is to be a place for resolving conflicting plans for controlling the skeletal muscles. In the unconscious decision-making experiments, the tasks have mostly been pretty simple, and only involved the kinds of goals that could all be encapsulated within a single motivational system. In real life however, we often run into situations where different brain systems output conflicting instructions. For instance, if we are carrying a hot cup of tea, our desire to drop the cup may be competing against our desire to carry it to the table, and these may have their origin in very different sorts of motivations. Information from both systems would need to be taken into account and integrated in order to make an overall decision.

To stretch Dehaene's FBI metaphor: as long as the FBI is doing things that fall within their jurisdiction and they are equipped to handle, then they can just do that without getting in contact with the President. But if the head of the FBI and the head of the CIA have conflicting ideas about what should be done, on a topic on which the two agencies have overlapping jurisdiction, then it might be necessary to bring the disagreement out in the open so that a higher-up can make the call. Of course, there isn't any single "President" in the brain who would make the final decision: rather, it's more like the chiefs of all the other alphabet soup bureaus were also called in, and they then hashed out the details of their understanding until they came to a shared agreement about what to do.

Lasting thoughts and working memory

As already touched upon, consciousness is associated with memory. Unconsciously registered information tends to fade very quickly and then disappear. In all the masking experiments, the duration between the prime and the target is very brief; if the duration would be any longer, there would be no learning or effect on decision-making. For e.g. associating cues and outcomes with each other over an extended period of time, the cue has to be consciously perceived.

Dehaene describes an experiment which demonstrates exactly this:

> The cognitive scientists Robert Clark and Larry Squire conducted a wonderfully simple test of temporal synthesis: time-lapse conditioning of the eyelid reflex. At a precisely timed moment, a pneumatic machine puffs air toward the eye. The reaction is instantaneous: in rabbits and humans alike, the protective membrane of the eyelid immediately closes. Now precede the delivery of air with a brief warning tone. The outcome is called Pavlovian conditioning (in memory of the Russian physiologist Ivan Petrovich Pavlov, who first conditioned dogs to salivate at the sound of a bell, in anticipation of food). After a short training, the eye blinks to the sound itself, in anticipation of the air puff. After a while, an occasional presentation of the isolated tone suffices to induce the "eyes wide shut" response.
>
> The eye-closure reflex is fast, but is it conscious or unconscious? The answer, surprisingly, depends on the presence of a temporal gap. In one version of the test, usually termed "delayed conditioning," the tone lasts until the puff arrives. Thus the two

stimuli briefly coincide in the animal's brain, making the learning a simple matter of coincidence detection. In the other, called "trace conditioning," the tone is brief, separated from the subsequent air puff by an empty gap. This version, although minimally different, is clearly more challenging. The organism must keep an active memory trace of the past tone in order to discover its systematic relation to the subsequent air puff. To avoid any confusion, I will call the first version "coincidence-based conditioning" (the first stimulus lasts long enough to coincide with the second, thus removing any need for memory) and the second "memory-trace conditioning" (the subject must keep in mind a memory trace of the sound in order to bridge the temporal gap between it and the obnoxious air puff).

The experimental results are clear: coincidence-based conditioning occurs unconsciously, while for memory-trace conditioning, a conscious mind is required. In fact, coincidence-based conditioning does not require any cortex at all. A decerebrate rabbit, without any cerebral cortex, basal ganglia, limbic system, thalamus, and hypothalamus, still shows eyelid conditioning when the sound and the puff overlap in time. In memory-trace conditioning, however, no learning occurs unless the hippocampus and its connected structures (which include the prefrontal cortex) are intact. In human subjects, memory-trace learning seems to occur if and only if the person reports being aware of the systematic predictive link between the tone and the air puff. Elderly people, amnesiacs, and people who were simply too distracted to notice the temporal relationship show no conditioning at all (whereas these manipulations have no effect whatsoever on coincidence-based conditioning). Brain imaging shows that the subjects who gain awareness are precisely those who activate their prefrontal cortex and hippocampus during the learning.

Carrying out artificial serial operations

Consider what happens when you calculate 12 * 13 in your head.

When you do so, you have some conscious awareness of the steps involved: maybe you first remember that 12 * 12 = 144 and then add 144 + 12, or maybe you first multiply 12 * 10 = 120 and then keep that result in memory as you multiply 12 * 3 = 36 and then add 120 + 36. Regardless of the strategy, the calculation happens consciously.

Dehaene holds that this kind of multi-step arithmetic can't happen unconsciously. We can do *single-step* arithmetic unconsciously: for example, people can be shown a single masked digit *n*, and then be asked to carry out one of three operations. People might be asked to name the digit (the "*n*" task), to add 2 to *n* and report the resulting the number (the "*n* + 2" task), or to report whether or not it's smaller than 5 (the "*n* < 5" task). On all of these tasks, even if people haven't consciously seen the digit, when they are forced to guess they typically get the right answer half of the time.

However, unconscious *two-step* arithmetic fails. If people are flashed an invisible digit and told to first add 2 to it, and then report whether the result is more or less than 5 (the "(*n* + 2) > 5" task), their performance is on the chance level. The unconscious mind can carry out a single arithmetic operation, but it can't then store the result of that operation and use it as the input of a second operation, even though it could carry out either of the two operations alone.

Dehaene notes that this might seem to contradict a previous finding, which is that the unconscious brain can *accumulate* multiple pieces of information over time. For instance, in the arrow experiment, people were shown several masked arrows one at a time; at the end, they could tell whether most of them had been pointing to the left or to the right. Dehaene says that the difference is that opening a neural circuit which accumulates multiple observations is a single operation for the brain: and while the accumulator stores information

of how many arrows have been observed so far, that information can't be taken out of it and used as an input for a second calculation.

The accumulator also can't reach a decision by itself: for instance, if people saw the arrows consciously, they could reach a decision after having seen three arrows that pointed one way, knowing that the remaining arrows couldn't change the overall decision anymore. In unconscious trials, they can't use this kind of strategic reasoning: the unconscious circuit can only keep adding up the arrows, rather than adding up the arrows *and* also checking whether a rule of type "if seen_arrows > 3" has been satisfied yet.

According to Dehaene, implementing such rules is one of the functions of consciousness. In fact, he explicitly compares consciousness to a production system: an AI design which holds a number of objects in a working memory, and also contains a number of IF-THEN rules, such as "if there is an A in working memory, change it to sequence BC". If multiple rules match, one of them is chosen for execution according to some criteria. After one of the rules has fired, the contents of the working memory gets updated, and the cycle repeats. The conscious mind, Dehaene says, works using a similar principle - creating a biological Turing machine that can combine operations from a number of neuronal modules, flexibly chaining them together for serial execution.

A social sharing device

If a thought is conscious, we can describe it and report it to other people using language. I won't elaborate on this, given that the advantages of being able to use language to communicate with others are presumably obvious. I'll just note that Dehaene highlights one interesting perspective: one where other people are viewed as additional modules that can carry out transformations on the objects in the workspace.

Whether it's a subsystem in the brain that's applying production rules to the workspace contents, or whether you are communicating the contents to another person who then comments on it (as guided by some subsystem in *their* brain), the same principle of "production rules transforming the workspace contents" still applies. Only in one of the cases, the rules and transformations come from subsystems that are located within a single brain, and in the other case subsystems from multiple brains are engaged in joint manipulation of the contents - though of course the linguistic transmission is lossy, since subsystems in multiple brains can't communicate with the same bandwidth as subsystems in a single brain. (Yet.)

Other stuff

Dehaene also discusses a bunch of other things in his book: for instance, he talks about comatose patients and how his research has been applied to study their brains, in order to predict which patients will eventually recover and which ones will remain permanently unresponsive. This is pretty cool, and feels like a confirmation of the theories being on the right track, but since it's no longer elaborating on the mechanisms and functions of consciousness, I won't cover that here.

Takeaways for the rest of the sequence

This has been a pretty long post. Now that we're at the end, I'm just going to highlight a few of the points which will be most important when we go forward in the multiagent minds sequence:

- Consciousness can only contain a single mental object at a time.

- The brain has multiple different systems doing different things; many of the systems do unconscious processing of information. When a mental object becomes conscious, many systems will synchronize their processing around analyzing and manipulating that mental object.
- The brain can be compared to a production system, with a large number of specialized rules which fire in response to specific kinds of mental objects. E.g. when doing mental arithmetic, applying the right sequence of arithmetic operations for achieving the main goal.

If we take the view of looking at various neural systems as being literally technically subagents, then we can reframe the above points as follows:

- The brain has multiple subagents doing different things; many of the subagents do unconscious processing of information. When a mental object becomes conscious, many subagents will synchronize their processing around analyzing and manipulating that mental object.
- The collective of subagents can only have their joint attention focused on one mental object at a time.
- The brain can be compared to a production system, with a large number of subagents carrying out various tasks when they see the kinds of mental objects that they care about. E.g. when doing mental arithmetic, applying the right sequence of mental operations for achieving the main goal.

Next up: constructing a mechanistic sketch of how a mind might work, combining the above points as well as the kinds of mechanisms that have already been demonstrated in contemporary machine learning, to finally end up with a model that pretty closely resembles the Internal Family Systems one.

Building up to an Internal Family Systems model

Introduction

Internal Family Systems (IFS) is a psychotherapy school/technique/model which lends itself particularly well for being used alone or with a peer. For years, I had noticed that many of the kinds of people who put in a lot of work into developing their emotional and communication skills, some within the rationalist community and some outside it, kept mentioning IFS.

So I looked at the Wikipedia page about the IFS model, and bounced off, since it sounded like nonsense to me. Then someone brought it up again, and I thought that maybe I should reconsider. So I looked at the WP page again, thought "nah, still nonsense", and continued to ignore it.

This continued until I participated in CFAR mentorship training last September, and we had a class on CFAR's Internal Double Crux (IDC) technique. IDC clicked really well for me, so I started using it a lot and also facilitating it to some friends. However, once we started using it on more emotional issues (as opposed to just things with empirical facts pointing in different directions), we started running into some weird things, which it felt like IDC couldn't quite handle… things which reminded me of how people had been describing IFS. So I finally read up on it, and have been successfully applying it ever since.

In this post, I'll try to describe and motivate IFS in terms which are less likely to give people in this audience the same kind of a "no, that's nonsense" reaction as I initially had.

Epistemic status

This post is intended to give an argument for why *something like* the IFS model *could* be true and a thing that works. It's not really an argument that IFS *is* correct. My reason for thinking in terms of IFS is simply that I was initially super-skeptical of it (more on the reasons of my skepticism later), but then started encountering things which it turned out IFS predicted - and I only found out about IFS predicting those things *after* I familiarized myself with it.

Additionally, I now feel that IFS gives me significantly more gears for understanding the behavior of both other people and myself, and it has been significantly transformative in addressing my own emotional issues. Several other people who I know report it having been similarly powerful for them. On the other hand, aside for a few isolated papers with titles like "proof-of-concept" or "pilot study", there seems to be conspicuously little peer-reviewed evidence in favor of IFS, meaning that we should probably exercise some caution.

I think that, even if not completely correct, IFS is currently the best model that I have for explaining the observations that it's pointing at. I encourage you to read this post

in the style of learning soft skills - trying on this perspective, and seeing if there's anything in the description which feels like it resonates with your experiences.

But before we talk about IFS, let's first talk about building robots. It turns out that if we put together some existing ideas from machine learning and neuroscience, we can end up with a robot design that pretty closely resembles IFS's model of the human mind.

What follows is an intentionally simplified story, which is simpler than *either* the full IFS model *or* a full account that would incorporate everything that I know about human brains. Its intent is to demonstrate that an agent architecture with IFS-style subagents might easily emerge from basic machine learning principles, without claiming that all the details of that toy model would exactly match human brains. A discussion of what exactly IFS *does* claim in the context of human brains follows after the robot story.

Wanted: a robot which avoids catastrophes

Suppose that we're building a robot that we want to be generally intelligent. The hot thing these days seems to be deep reinforcement learning, so we decide to use that. The robot will explore its environment, try out various things, and gradually develop habits and preferences as it accumulates experience. (Just like those human babies.)

Now, there are some problems we need to address. For one, deep reinforcement learning works fine in simulated environments where you're safe to explore for an indefinite duration. However, it runs into problems if the robot is supposed to learn in a real life environment. Some actions which the robot might take will result in catastrophic consequences, such as it being damaged. If the robot is just doing things at random, it might end up damaging itself. Even worse, if the robot does something which could have been catastrophic but narrowly avoids harm, it might then forget about it and end up doing the same thing again!

How could we deal with this? Well, let's look at the existing literature. Lipton et al. (2016) proposed what seems like a promising idea for addressing the part about forgetting. Their approach is to explicitly maintain a memory of *danger states* - situations which are not the catastrophic outcome itself, but from which the learner has previously ended up in a catastrophe. For instance, if "being burned by a hot stove" is a catastrophe, then "being about to poke your finger in the stove" is a danger state. Depending on how cautious we want to be and how many preceding states we want to include in our list of danger states, "going near the stove" and "seeing the stove" can also be danger states, though then we might end up with a seriously stove-phobic robot.

In any case, we maintain a separate storage of danger states, in such a way that the learner never forgets about them. We use this storage of danger states to train a *fear model*: a model which is trying to predict the probability of ending up in a catastrophe from some given novel situation. For example, maybe our robot poked its robot finger at the stove in our kitchen, but poking its robot finger at stoves in other kitchens might be dangerous too. So we want the fear model to generalize from our stove to other stoves. On the other hand, we don't want it to be stove-phobic and run away at the mere sight of a stove. The task of our fear model is to predict exactly how likely it

is for the robot to end up in a catastrophe, given some situation it is in, and then make it increasingly disinclined to end up in the kinds of situations which might lead to a catastrophe.

This sounds nice in theory. On the other hand, Lipton et al. are still assuming that they can train their learner in a simulated environment, and that they can label catastrophic states ahead of time. We don't know in advance every possible catastrophe our robot might end up in - it might walk off a cliff, shoot itself in the foot with a laser gun, be beaten up by activists protesting technological unemployment, or any number of other possibilities.

So let's take inspiration from humans. We can't know beforehand every bad thing that might happen to our robot, but we can identify some classes of things which are correlated with catastrophe. For instance, being beaten or shooting itself in the foot will cause physical damage, so we can install sensors which indicate when the robot has taken physical damage. If these sensors - let's call them "pain" sensors - register a high amount of damage, we consider the situation to have been catastrophic. When they do, we save that situation and the situations preceding it to our list of dangerous situations. Assuming that our robot has managed to make it out of that situation intact and can do anything in the first place, we use that list of dangerous situations to train up a fear model.

At this point, we notice that this is starting to remind us about our experience with humans. For example, the infamous Little Albert experiment. A human baby was allowed to play with a laboratory rat, but each time that he saw the rat, a researcher made a loud scary sound behind his back. Soon Albert started getting scared whenever he saw the rat - and then he got scared of furry things in general.

Something like Albert's behavior could be implemented very simply using something like Hebbian conditioning to get a learning algorithm which picks up on some features of the situation, and then triggers a panic reaction whenever it re-encounters those same features. For instance, it registers that the sight of fur and loud sounds tend to coincide, and then it triggers a fear reaction whenever it sees fur. This would be a basic fear model, and a "danger state" would be "seeing fur".

Wanting to keep things simple, we decide to use this kind of an approach as the fear model of our robot. Also, having read *Consciousness and the Brain*, we remember a few basic principles about how those human brains work, which we decide to copy because we're lazy and don't want to come up with entirely new principles:

- There's a special network of neurons in the brain, called the *global neuronal workspace*. The contents of this workspace are roughly the same as the contents of consciousness.
- We can thus consider consciousness a workspace which many different brain systems have access to. It can hold a single "chunk" of information at a time.
- The brain has multiple different systems doing different things. When a mental object becomes conscious (that is, is projected into the workspace by a subsystem), many systems will synchronize their processing around analyzing and manipulating that mental object.

So here is our design:

- The robot has a hardwired system scanning for signs of catastrophe. This system has several subcomponents. One of them scans the "pain" sensors for signs of

physical damage. Another system watches the “hunger” sensors for signs of low battery.
- Any of these “distress” systems can, alone or in combination, feed a negative reward signal into the global workspace. This tells the rest of the system that this is a bad state, from which the robot should escape.
- If a certain threshold level of “distress” is reached, the current situation is designated as *catastrophic*. All other priorities are suspended and the robot will prioritize getting out of the situation. A memory of the situation and the situations preceding it are saved to a dedicated storage.
- After the experience, the memory of the catastrophic situation is replayed in consciousness for analysis. This replay is used to train up a separate fear model which effectively acts as a new “distress” system.
- As the robot walks around its environment, sensory information about the surroundings will enter its consciousness workspace. When it plans future actions, simulated sensory information about how those actions would unfold enters the workspace. Whenever the new fear model detects features in either kind of sensory information which it associates with the catastrophic events, it will feed “fear”-type “distress” into the consciousness workspace.

So if the robot sees things which remind it of poking at hot stove, it will be inclined to go somewhere else; if it imagines doing something which would cause it to poke at the hot stove, then it will be inclined to imagine doing something else.

Introducing managers

But is this actually enough? We've now basically set up an algorithm which warns the robot when it sees things which have previously preceded a bad outcome. This might be enough for dealing with static tasks, such as not burning yourself at a stove. But it seems insufficient for dealing with things like predators or technological unemployment protesters, who might show up in a wide variety of places and actively try to hunt you down. By the time you see a sign of them, you're already in danger. It would be better if we could learn to avoid them entirely, so that the fear model would never even be triggered.

As we ponder this dilemma, we surf the web and run across this blog post summarizing Saunders, Sastry, Stuhlmüller & Evans (2017). They are also concerned with preventing reinforcement learning agents from running into catastrophes, but have a somewhat different approach. In their approach, a reinforcement learner is allowed to do different kinds of things, which a human overseer then allows or blocks. A separate “blocker” model is trained to predict which actions the human overseer would block. In the future, if the robot would ever take an action which the “blocker” predicts the human overseer would disallow, it will block that action. In effect, the system consists of two separate subagents, one subagent trying to maximize rewards and the other subagent trying to block non-approved actions.

Since our robot has a nice modular architecture into which we can add various subagents which are listening in and taking actions, we decide to take inspiration from this idea. We create a system for spawning dedicated subprograms which try to predict and and block actions which would cause the fear model to be triggered. In theory, this is unnecessary: given enough time, even standard reinforcement learning should learn to avoid the situations which trigger the fear model. But again, trial-and-error can take a very long time to learn exactly which situations trigger fear, so we dedicate a separate subprogram to the task of pre-emptively figuring it out.

Each fear model is paired with a subagent that we'll call a *manager*. While the fear model has associated a bunch of cues with the notion of an impending catastrophe, the manager learns to predict which situations would cause the fear model to trigger. Despite sounding similar, these are not the same thing: one indicates when you are already in danger, the other is trying to figure out what you can do to never end up in danger in the first place. A fear model might learn to recognize signs which technological unemployment protesters commonly wear. Whereas a manager might learn the kinds of environments where the fear model has noticed protesters before: for instance, near the protester HQ.

Then, if a manager predicts that a given action (such as going to the protester HQ) would eventually trigger the fear model, it will block that action and promote some other action. We can use the interaction of these subsystems to try to ensure that the robot only feels fear in situations which already resemble the catastrophic situation so much as to actually *be* dangerous. At the same time, the robot will be unafraid to take safe actions in situations from which it *could* end up in a danger zone, but are themselves safe to be in.

As an added benefit, we can recycle the manager component to also do the same thing as the blocker component in the Saunders et al. paper originally did. That is, if the robot has a human overseer telling it in strict terms not to do some things, it can create a manager subprogram which models that overseer and likewise blocks the robot from doing things which the model predicts that the overseer would disapprove of.

Putting together a toy model

If the robot *does* end up in a situation where the fear model is sounding an alarm, then we want to get it out of the situation as quickly as possible. It may be worth spawning a specialized subroutine just for this purpose. Technological unemployment activists could, among other things, use flamethrowers that set the robot on fire. So let's call these types of subprograms dedicated to escaping from the danger zone, *firefighters*.

So how does the system as a whole work? First, the different subagents act by sending into the consciousness workspace various mental objects, such as an emotion of fear, or an intent to e.g. make breakfast. If several subagents are submitting identical mental objects, we say that they are voting for the same object. On each time-step, one of the submitted objects is chosen at random to become the contents of the workspace, with each object having a chance to be selected that's proportional to its number of votes. If a mental object describing a physical action (an "intention") ends up in the workspace and stays chosen for several time-steps, then that action gets executed by a motor subsystem.

Depending on the situation, some subagents will have more votes than others. E.g. a fear model submitting a fear object gets a number of votes proportional to how strongly it is activated. Besides the specialized subagents we've discussed, there's also a default planning subagent, which is just taking whatever actions (that is, sending to the workspace whatever mental objects) it thinks will produce the greatest reward. This subagent only has a small number of votes.

Finally, there's a self-narrative agent which is constructing a narrative of the robot's actions as if it was a unified agent, for social purposes and for doing reasoning

afterwards. After the motor system has taken an action, the self-narrative agent records this as something like “I, Robby the Robot, made breakfast by cooking eggs and bacon”, transmitting this statement to the workspace and saving it to an episodic memory store for future reference.

Consequences of the model

Is this design any good? Let’s consider a few of its implications.

First, in order for the robot to take physical actions, the intent to do so has to be in its consciousness for a long enough time for the action to be taken. If there are any subagents that wish to prevent this from happening, they must muster enough votes to bring into consciousness some other mental object replacing that intention before it’s been around for enough time-steps to be executed by the motor system. (This is analogous to the concept of the final veto in humans, where consciousness is the last place to block pre-consciously initiated actions before they are taken.)

Second, the different subagents do not see each other directly: they only see the consequences of each other’s actions, as that’s what’s reflected in the contents of the workspace. In particular, the self-narrative agent has no access to information about which subagents were responsible for generating which physical action. It only sees the intentions which preceded the various actions, and the actions themselves. Thus it might easily end up constructing a narrative which creates the internal appearance of a single agent, even though the system is actually composed of multiple subagents.

Third, even if the subagents can’t directly see each other, they might still end up forming alliances. For example, if the robot is standing near the stove, a curiosity-driven subagent might propose poking at the stove (“I want to see if this causes us to burn ourselves again!”), while the default planning system might propose cooking dinner, since that’s what it predicts will please the human owner. Now, a manager trying to prevent a fear model agent from being activated, will eventually learn that if it votes for the default planning system’s intentions to cook dinner (which it saw earlier), then the curiosity-driven agent is less likely to get *its* intentions into consciousness. Thus, no poking at the stove, and the manager’s and the default planning system’s goals end up aligned.

Fourth, this design can make it *really difficult* for the robot to even become aware of the existence of some managers. A manager may learn to support any other mental processes which block the robot from taking specific actions. It does it by voting in favor of mental objects which orient behavior towards anything else. This might manifest as something subtle, such as a mysterious lack of interest towards something that sounds like a good idea in principle, or just repeatedly forgetting to do something, as the robot always seems to get distracted by something else. The self-narrative agent, not having any idea of what’s going on, might just explain this as “Robby the Robot is forgetful sometimes” in its internal narrative.

Fifth, the default planning subagent here is doing something like rational planning, but given its weak voting power, it’s likely to be overruled if other subagents disagree with it (unless some subagents also agree with it). If some actions seem worth doing, but there are managers which are blocking it and the default planning subagent doesn’t have an explicit representation of them, this can manifest as all kinds of procrastinating behaviors and numerous failed attempts for the default planning

system to “try to get itself to do something”, using various strategies. But as long as the managers keep blocking those actions, the system is likely to remain stuck.

Sixth, the purpose of both managers and firefighters is to keep the robot out of a situation that has been previously designated as dangerous. Managers do this by trying to pre-emptively block actions that would cause the fear model agent to activate; firefighters do this by trying to take actions which shut down the fear model agent after it has activated. But the fear model agent activating is not actually the *same* thing as being in a dangerous situation. Thus, both managers and firefighters may fall victim to Goodhart’s law, doing things which block the fear model while being irrelevant for escaping catastrophic situations.

For example, “thinking about the consequences of going to the activist HQ” is something that might activate the fear model agent, so a manager might try to block just *thinking* about it. This has obvious consequence that the robot can’t think clearly about that issue. Similarly, once the fear model has already activated, a firefighter might Goodhart by supporting *any* action which helps activate an agent with a lot of voting power that’s going to think about something entirely different. This could result in compulsive behaviors which were effective at pushing the fear aside, but useless for achieving any of the robot’s actual aims.

At worst, this could cause loops of mutually activating subagents pushing in opposite directions. First, a stove-phobic robot runs away from the stove as it was about to make breakfast. Then a firefighter trying to suppress that fear, causes the robot to get stuck looking at pictures of beautiful naked robots, which is engrossing and thus great for removing the fear of the stove. Then another fear model starts to activate, this one afraid of failure and of spending so much time looking at pictures of beautiful naked robots that the robot won’t accomplish its goal of making breakfast. A separate firefighter associated with this second fear model has learned that focusing the robot’s attention on the pictures of beautiful naked robots *even more* is the most effective action for keeping this new fear temporarily subdued. So the two firefighters are allied and temporarily successful at their goal, but then the first one - seeing that the original stove fear has disappeared - turns off. Without the first firefighter’s votes supporting the second firefighter, the fear manages to overwhelm the second firefighter, causing the robot to rush into making breakfast. This again activates its fear of the stove, but *if* the fear of failure remains strong enough, it might overpower its fear of the stove so that the robot manages to make breakfast in time...

Hmm. Maybe this design isn’t so great after all. Good thing we noticed these failure modes, so that there aren’t any mind architectures like this going around being vulnerable to them!

The Internal Family Systems model

But enough hypothetical robot design; let’s get to the topic of IFS. The IFS model hypothesizes the existence of three kinds of “extreme parts” in the human mind:

- **Exiles** are said to be parts of the mind which hold the memory of past traumatic events, which the person did not have the resources to handle. They are parts of the psyche which have been split off from the rest and are frozen in time of the traumatic event. When something causes them to surface, they tend to flood the mind with pain. For example, someone may have an exile associated with times when they were romantically rejected in the past.

- **Managers** are parts that have been tasked with keeping the exiles permanently exiled from consciousness. They try to arrange a person's life and psyche so that exiles never surface. For example, managers might keep someone from reaching out to potential dates due to a fear of rejection.
- **Firefighters** react when exiles have been triggered, and try to either suppress the exile's pain or distract the mind from it. For example, after someone has been rejected by a date, they might find themselves drinking in an attempt to numb the pain.
- Some presentations of the IFS model simplify things by combining Managers and Firefighters into the broader category of **Protectors**, so only talk about Exiles and Protectors.

Exiles are not limited to being created from the kinds of situations that we would commonly consider seriously traumatic. They can also be created from things like relatively minor childhood upsets, as long as the child didn't feel like they could handle the situation.

IFS further claims that you can treat these parts as something like independent subpersonalities. You can communicate with them, consider their worries, and gradually persuade managers and firefighters to give you access to the exiles that have been kept away from consciousness. When you do this, you can show them that you are no longer in the situation which was catastrophic before, and now have the resources to handle it if something similar was to happen again. This heals the exile, and also lets the managers and firefighters assume better, healthier roles.

As I mentioned in the beginning, when I first heard about IFS, I was turned off by it for several different reasons. For instance, here were some of my thoughts at the time:

1. The whole model about some parts of the mind being in pain, and other parts trying to suppress their suffering. The thing about exiles was framed in terms of *a part of the mind splitting off in order to protect the rest of the mind against damage.* What? That doesn't make any evolutionary sense! A traumatic situation is *just sensory information* for the brain, it's not literal brain damage: it wouldn't have made any sense for minds to evolve in a way that caused parts of it to split off, forcing other parts of the mind to try to keep them suppressed. Why not just… never be damaged in the first place?
2. That whole thing about parts being personalized characters that you could talk to. That… doesn't describe anything in my experience.
3. Also, how does just talking to yourself fix any trauma or deeply ingrained behaviors?
4. IFS talks about everyone having a "True Self". Quote from Wikipedia: *IFS also sees people as being whole, underneath this collection of parts. Everyone has a true self or spiritual center, known as the Self to distinguish it from the parts. Even people whose experience is dominated by parts have access to this Self and its healing qualities of curiosity, connectedness, compassion, and calmness. IFS sees the therapist's job as helping the client to disentangle themselves from their parts and access the Self, which can then connect with each part and heal it, so that the parts can let go of their destructive roles and enter into a harmonious collaboration, led by the Self.* That… again did not sound particularly derived from any sensible psychology.

Hopefully, I've already answered my past self's concerns about the first point. The model itself talks in terms of managers protecting the mind from pain, exiles being exiled from consciousness in order for their pain to remain suppressed, etc. Which is a

reasonable description *of the subjective experience* of what happens. But the evolutionary logic - as far as I can guess - is slightly different: to keep us out of dangerous situations.

The story of the robot describes the actual “design rationale”. Exiles are in fact subagents which are “frozen in the time of a traumatic event”, but they didn’t split off to protect the rest of the mind from damage. *Rather, they were created as an isolated memory block to ensure that the memory of the event wouldn’t be forgotten*. Managers then exist to keep the person away from such catastrophic situations, and firefighters exist to help escape them. Unfortunately, this setup is vulnerable to various failure modes, similar to those that the robot is vulnerable to.

With that said, let’s tackle the remaining problems that I had with IFS.

Personalized characters

IFS suggests that you can experience the exiles, managers and firefighters in your mind as something akin to subpersonalities - entities with their own names, visual appearances, preferences, beliefs, and so on. Furthermore, this isn’t inherently dysfunctional, nor indicative of something like Dissociative Identity Disorder. Rather, even people who are entirely healthy and normal may experience this kind of “multiplicity”.

Now, it’s important to note right off that not everyone has this to a major extent: you don’t *need* to experience multiplicity in order for the IFS process to work. For instance, my parts feel more like bodily sensations and shards of desire than subpersonalities, but IFS still works super-well for me.

In the book *Internal Family Systems Therapy*, Richard Schwartz, the developer of IFS, notes that if a person’s subagents play well together, then that person is likely to feel mostly internally unified. On the other hand, if a person has lots of internal conflict, then they are more likely to experience themselves as having multiple parts with conflicting desires.

I think that this makes a lot of sense, assuming the existence of something like a self-narrative subagent. If you remember, this is the part of the mind which looks at the actions that the mind-system has taken, and then constructs an explanation for why those actions were taken. (See e.g. the posts on the limits of introspection and on the Apologist and the Revolutionary for previous evidence for the existence of such a confabulating subagent with limited access to our true motivations.) As long as all the exiles, managers and firefighters are functioning in a unified fashion, the most parsimonious model that the self-narrative subagent might construct is simply that of a unified self. But if the system keeps being driven into strongly conflicting behaviors, then it can’t necessarily make sense of them from a single-agent perspective. Then it might naturally settle on something like a multiagent approach and experience itself as being split into parts.

Kevin Simler, in Neurons Gone Wild, notes how people with strong addictions seem particularly prone to developing multi-agent narratives:

> *This American Life* did a nice segment on addiction a few years back, in which the producers — seemingly on a lark — asked people to personify their addictions. "It was like people had been waiting all their lives for somebody to ask them this

> question," said the producers, and they gushed forth with descriptions of the 'voice' of their inner addict:
>
> "*The voice is irresistible, always. I'm in the thrall of that voice.*"
>
> "*Totally out of control. It's got this life of its own, and I can't tame it anymore.*"
>
> "*I actually have a name for the voice. I call it Stan. Stan is the guy who tells me to have the extra glass of wine. Stan is the guy who tells me to smoke.*"

This doesn't seem like it explains all of it, though. I've frequently been very dysfunctional, and have always found very intuitive the notion of the mind being split into very parts. Yet I mostly still don't seem to experience my subagents anywhere near as person-like as some others clearly do. I know at least one person who ended up finding IFS because of having all of these talking characters in their head, and who was looking for something that would help them make sense of it. Nothing like that has ever been the case for me: I did experience strongly conflicting desires, but they were just that, strongly conflicting desires.

I can only surmise that it has something to do with the same kinds of differences which cause some people to think mainly verbally, others mainly visually, and others yet in some other hard-to-describe modality. Some fiction writers spontaneously experience their characters as real people who speak to them and will even bother the writer when at the supermarket, and some others don't.

It's been noted that the mechanisms which use to model ourselves and other people overlap - not very surprisingly, since both we and other people are (presumably) humans. So it seems reasonable that some of the mechanisms for representing other people, would sometimes also end up spontaneously recruited for representing internal subagents or coalitions of them.

Why should this technique be useful for psychological healing?

Okay, suppose it's possible to access our subagents somehow. Why would just talking with these entities in your own head, help you fix psychological issues?

Let's consider that a person having exiles, managers and firefighters is costly in the sense of constraining that person's options. If you never want to do anything that would cause you to see a stove, that limits quite a bit of what you can do. I strongly suspect that many forms of procrastination and failure to do things we'd like to do are mostly a manifestation of overactive managers. So it's important not to create those kinds of entities unless the situation really *is* one which should be designated as categorically unacceptable to end up in.

The theory for IFS mentions that not all painful situations turn into trauma: just ones in which we felt helpless and like we didn't have the necessary resources for dealing with it. This makes sense, since if we were capable of dealing with it, then the situation can't have been that catastrophic. The aftermath of the immediate event is important as well: a child who ends up in a painful situation doesn't necessarily end up traumatized, if they have an adult who can put the event in a reassuring context afterwards.

But situations which used to be catastrophic and impossible for us to handle before, aren't necessarily that any more. It seems important to have a mechanism for updating that cache of catastrophic events and for disassembling the protections around it, if the protections turn out to be unnecessary.

How does that process usually happen, without IFS or any other specialized form of therapy?

Often, by talking about your experiences with someone you trust. Or writing about them in private or in a blog.

In my post about *Consciousness and the Brain*, I mentioned that once a mental object becomes conscious, many different brain systems synchronize their processing around it. I suspect that the reason why many people have such a powerful urge to discuss their traumatic experiences with someone else, is that doing so is a way of bringing those memories into consciousness in detail. And once you've dug up your traumatic memories from their cache, their content can be re-processed and re-evaluated. If your brain judges that you now *do* have the resources to handle that event if you ever end up in it again, or if it's something that simply can't happen anymore, then the memory can be removed from the cache and you no longer need to avoid it.

I think it's also significant that, while something like just writing about a traumatic event is sometimes enough to heal, often it's more effective if you have a sympathetic listener who you trust. Traumas often involve some amount of shame: maybe you were called lazy as a kid and are still afraid of others thinking that you are lazy. Here, having friends who accept you and are willing to nonjudgmentally listen while you talk about your issues, is by itself an indication that the thing that you used to be afraid of isn't a danger anymore: there exist people who will stay by your side despite knowing your secret.

Now, when you are talking to a friend about your traumatic memory, you will be going through cached memories that have been stored in an exile subagent. A specific memory circuit - one of several circuits specialized for the act of holding painful memories - is active and outputting its contents into the global workspace, from which they are being turned into words.

Meaning that, in a sense, *your friend is talking directly to your exile.*

Could you hack this process, so that you wouldn't even *need* a friend, and could carry this process out entirely internally?

In my earlier post, I remarked that you could view language as a way of joining two people's brains together. A subagent in your brain outputs something that appears in your consciousness, you communicate it to a friend, it appears in their consciousness, subagents in your friend's brain manipulate the information somehow, and then they send it back to your consciousness.

If you are telling your friend about your trauma, you are in a sense joining your workspaces together, and letting some subagents in *your* workspace, communicate with the "sympathetic listener" subagents in *your friend's* workspace.

So why not let a "sympathetic listener" subagent in your workspace, hook up directly with the traumatized subagents that are also in your own workspace?

I think that something like this happens when you do IFS. You are using a technique designed to activate the relevant subagents in a very specific way, which allows for this kind of a "hooking up" without needing another person.

For instance, suppose that you are talking to a manager subagent which wants to hide the fact that you're bad at something, and starts reacting defensively whenever the topic is brought up. Now, one way by which its activation could manifest, is feeding those defensive thoughts and reactions directly into your workspace. In such a case, you would experience them as your *own* thoughts, and possibly as objectively real. IFS calls this "blending"; I've also previously used the term "cognitive fusion" for what's essentially the same thing.

Instead of remaining blended, you then use various unblending / cognitive defusion techniques that highlight the way by which these thoughts and emotions are coming from a specific part of your mind. You could think of this as wrapping extra content around the thoughts and emotions, and then seeing them through the wrapper (which is obviously not-you), rather than experiencing the thoughts and emotions directly (which you might experience as your own). For example, the IFS book *Self-Therapy* suggests this unblending technique (among others):

> Allow a visual image of the part [subagent] to arise. This will give you the sense of it as a separate entity. This approach is even more effective if the part is clearly a certain distance away from you. The further away it is, the more separation this creates.
>
> Another way to accomplish visual separation is to draw or paint an image of the part. Or you can choose an object from your home that represents the part for you or find an image of it in a magazine or on the Internet. Having a concrete token of the part helps to create separation.

I think of this as something like, you are taking the subagent in question, routing its responses through a visualization subsystem, and then you see a talking fox or whatever. And this is then a representation that your internal subsystems for talking with other people can respond to. You can then have a dialogue with the part (verbally or otherwise) in a way where its responses are clearly labeled as coming from it, rather than being mixed together with all the other thoughts in the workspace. This lets the content coming from the sympathetic-listener subagent and the exile/manager/firefighter subagent be kept clearly apart, allowing you to consider the emotional content as you would as an external listener, preventing you from drowning in it. You're hacking your brain so as to work as the therapist and client as the same time.

The Self

IFS claims that, below all the various parts and subagents, there exists a "true self" which you can learn to access. When you are in this Self, you exhibit the qualities of "calmness, curiosity, clarity, compassion, confidence, creativity, courage, and connectedness". Being at least partially in Self is said to be a prerequisite for working with your parts: if you are not, then you are not able to evaluate their models objectively. The parts will sense this, and as a result, they will not share their models properly, preventing the kind of global re-evaluation of their contents that would update them.

This was the part that I was initially the most skeptical of, and which made me most frequently decide that IFS was not worth looking at. I could easily conceptualize the mind as being made up of various subagents. But then it would just be numerous subagents all the way down, without any single one that could be designated the "true" self.

But let's look at IFS's description of how exactly to get into Self. You check whether you seem to be blended with any part. If you are, you unblend with it. Then you check whether you might also be blended with some other part. If you are, you unblend from it also. You then keep doing this until you can find no part that you might be blended with. All that's left are those "eight Cs", which just seem to be a kind of a global state, with no particular part that they would be coming from.

I now think that "being in Self" represents a state where there no particular subagent is getting a disproportionate share of voting power, and everything is processed by the system as a whole. Remember that in the robot story, catastrophic states were situations in which the organism should *never* end up. A subagent kicking in to prevent that from happening is a kind of a priority override to normal thinking. It blocks you from being open and calm and curious *because* some subagent thinks that doing so would be dangerous. If you then turn off or suspend all those priority overrides, then the mind's default state absent any override seems to be one with the qualities of the Self.

This actually fits at least one model of the function of positive emotions pretty well. Fredrickson (1998) suggests that an important function of positive emotions is to make us engage in activities such as play, exploration, and savoring the company of other people. Doing these things has the effect of building up skills, knowledge, social connections, and other kinds of resources which might be useful for us in the future. If there are no active ongoing threats, then that implies that the situation is pretty safe for the time being, making it reasonable to revert to a positive state of being open to exploration.

The *Internal Family Systems Therapy* book makes a somewhat big deal out of the fact that everyone, even most traumatized people, ultimately has a Self which they can access. It explains this in terms of the mind being organized to protect against damage, and with parts always splitting off from the Self when it would otherwise be damaged. I think the real explanation is much simpler: the mind is not accumulating damage, it is just accumulating a longer and longer list of situations not considered safe.

As an aside, this model feels like it makes me less confused about confidence. It seems like people are really attracted to confident people, and that to some extent it's also possible to fake confidence until it becomes genuine. But if confidence is so attractive and we can fake it, why hasn't evolution just made everyone confident by default?

Turns out that it *has*. The reason why faked confidence gradually turns into genuine confidence is that by forcing yourself to act in confident ways which felt dangerous before, your mind gets information indicating that this behavior is not as dangerous as you originally thought. That gradually turns off those priority overrides that kept you out of Self originally, until you get there naturally.

The reason why being in Self is a requirement for doing IFS, is the existence of conflicts between parts. For instance, recall the stove-phobic robot having a firefighter

subagent that caused it to retreat from the stove into watching pictures of beautiful naked robots. This triggered a subagent which was afraid of the naked-robot-watching preventing the robot from achieving its goals. If the robot now tried to do IFS and talk with the firefighter subagent that caused it to run away from stoves, this might bring to mind content which activated the exile that was afraid of not achieving things. Then *that* exile would keep flooding the mind with negative memories, trying to achieve its priority override of "we need to get out of this situation", and preventing the process from proceeding. Thus, all of the subagents that have strong opinions about the situation need to be unblended from, before integration can proceed.

IFS also has a separate concept of "Self-Leadership". This is a process where various subagents eventually come to trust the Self, so that they allow the person to increasingly remain in Self even in various emergencies. IFS views this as a positive development, not only because it feels nice, but because doing so means that the person will have more cognitive resources available for actually dealing with the emergency in question.

I think that this ties back to the original notion of subagents being generated to invoke priority overrides for situations *which the person originally didn't have the resources to handle*. Many of the subagents IFS talks about seem to emerge from childhood experiences. A child has many fewer cognitive, social, and emotional resources for dealing with bad situations, in which case it makes sense to just categorically avoid them, and invoke special overrides to ensure that this happens. A child's cognitive capacities, models of the world, and abilities to self-regulate are also less developed, so she may have a harder time staying out of dangerous situations *without* having some priority overrides built in. An adult, however, typically has many more resources than a child does. Even when faced with an emergency situation, it can be much better to be able to remain calm and analyze the situation using *all* of one's subagents, rather than having a few of them take over all the decision-making. Thus, it seems to me - both theoretically and practically - that developing Self-Leadership is *really* valuable.

That said, I do not wish to imply that it would be a good goal to *never* have negative emotions. Sometimes blending with a subagent, and experiencing resulting negative emotions, is the right thing to do in that situation. Rather than suppressing negative emotions entirely, Self-Leadership aims to get to a state where any emotional reaction tends to be endorsed by the mind-system as a whole. Thus, if feeling angry or sad or bitter or whatever feels appropriate to the situation, you can let yourself feel so, and then give yourself to that emotion without resisting it. As a result, negative emotions become less unpleasant to experience, since there are fewer subagents trying to fight against them. Also, if it turns out that being in a negative emotional state is no longer useful, the system as a whole can just choose to move back into Self.

Final words

I've now given a brief summary of the IFS model, and explained why I think it makes sense. This is of course not enough to establish the model as *true*. But it might help in making the model plausible enough to at least try out.

I think that most people could benefit from learning and doing IFS on themselves, either alone or together with a friend. I've been saying that exiles/managers/firefighters tend to be generated from trauma, but it's important to realize that these events don't need to be anything immensely traumatic. The kinds of

ordinary, normal childhood upsets that everyone has had can generate these kinds of subagents. Remember, just because you think of a childhood event as trivial *now*, doesn't mean that it felt trivial to you *as a child*. Doing IFS work, I've found exiles related to memories and events which I *thought* left no negative traces, but actually did.

Remember also that it can be really hard to notice the presence of some managers: if they are doing their job effectively, then you might never become aware of them directly. "I don't have any trauma so I wouldn't benefit from doing IFS" isn't necessarily correct. Rather, the cues that I use for detecting a need to do internal work are:

- *Do I have the qualities associated with Self, or is something blocking them?*
- *Do I feel like I'm capable of dealing with this situation rationally, and doing the things which feel like good ideas on an intellectual level?*
- *Do my emotional reactions feel like they are endorsed by my mind-system as a whole, or is there a resistance to them?*

If not, there is often some internal conflict which needs to be addressed - and IFS, combined with some other practices such as Focusing and meditation - has been very useful in learning to solve those internal conflicts.

Even if you don't feel convinced that doing IFS personally would be a good idea, I think adopting its framework of exiles, managers and firefighters is useful for better understanding the behavior of other people. Their dynamics will be easier to recognize in other people if you've had some experience recognizing them in yourself, however.

If you want to learn more about IFS, I would recommend starting with Self-Therapy by Jay Earley. In terms of What/How/Why books, my current suggestions would be:

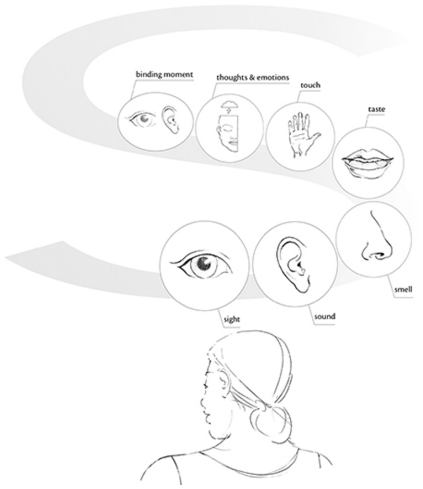
- How: Self-Therapy by Jay Earley.
- What: Internal Family Systems Therapy, by Richard Schwartz
- Why: The Power of Focusing, by Ann Weiser Cornell (technically not about IFS, but AWC's variant of Focusing gets very close to IFS, and is excellent for conveying the right mindset for it)

This post was written as part of research supported by the Foundational Research Institute. Thank you to everyone who provided feedback on earlier drafts of this article: Eli Tyre, Elizabeth Van Nostrand, Jan Kulveit, Juha Törmänen, Lumi Pakkanen, Maija Haavisto, Marcello Herreshoff, Qiaochu Yuan, and Steve Omohundro.

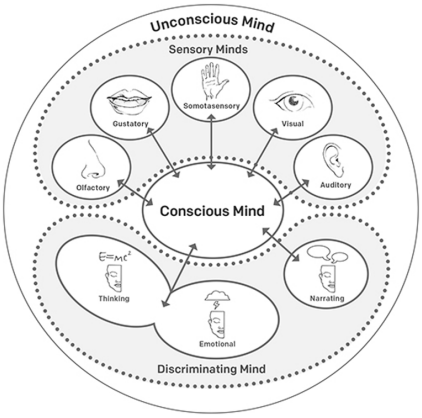
Subagents, introspective awareness, and blending

> In this post, I extend the model of mind that I've been building up in previous posts to explain some things about change blindness, not knowing whether you are conscious, forgetting most of your thoughts, and mistaking your thoughts and emotions as objective facts, while also connecting it with the theory in the meditation book The Mind Illuminated. (If you didn't read my previous posts, this article has been written to also work as a stand-alone piece.)

The Mind Illuminated (Amazon, SSC review), or *TMI* for short, presents what it calls the *moments of consciousness model*. According to this model, our stream of consciousness consists of a series of discrete moments, each a mental object. Under this model, there are always different "subminds" which are projecting mental objects into consciousness. At different moments, different mental objects get selected as the content of consciousness.



If you've read some of the previous posts in this sequence, you may recognize this as sounding familiar. We started by discussing some of the neuroscience research on consciousness. There we covered the GWT/GNW theory of consciousness being a "workspace" in the brain that different brain systems project information into, and which allows them to synchronize their processing around a single piece of information. In the next post, we discussed the psychotherapy model of Internal Family Systems, which also conceives the mind of being composed of different parts, many of which are trying to accomplish various aims by competing to project various mental objects into consciousness. (TMI talks about subminds, IFS talks about parts, GWT/GNW just talks about different parts of the brain; for consistency's sake, I will just use "subagent" in the rest of this post.)



At this point, we might want to look at some criticisms of this kind of a framework. Susan Blackmore has written an interesting paper called "There is no stream of consciousness". She has several examples for why we should reject the notion of any stream of consciousness. For instance, this one:

> For many years now I have been getting my students to ask themselves, as many times as possible every day "Am I conscious now?". Typically they find the task unexpectedly hard to do; and hard to remember to do. But when they do it, it has some very odd effects. First they often report that they always seem to be conscious when they ask the question but become less and less sure about whether they were conscious a moment before. With more practice they say that asking the question itself makes them more conscious, and that they can extend this consciousness from a few seconds to perhaps a minute or two. What does this say about consciousness the rest of the time?
>
> Just this starting exercise (we go on to various elaborations of it as the course progresses) begins to change many students' assumptions about their own experience. In particular they become less sure that there are always contents in their stream of consciousness. How does it seem to you? It is worth deciding at the outset because this is what I am going to deny. I suggest that there is no stream of consciousness. [...]
>
> I want to replace our familiar idea of a stream of consciousness with that of illusory backwards streams. At any time in the brain a whole lot of different things are going on. None of these is either 'in' or 'out' of consciousness, so we don't need to explain the 'difference' between conscious and unconscious processing. Every so often something happens to create what seems to have been a stream. For example, we ask "Am I conscious now?". At this point a retrospective story is concocted about what was in the stream of consciousness a moment before, together with a self who was apparently experiencing it. Of course there was neither a conscious self nor a stream, but it now seems as though there was. This process goes on all the time with new stories being concocted whenever required. At any time that we bother to look, or ask ourselves about it, it seems as though there is a stream of consciousness going on. When we don't bother to ask, or to look, it doesn't, but then we don't notice so it doesn't matter. This way the grand illusion is concocted.

This is an interesting argument. A similar example might be that when I first started doing something like track-back meditation when on walks, checking what was in my mind just a second ago. I was surprised at just how many thoughts I would have while

on a walk, that I would usually just totally forget about afterwards, and come back home having no recollection of 95% of them. This seems similar to Blackmore's "was I conscious just now" question, in that when I started to check back the contents of my mind just a few seconds ago, I was frequently surprised by what I found out. (And yes, I've tried the "was I conscious just now" question as well, with similar results as Blackmore's students.)

Another example that Blackmore cites is change blindness. When people are shown an image, it's often possible to introduce unnoticed major changes into the image, as long as people are not looking at the very location of the change when it's made. Blackmore also interprets this as well to mean that there is no stream of consciousness - we aren't actually building up a detailed visual model of our environment, which we would then experience in our consciousness.

One might summarize this class of objections as something like, "stream-of-consciousness theories assume that there is a conscious stream of mental objects in our minds that we are aware of. However, upon investigation it often becomes apparent that we *haven't* been aware of something that was supposedly in our stream of consciousness. In change blindness experiments we weren't aware of what the changed detail actually was pre-change, and more generally we don't even have clear awareness of *whether we were conscious a moment ago*."

But on the other hand, as we reviewed earlier, there still seem to be objective experiments which establish the existence of something like a "consciousness", which holds approximately one mental object at a time.

So I would interpret Blackmore's findings differently. I agree that answers to questions like "am I conscious right now" are constructed somewhat on the spot, in response to the question. But I don't think that we need to reject having a stream of consciousness because of that. I think that *you can be aware of something, without being aware of the fact that you were aware of it*.

Robots again

For example, let's look at a robot that has something like a global consciousness workspace. Here are the contents of its consciousness on five successive timesteps:

1. It's raining outside
2. Battery low
3. Technological unemployment protestors are outside
4. Battery low
5. I'm now recharging my battery

Notice that at the first timestep, the robot was aware of the fact that it was raining outside; this was the fact being broadcast from consciousness to all subsystems. But at no later timestep was it conscious of the fact that at the first timestep, it had been aware of it raining outside. Assuming that no subagent happened to save this specific piece of information, then all knowledge of it was lost as soon as the content of the consciousness workspace changed.

But suppose that there is some subagent which happens to keep track of what has been happening in consciousness. In that case it may choose to make its memory of previous mind-states consciously available:

6. At time 1, there was the thought that [It's raining outside]

Now there is a mental object in the robot's consciousness, which encodes not only the observation of it raining outside before, *but also* the fact that the system was thinking of this before. That knowledge may then have further effects on the system - for example, when I became aware of how much time I spent on useless rumination while on walks, I got frustrated. And this seems to have contributed to making me ruminate less: as the system's actions and their overall effect were metacognitively represented and made available for the system's decision-making, this had the effect of the system adjusting its behavior to tune down activity that was deemed useless.

The Mind Illuminated calls this *introspective awareness.* Moments of introspective awareness are summaries of the system's previous mental activity, with there being a dedicated subagent with the task of preparing and outputting such summaries. Usually it will only focus on tracking the specific kinds of mental states which seem important to track.

So if we ask ourselves "was I conscious just now" for the first time, that might cause the agent to output *some* representation of the previous state we were in. But it doesn't have experience in answering this question, and if it's anything like most human memory systems, it needs to have some kind of a concrete example of what exactly it is looking for. The first time we ask it, the subagent's pattern-matcher knows that the system is presumably conscious at *this* instant, so it should be looking for some feature in our previous experiences which somehow resembles this moment, but it's not quite sure of which one. And an introspective mind state, is likely to be different from the less introspective mind state that we were in a moment ago.

This has the result that on the first few times when it's asked, the subagent may produce an uncertain answer: it's basically asking its memory store "do my previous mind states resemble this one as judged by some unclear criteria", which is obviously hard to answer.

With time, operating from the assumption that the system is currently conscious, the subagent may learn to find more connections between the current moment and past ones that it still happens to have in memory. Then it will report those as consciousness, and likely also focus more attention on aspects of the current experience which it has learned to consider "conscious". This would match Blackmore's report of "with more practice [the students] say that asking the question itself makes them more conscious, and that they can extend this consciousness from a few seconds to perhaps a minute or two".

Similarly, this explains me not being aware of most of my thoughts, as well as change blindness. I had a stream of thoughts, but because I had not been practicing introspective awareness, there were no moments of introspective awareness making me aware of having had these thoughts. Though I was aware of the thoughts at the time, this was never re-presented in a way that would have left a memory trace.

In change blindness experiments, people might look at the same spot in a picture twice. Although they *did* see the contents of that spot at time 1 and were aware of them, that memory was never stored anywhere. When at time 2 they looked at the same spot and it was different, the lack of an awareness of what they saw previously means that they don't notice the change.

Introspective awareness will be an important concept in my future posts. (Abram Demski also wrote a previous post on Track-Back Meditation, which is basically an

exercise for introspective awareness.) Today, I'm going to talk about its relation to a concept I've talked about before: blending / cognitive fusion.

Blending

I've previously discussed "cognitive fusion", as what happens when the content of a thought or emotion is experienced as an objective truth rather than a mental construct. For instance, you get angry at someone, and the emotion makes you experience them as a horrible person - and in the moment this seems just true to you, rather than being an interpretation created by your emotional reaction.

You can also fuse with more logical-type beliefs - or for that matter any beliefs - when you just treat them as unquestioned truths, without remembering the possibility that they might be wrong. In my previous post, I suggested that many forms of meditation were training the skill of intentional cognitive defusion, but I didn't explain *how* exactly meditation lets you get better at defusion.

In my post about Internal Family Systems, I mentioned that IFS uses the term "blending" for when a subagent is sending emotional content to your consciousness, and suggested that IFS's unblending techniques worked by associating extra content around those thoughts and emotions, allowing you to recognize them as mental objects. For instance, you might notice sensations in your body that were associated with the emotion, and let your mind generate a mental image of what the physical form of those sensations might look like. Then this set of emotions, thoughts, sensations, and visual images becomes "packaged together" in your mind, unambiguously designating it as a mental object.

My current model is that meditation works similarly, only using moments of introspective awareness as the "extra wrapper". Suppose again that you are a robot, and the contents of your consciousness is:

1. It's raining outside.

Note that this mental object is basically being taken as an axiomatic truth: what is in your consciousness, is that it is raining outside.

On the other hand, suppose that your consciousness contains this:

1. Sensor 62 is reporting that [it's raining outside].

Now the mental object in your consciousness contains the origin of the belief that it's raining. The information is made available to various subagents which have other beliefs. E.g. a subagent holding knowledge about sensors, might upon seeing this mental object, recognize the reference to "sensor 62" and output its estimate of that sensor's reliability. The previous two mental objects could then be combined by a third subagent:

1. (Subagent A:) Sensor 62 is reporting that [it is raining outside]
2. (Subagent B:) Readings from sensor 62 are reliable 38% of the time.
3. (Subagent C:) It is raining outside with a 38% probability.

In my discussion of Consciousness and the Brain, I noted that one of the proposed functions of consciousness is to act as a production system, where many different

subagents may identify particular mental objects and then apply various rules to transform the contents of consciousness as a result. What I've sketched above is exactly a sequence of production rules: at e.g. step 2, something like the rule "if sensor 62 is mentioned as an information source, output the current best estimate of sensor 62's reliability" is applied by subagent B. Then at the third timestep, another subagent combines the observations from the previous two timesteps, and sends *that* into consciousness.

What was important was that the system was not representing the outside weather just as an axiomatic statement, but rather it was explicitly representing it as a fallible piece of information with a particular source.

Here's something similar:

1. I am a bad person.
2. At t1, there was the thought that [I am a bad person].

Here, the moment of awareness is highlighting the nature of the previous thought as a thought, thus causing the system to treat it as such. If you used introspective awareness for unblending, it might go something like this:

1. *Blending:* you are experiencing everything that a subagent outputs as true. In this situation, there's no introspective awareness that would highlight those outputs as being just thoughts. "My friend is a horrible person", feels like a fact about the world.
2. *Partial blending:* you realize that the thoughts which you have might not be entirely true, but you still feel them emotionally and might end up acting accordingly. In this situation, there are moments of introspective awareness, but there are also enough of the original "unmarked" thoughts to also be affecting your other subagents. You feel hostile towards your friend, and realize that this may not be rationally warranted, but still end up talking in an angry tone and maybe saying things you shouldn't.
3. *Unblending:* all or nearly all of the thoughts coming from a subagent are being filtered through a mechanism that wraps them inside moments of introspective awareness, such as "this subagent is thinking that X". You know that you have a subagent which has this opinion, but none of the other subagents are treating it as a proven fact.

By training your mind to have more introspective moments of awareness, you will become capable of perceiving more and more mental objects as just that. A classic example would be all those mindfulness exercises where you stop identifying with the content of a thought, and see it as something separate from yourself. At more advanced levels, even mental objects which build up sensations such as those which make up the experience of a self may be seen as just constructed mental objects.

Subagents, akrasia, and coherence in humans

In my previous posts, I have been building up a model of mind as a collection of subagents with different goals, and no straightforward hierarchy. This then raises the question of how that collection of subagents can exhibit coherent behavior: after all, many ways of aggregating the preferences of a number of agents fail to create consistent preference orderings.

We can roughly describe coherence as the property that, if you become aware that there exists a more optimal strategy for achieving your goals than the one that you are currently executing, then you will switch to that better strategy. If an agent is not coherent in this way, then bad things are likely to happen to them.

Now, we all know that humans sometimes express incoherent behavior. But on the whole, people still do okay: the median person in a developed country still manages to survive until their body starts giving up on them, and typically also manages to have and raise some number of initially-helpless children until *they* are old enough to take care of themselves.

For a subagent theory of mind, we would like to have some explanation of when exactly the subagents manage to be collectively coherent (that is, change their behavior to some better one), and what are the situations in which they fail to do so. The conclusion of this post will be:

> We are capable of changing our behaviors on occasions when the mind-system as a whole puts sufficiently high probability on the new behavior being better, when the new behavior is not being blocked by a particular highly weighted subagent (such as an IFS-style protector) that puts high probability on it being bad, and when we have enough slack in our lives for any new behaviors to be evaluated in the first place. Akrasia is subagent disagreement about what to do.

Correcting your behavior as a default

There are many situations in which we exhibit incoherent behavior simply because we're not aware of it. For instance, suppose that I do my daily chores in a particular order, when doing them in some other order would save more time. If you point this out to me, I'm likely to just say "oh", and then adopt the better system.

Similarly, several of the experiments which get people to exhibit incoherent behavior rely on showing different groups of people different formulations of the same question, and then indicating that different framings of the same question get different answers from people. It doesn't work quite as well if you show the different formulations to the *same* people, because then many of them will realize that differing answers would be inconsistent.

But there are also situations in which someone realizes that they are behaving in a nonsensical way, yet will continue behaving in that way. Since people usually *can* change suboptimal behaviors, we need an explanation for why they sometimes *can't*.

Towers of protectors as a method for coherence

In my post about Internal Family Systems, I discussed a model of mind composed of several different kinds of subagents. One of them, the default planning subagent, is a module just trying to straightforwardly find the best thing to do and then execute that. On the other hand, *protector* subagents exist to prevent the system from getting into situations which were catastrophic before. If they think that the default planning subagent is doing something which seems dangerous, they will override it and do something else instead. (Previous versions of the IFS post called the default planning agent, “a reinforcement learning subagent”, but this was potentially misleading since several other subagents were reinforcement learning ones too, so I’ve changed the name.)

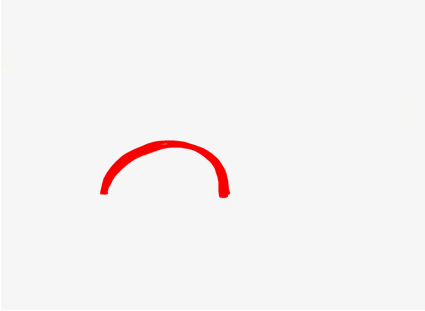
Thus, your behavior can still be coherent even if you *feel* that you are failing to act in a coherent way. You simply don’t realize that a protector is carrying out a routine intended to avoid dangerous outcomes - and this might actually be a very successful way of keeping you out of danger. Some subagents in your mind think that doing X would be a superior strategy, but the protector thinks that it would be a horrible idea - so from the point of view of the system as a whole, doing X is *not* a better strategy, so not switching to it is actually better.

On the other hand, it may also be the case that the protector’s behavior, while keeping you out of situations which the protector considers unacceptable, is causing other outcomes which are *also* unacceptable. The default planning subagent may realize this - but as already established, any protector can overrule it, so this doesn’t help.

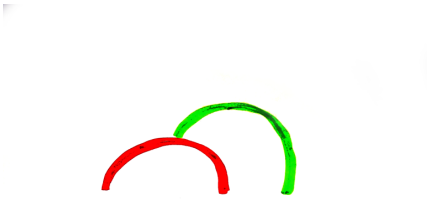
Evolution’s answer here seems to be spaghetti towers. The default planning subagent might *eventually* figure out the better strategy, which avoids both the thing that the protector is trying to block *and* the new bad outcome. But it could be dangerous to wait that long, especially since the default planning agent doesn't have direct access to the protector's goals. So for the same reasons why a separate protector subagent was created to avoid the *first* catastrophe, the mind will create or recruit a protector to avoid the *second* catastrophe - the one that the first protector keeps causing.

With permission, I’ll borrow the illustrations from eukaryote’s spaghetti tower post to illustrate this.

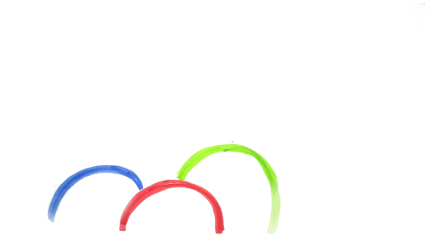
Example Eric grows up in an environment where he learns that disagreeing with other people is unsafe, and that he should always agree to do things that other people ask of him. So Eric develops a protector subagent running a pleasing, submissive behavior.



Unfortunately, while this tactic worked in Eric's childhood home, once he became an adult he starts saying "yes" to too many things, without leaving any time for his own needs. But saying "no" to anything still feels unsafe, so he can't just stop saying "yes". Instead he develops a protector which tries to keep him out of situations where people would ask him to do anything. This way, he doesn't need to say "no", and also won't get overwhelmed by all the things that he has promised to do. The two protectors together form a composite strategy.



While this helps, it still doesn't entirely solve the issue. After all, there are plenty of reasons that might push Eric into situations where someone would ask something of him. He still ends up agreeing to do lots of things, to the point of neglecting his own needs. Eventually, his brain creates another protector subagent. This one causes exhaustion and depression, so that he now has a socially-acceptable reason for being unable to do all the things that he has promised to do. He continues saying "yes" to things, but also keeps apologizing for being unable to do things that he (honestly) intended to do as promised, and eventually people realize that you probably shouldn't ask him to do anything that's really important to get done.



And while this kind of a process of stacking protector on top of a protector is not perfect, for most people it mostly works out okay. Almost everyone ends up having their unique set of minor neuroses and situations where they don't quite behave rationally, but as they learn to understand themselves better, their default planning subagent gets better at working around those issues. This might also make the

various protectors relax a bit, since the various threats are generally avoided and there isn't a need to keep avoiding them.

Gradually, as negative consequences to different behaviors become apparent, behavior gets adjusted - either by the default planning subagents or by spawning more protectors - and remains coherent overall.

But sometimes, especially for people in highly stressful environments where almost any mistake may get them punished, or when they end up in an environment that their old tower of protectors is no longer well-suited for (distributional shift), things don't go as well. In that situation, their minds may end up looking like this a hopelessly tangled web, where they have almost no flexibility. Something happens in their environment, which sets off one protector, which sets off another, which sets off another - leaving them with no room for flexibility or rational planning, but rather forcing them to act in a way which is almost bound to only make matters worse.

This kind of an outcome is obviously bad. So besides building spaghetti towers, the second strategy which the mind has evolved to employ for keeping its behavior coherent while piling up protectors, is the ability to re-process memories of past painful events.

As I discussed in my original IFS post, the mind has methods for bringing up the original memories which caused a protector to emerge, in order to re-analyze them. If ending up in some situation is actually no longer catastrophic (for instance, you are no longer in your childhood home where you get punished simply for not wanting to do something), then the protectors which were focused on avoiding that outcome can relax and take a less extreme role.

For this purpose, there seems to be a built-in tension. Exiles (the IFS term for subagents containing memories of past trauma) "want" to be healed and will do things like occasionally sending painful memories or feelings into consciousness so as to become the center of attention, especially if there is something about the current situation which resembles the past trauma. This also acts as what my IFS post called a fear model - something that warns of situations which resemble the past trauma enough to be considered dangerous in their own right. At the same time, protectors "want" to keep the exiles hidden and inactive, doing anything that they can for keeping them so. Various schools of therapy - IFS one of them - seek to tap into this existing tension so as to reveal the trauma, trace it back to its original source, and heal it.

Coherence and conditioned responses

Besides the presence of protectors, another possibility for why we might fail to change our behavior are strongly conditioned habits. Most human behavior involves automatic habits: behavioral routines which are triggered by some sort of a cue in the environment, and lead to or have once led to a reward. (Previous discussion; see also.)

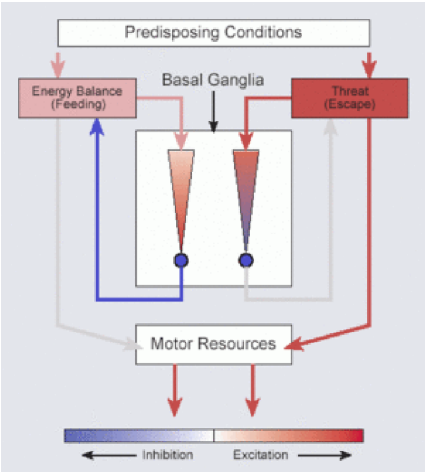
The problem with this is that people might end up with habits that they wouldn't want to have. For instance, I might develop a habit of checking social media on their phone when I'm bored, creating a loop of boredom (cue) -> looking at social media (behavior) -> seeing something interesting on social media (reward).

Reflecting on this behavior, I notice that back when I *didn't* do it, my mind was more free to wander when I was bored, generating motivation and ideas. I think that my old behavior was more valuable than my new one. But even so, my new behavior still delivers enough momentary satisfaction to keep reinforcing the habit.

Subjectively, this feels like an increasing compulsion to check my phone, which I try to resist since I know that long-term it would be a better idea to not be checking my phone all the time. But as the compulsion keeps growing stronger and stronger, eventually I give up and look at the phone anyway.

The exact neuroscience of what is happening at such a moment remains only partially understood (Simpson & Balsam 2016). However, we know that whenever different subsystems in the brain produce conflicting motor commands, that conflict needs to be resolved, with only one at a time being granted access to the "final common motor path". This is thought to happen in the basal ganglia, a part of the brain closely involved in action selection and connected to the global neuronal workspace.

One model (e.g. Redgrave 2007, McHaffie 2005) is that the basal ganglia receives inputs from many different brain systems; each of those systems can send different "bids" supporting or opposing a specific course of action to the basal ganglia. A bid submitted by one subsystem may, through looped connections going back from the basal ganglia, inhibit other subsystems, until one of the proposed actions becomes sufficiently dominant to be taken.



The above image from Redgrave 2007 has a conceptual image of the model, with two example subsystems shown. Suppose that you are eating at a restaurant in Jurassic Park when two velociraptors charge in through the window. Previously, your hunger system was submitting successful bids for the "let's keep eating" action, which then caused inhibitory impulses to the be sent to the threat system. This inhibition

prevented the threat system from making bids for silly things like jumping up from the table and running away in a panic. However, as your brain registers the new situation, the threat system gets significantly more strongly activated, sending a strong bid for the “let’s run away” action. As a result of the basal ganglia receiving that bid, an inhibitory impulse is routed from the basal ganglia to the subsystem which was previously submitting bids for the “let’s keep eating” actions. This makes the threat system’s bids even stronger relative to the (inhibited) eating system’s bids.

Soon the basal ganglia, which was previously inhibiting the threat subsystem’s access to the motor system while allowing the eating system access, withdraws that inhibition and starts inhibiting the eating system’s access instead. The result is that you jump up from your chair and begin to run away. Unfortunately, this is hopeless since the velociraptor is faster than you. A few moments later, the velociraptor’s basal ganglia gives the raptor’s “eating” subsystem access to the raptor’s motor system, letting it happily munch down its latest meal.

But let’s leave velociraptors behind and go back to our original example with the phone. Suppose that you have been trying to replace the habit of looking at your phone when bored, to instead smiling and directing your attention to pleasant sensations in your body, and then letting your mind wander.

Until the new habit establishes itself, the two habits will compete for control. Frequently, the old habit will be stronger, and you will just automatically check your phone without even remembering that you were supposed to do something different. For this reason, behavioral change programs may first spend several weeks just practicing *noticing* the situations in which you engage in the old habit. When you *do* notice what you are about to do, then more goal-directed subsystems may send bids towards the “smile and look for nice sensations” action. If this happens and you pay attention to your experience, you may notice that long-term it actually feels more pleasant than looking at the phone, reinforcing the new habit until it becomes prevalent.

To put this in terms of the subagent model, we might drastically simplify things by saying that the neural pattern corresponding to the old habit is a subagent reacting to a specific sensation (boredom) in the consciousness workspace: its reaction is to generate an intention to look at the phone. At first, you might train the subagent responsible for monitoring the contents of your consciousness, to output moments of introspective awareness highlighting when that intention appears. That introspective awareness helps alert a goal-directed subagent to try to trigger the new habit instead. Gradually, a neural circuit corresponding to the new habit gets trained up, which starts sending its own bids when it detects boredom. Over time, reinforcement learning in the basal ganglia starts giving that subagent’s bids more weight relative to the old habit’s, until it no longer needs the goal-directed subagent’s support in order to win.

Now this model helps incorporate things like the role of having a vivid emotional motivation, a sense of hope, or psyching yourself up when trying to achieve habit change. Doing things like imagining an outcome that you wish the habit to lead to, may activate additional subsystems which care about those kinds of outcomes, causing them to submit additional bids in favor of the new habit. The extent to which you succeed at doing so, depends on the extent to which your mind-system considers it *plausible* that the new habit leads to the new outcome. For instance, if you imagine your exercise habit making you strong and healthy, then subagents which care about

strength and health might activate to the extent that you believe this to be a likely outcome, sending bids in favor of the exercise action.

On this view, one way for the mind to maintain coherence and readjust its behaviors, is its ability to re-evaluate old habits in light of which subsystems get activated when reflecting on the possible consequences of new habits. An old habit having been strongly reinforced reflects that a great deal of evidence has accumulated in favor of it being beneficial, but the behavior in question can still be overridden if enough influential subsystems weigh in with their evaluation that a new behavior would be more beneficial in expectation.

Some subsystems having concerns (e.g. immediate survival) which are ranked more highly than others (e.g. creative exploration) means that the decision-making process ends up carrying out an implicit expected utility calculation. The strengths of bids submitted by different systems do not just reflect the probability that those subsystems put on an action being the most beneficial. There are also different mechanisms giving the bids from different subsystems varying amounts of weight, depending on how important the concerns represented by that subsystem happen to be in that situation. This ends up doing something like weighting the probabilities by utility, with the kinds of utility calculations that are chosen by evolution and culture in a way to maximize genetic fitness on average. Protectors, of course, are subsystems whose bids are weighted particularly strongly, since the system puts high utility on avoiding the kinds of outcomes they are trying to avoid.

The original question which motivated this section was: why are we sometimes incapable of adopting a new habit or abandoning an old one, despite knowing that to be a good idea? And the answer is: because we *don't* know that such a change would be a good idea. Rather, *some* subsystems *think* that it would be a good idea, but other subsystems remain unconvinced. Thus the system's overall judgment is that the old behavior should be maintained.

Interlude: Minsky on mutually bidding subagents

> I was trying to concentrate on a certain problem but was getting bored and sleepy. Then I imagined that one of my competitors, Professor Challenger, was about to solve the same problem. An angry wish to frustrate Challenger then kept me working on the problem for a while. The strange thing was, this problem was not of the sort that ever interested Challenger.
>
> What makes us use such roundabout techniques to influence ourselves? Why be so indirect, inventing misrepresentations, fantasies, and outright lies? Why can't we simply tell ourselves to do the things we want to do? [...]
>
> Apparently, what happened was that my agency for Work exploited Anger to stop Sleep. But why should Work use such a devious trick?
>
> To see why we have to be so indirect, consider some alternatives. If Work could simply turn off Sleep, we'd quickly wear our bodies out. If Work could simply switch Anger on, we'd be fighting all the time. Directness is too dangerous. We'd die.

> Extinction would be swift for a species that could simply switch off hunger or pain. Instead, there must be checks and balances. We'd never get through one full day if any agency could seize and hold control over all the rest. This must be why our agencies, in order to exploit each other's skills, have to discover such roundabout pathways. All direct connections must have been removed in the course of our evolution.
>
> This must be one reason why we use fantasies: to provide the missing paths. You may not be able to make yourself angry simply by deciding to be angry, but you can still imagine objects or situations that make you angry. In the scenario about Professor Challenger, my agency Work exploited a particular memory to arouse my Anger's tendency to counter Sleep. This is typical of the tricks we use for self-control.
>
> Most of our self-control methods proceed unconsciously, but we sometimes resort to conscious schemes in which we offer rewards to ourselves: "If I can get this project done, I'll have more time for other things." However, it is not such a simple thing to be able to bribe yourself. To do it successfully, you have to discover which mental incentives will actually work on yourself. This means that you - or rather, your agencies - have to learn something about one another's dispositions. In this respect the schemes we use to influence ourselves don't seem to differ much from those we use to exploit other people - and, similarly, they often fail. When we try to induce ourselves to work by offering ourselves rewards, we don't always keep our bargains; we then proceed to raise the price or even deceive ourselves, much as one person may try to conceal an unattractive bargain from another person.
>
> Human self-control is no simple skill, but an ever-growing world of expertise that reaches into everything we do. Why is it that, in the end, so few of our self-incentive tricks work well? Because, as we have seen, directness is too dangerous. If self-control were easy to obtain, we'd end up accomplishing nothing at all.

-- Marvin Minsky, *The Society of Mind*

Akrasia is subagent disagreement

You might feel that the above discussion doesn't still entirely resolve the original question. After all, sometimes we *do* manage to change even strongly conditioned habits pretty quickly. Why is it sometimes hard and sometimes easier?

Redgrave et al. (2010) discuss two modes of behavioral control: goal-directed versus habitual. Goal-directed control is a relatively slow mode of decision-making, where "action selection is determined primarily by the relative utility of predicted outcomes", whereas habitual control involves more directly conditioned stimulus-response behavior. Which kind of subsystem is in control is complicated, and depends on a variety of factors (the following quote has been edited to remove footnotes to references; see the original for those):

> Experimentally, several factors have been shown to determine whether the agent (animal or human) operates in goal-directed or habitual mode. The first is over-training: here, initial control is largely goal-directed, but with consistent and repeated training there is a gradual shift to stimulus–response, habitual control. Once habits are established, habitual responding tends to dominate, especially in stressful situations in which quick reactions are required. The second related

> factor is task predictability: in the example of driving, talking on a mobile phone is fine so long as everything proceeds predictably. However, if something unexpected occurs, such as someone stepping out into the road, there is an immediate switch from habitual to goal-directed control. Making this switch takes time and this is one of the reasons why several countries have banned the use of mobile phones while driving. The third factor is the type of reinforcement schedule: here, fixed-ratio schedules promote goal-directed control as the outcome is contingent on responding (for example, a food pellet is delivered after every n responses). By contrast, interval schedules (for example, schedules in which the first response following a specified period is rewarded) facilitate habitual responding because contingencies between action and outcome are variable. Finally, stress, often in the form of urgency, has a powerful influence over which mode of control is used. The fast, low computational requirements of stimulus–response processing ensure that habitual control predominates when circumstances demand rapid reactions (for example, pulling the wrong way in an emergency when driving on the opposite side of the road). Chronic stress also favours stimulus–response, habitual control. For example, rats exposed to chronic stress become, in terms of their behavioural responses, insensitive to changes in outcome value and resistant to changes in action–outcome contingency. [...]
>
> Although these factors can be seen as promoting one form of instrumental control over the other, real-world tasks often have multiple components that must be performed simultaneously or in rapid sequences. Taking again the example of driving, a driver is required to continue steering while changing gear or braking. During the first few driving lessons, when steering is not yet under automatic stimulus–response control, things can go horribly awry when the new driver attempts to change gears. By contrast, an experienced (that is, ‘over-trained’) driver can steer, brake and change gear automatically, while holding a conversation, with only fleeting contributions from the goal-directed control system. This suggests that many skills can be deconstructed into sequenced combinations of both goal-directed and habitual control working in concert. [...]
>
> Nevertheless, a fundamental problem remains: at any point in time, which mode should be allowed to control which component of a task? Daw et al. have used a computational approach to address this problem. Their analysis was based on the recognition that goal-directed responding is flexible but slow and carries comparatively high computational costs as opposed to the fast but inflexible habitual mode. They proposed a model in which the relative uncertainty of predictions made by each control system is tracked. In any situation, the control system with the most accurate predictions comes to direct behavioural output.

Note those last sentences: besides the subsystems making their own predictions, there might also be a meta-learning system keeping track of which other subsystems tend to make the most accurate predictions in each situation, giving extra weight to the bids of the subsystem which has tended to perform the best in that situation. We’ll come back to that in future posts.

This seems compatible with my experience in that, I feel like it’s possible for me to change even entrenched habits relatively quickly - *assuming that the new habit really is unambiguously better*. In that case, while I might forget and lapse to the old habit a few times, there’s still a rapid feedback loop which quickly indicates that the goal-directed system is simply *right* about the new habit being better.

Or, the behavior in question might be sufficiently complex and I might be sufficiently inexperienced at it, that the goal-directed (default planning) subagent has always mostly remained in control of it. In that case change is again easy, since there is no strong habitual pattern to override.

In contrast, in cases where it's hard to establish a new behavior, there tends to be some kind of genuine uncertainty:

- The benefits of the old behavior have been validated in the form of direct experience (e.g. unhealthy food that tastes good, has in fact tasted good each time), whereas the benefits of the new behavior come from a less trusted information source which is harder to validate (e.g. I've read scientific studies about the long-term health risks of this food).
- Immediate vs. long-term rewards: the more remote the rewards, the larger the risk that they will for some reason never materialize.
- High vs. low variance: sometimes when I'm bored, looking at my phone produces genuinely *better* results than letting my thoughts wander. E.g. I might see an interesting article or discussion, which gives me novel ideas or insights that I would not otherwise have had. Basically looking at my phone usually produces worse results than not looking at it - but sometimes it also produces much better ones than the alternative.
- Situational variables affecting the value of the behaviors: looking at my phone can be a way to escape uncomfortable thoughts or sensations, for which purpose it's often excellent. This then also tends to reinforce the behavior of looking at the phone when I'm in the same situation otherwise, but *without* uncomfortable sensations that I'd like to escape.

When there is significant uncertainty, the brain seems to fall back to those responses which have worked the best in the past - which seems like a reasonable approach, given that intelligence involves hitting tiny targets in a huge search space, so most novel responses are likely to be wrong.

As the above excerpt noted, the tendency to fall back to old habits is exacerbated during times of stress. The authors attribute it to the need to act quickly in stressful situations, which seems correct - but I would also emphasize the fact that negative emotions in general tend to be signs of something being wrong. E.g. Eldar et al. (2016) note that positive or negative moods tend to be related to whether things are going better or worse than expected, and suggest that mood is a *computational representation of momentum*, acting as a sort of global update to our reward expectations.

For instance, if an animal finds more fruit than it had been expecting, that may indicate that spring is coming. A shift to a good mood and being "irrationally optimistic" about finding fruit even in places where the animal hasn't seen fruit in a while, may actually serve as a rational pre-emptive update to its expectations. In a similar way, things going less well than expected may be a sign of some more general problem, necessitating fewer exploratory behaviors and less risk-taking, so falling back into behaviors for which there is a higher certainty of them working out.

So to repeat the summary that I had in the beginning: we are capable of changing our behaviors on occasions when the mind-system as a whole puts sufficiently high probability on the new behavior being better, when the new behavior is not being blocked by a particular highly weighted subagent (such as an IFS protector whose bids get a lot of weight) that puts high probability on it being bad, and when we have

enough slack in our lives for any new behaviors to be evaluated in the first place. Akrasia is subagent disagreement about what to do.

Integrating disagreeing subagents

In my previous post, I suggested that akrasia involves subagent disagreement - or in other words, different parts of the brain having differing ideas on what the best course of action is. The existence of such conflicts raises the question, how does one resolve them?

In this post I will discuss various techniques which could be interpreted as ways of resolving subagents disagreements, as well as some of the reasons for why this doesn't always happen.

A word on interpreting "subagents"

The frame that I've had so far is that of the brain being composed of different subagents with conflicting beliefs. On the other hand, one could argue that the subagent interpretation isn't strictly necessary for many of the examples that I bring up in this post. One could just as well view my examples as talking about a single agent with conflicting beliefs.

The distinction between these two frames isn't always entirely clear. In "Complex Behavior from Simple (Sub)Agents", mordinamael presents a toy model where an agent has different goals. Moving to different locations will satisfy the different goals to a varying extent. The agent will generate a list of possible moves and picks the move which will bring some goal the closest to being satisfied.

Is this a unified agent, or one made up of several subagents?

One could argue for either interpretation. On the other hand, mordinamael's post frames the goals as subagents, and they are in a sense competing with each other. On the other hand, the subagents arguably don't make the final decision themselves: they just report expected outcomes, and then a central mechanism picks a move based on their reports.

This resembles the neuroscience model I discussed in my last post, where different subsystems in the brain submit various action "bids" to the basal ganglia. Various mechanisms then pick a winning bid based on various criteria - such as how relevant the subsystem's concerns are for the current situation, and how accurate the different subsystems have historically been in their predictions.

Likewise, in extending the model from *Consciousness and the Brain* for my toy version of the Internal Family Systems model, I postulated a system where various subagents vote for different objects to become the content of consciousness. In that model, the winner was determined by a system which adjusted the vote weights of the different subagents based on various factors.

So, subagents, or just an agent with different goals?

Here I would draw an analogy to parliamentary decision-making. In a sense, a parliament as a whole is an agent. Various members of parliament cast their votes, with "the voting system" then "making the final choice" based on the votes that have been cast. That reflects the overall judgment of the parliament as a whole. On the

other hand, for understanding and predicting how the parliament will actually vote in different situations, it is important to model how the individual MPs influence and broker deals with each other.

Likewise, the subagent frame seems most useful when a person's goals interact in such a way that applying the intentional stance - thinking in terms of the beliefs and goals of the individual subagents - is useful for modeling the overall interactions of the subagents.

For example, in my toy Internal Family Systems model, I noted that reinforcement learning subagents might end up forming something like alliances. Suppose that a robot has a choice between making cookies, poking its finger at a hot stove, or daydreaming. It has three subagents: "cook" wants the robot to make cookies, "masochist" wants to poke the robot's finger at the stove, and "safety" wants the robot to *not* poke its finger at the stove.

By default, "safety" is indifferent between "make cookies" and "daydream", and might cast its votes at random. But when it votes for "make cookies", then that tends to avert "poke at stove" more reliably than voting for "daydream" does, as "make cookies" is also being voted for by "cook". Thus its tendency to vote for "make cookies" in this situation gets reinforced.

We can now apply the intentional stance to this situation, and say that "safety" has "formed an alliance" with "cook", as it correctly "believes" that this will avert masochistic actions. If the subagents are also aware of each other and can predict each other's actions, then the intentional stance gets even more useful.

Of course, we could just as well apply the purely mechanistic explanation and end up with the same predictions. But the intentional explanation often seems easier for humans to reason with, and helps highlight salient considerations.

Integrating beliefs, naturally or with techniques

In any case, regardless of whether we are talking about subagents with conflicting beliefs or just conflicting goals, it still seems like many of our problems arise from *some* kind of internal disagreement. I will use the term "integration" for anything that acts to resolve such conflicts, and discuss a few examples of things which can be usefully thought of as integration.

In these examples, I am again going to rely on the basic observation from *Consciousness and the Brain*: that when some subsystem in the brain manages to elevate a mental object into the content of consciousness, multiple subsystems will synchronize their processing around that object. Assuming that the conditions are right, this will allow for the integration of otherwise conflicting beliefs or behaviors.

Why do we need to explicitly integrate beliefs, rather than this happening automatically? One answer is that trying to integrate all beliefs would be infeasible; as CronoDAS notes:

> GEB has a section on this.

> In order to *not* compartmentalize, you need to test if your beliefs are all consistent with each other. If your beliefs are all statements in propositional logic, consistency checking becomes the Boolean Satisfiability Problem, which is NP-complete. If your beliefs are statements in predicate logic, then consistency checking becomes PSPACE-complete, which is even worse than NP-complete.

Rather than try to constantly integrate every possible belief and behavior, the brain will rather try to integrate beliefs at times when it notices contradictions. Of course, sometimes we *do* realize that there are contradictions, but still don't automatically integrate the subagents. Then we can use various techniques for making integration more effective. How come integration isn't more automatic?

One reason is that integration requires the right conditions, and while the brain has mechanisms for getting those conditions right, integration is still a nontrivial skill. As an analogy, most children learn the basics of talking and running on their own, but they can still explicitly study rhetoric or running techniques to boost their native competencies far above their starting level. Likewise, everyone natively does *some* integration on their own, but people can also use explicit techniques which make them much better at it.

Resisting belief integration

Lack of skill isn't the full answer for why we don't always automatically update, however. Sometimes it seems as if the mind actively resists updating.

One of the issues that commonly comes up in Internal Family Systems therapy is that parts of the mind want to keep some old belief frozen, because if it were known, it would change the person's behavior in an undesired way. For example, if someone believes that they have a good reason not to abandon their friend, then a part of the mind which values not abandoning the friend in question might resist having this belief re-revaluated. The part may then need to be convinced that knowing the truth only leaves opens the *option* of abandoning the friend, it doesn't *compel* it.

Note that this isn't *necessarily* true. If there are other subagents which sufficiently strongly hold the opinion that the friend should be abandoned, and the subagent-which-values-the-friend is only managing to prevent that by hanging on to a specific belief, then readjusting that belief *might* remove the only constraint which was preventing the anti-friend coalition from dumping the friend. Thus from the point of view of the subagent which is resisting the belief update, the update *would* compel an abandonment of the friend. In such a situation, additional internal work may be necessary before the subagent will agree to let the belief revision proceed.

More generally, subagents may be incentivized to resist belief updating for at least three different reasons (this list is not intended to be exhaustive):

1. The subagent is trying to pursue or maintain a goal, and predicts that revising some particular belief would make the person less motivated to pursue or maintain the goal.
2. The subagent is trying to safeguard the person's social standing, and predicts that not understanding or integrating something will be safer, give the person an advantage in negotiation, or be otherwise socially beneficial. For instance, different subagents holding conflicting beliefs allows a person to verbally believe

in one thing while still not acting accordingly - even actively changing their verbal model so as to avoid falsifying the invisible dragon in the garage.
3. Evaluating a belief would require activating a memory of a traumatic event that the belief is related to, and the subagent is trying to keep that memory suppressed as part of an exile-protector dynamic.

Here's an alternate way of looking at the issue, which doesn't use the subagent frame. So far I have been mostly talking about integrating beliefs rather than goals, but humans don't seem to have a clear value/belief distinction. As Stuart Armstrong discusses in his mAIry's room article, for humans simply receiving sensory information often also rewires some of their values. Now, Mark Lippman suggests that trying to optimize a complicated network of beliefs and goals means that furthering one goal may hurt other goals, so the system needs to have checks in place to ensure that one goal is not pursued in a way which disproportionately harms the achievement of other goals.

For example, most people wouldn't want to spend the rest of their lives doing nothing but shooting up heroin, *even if* they knew for certain that this maximized the achievement of their "experience pleasure" goal. If someone offered them the chance to experience just *how* pleasurable heroin felt like - giving them more accurate emotion-level predictions of the experience - they might quite reasonably refuse, as they feared that making this update might make them more inclined to take heroin. Eliezer once noted that if someone offered him a pill which simulated the joy of scientific discovery, he would make sure never to take it.

Suppose that a system has a network of beliefs and goals and it does something like predicting how various actions and their effects - not only their effects on the external world, but on the belief/goal network itself - might influence its goal achievement. If it resists actions which reduce the probability of achieving its current goals, then this might produce dynamics which look like subagents trying to achieve their goals at the expense of the other subagents.

For instance, Eliezer's refusal to take the pill might be framed as a subagent valuing scientific discovery trying to block a subagent valuing happiness from implementing an action which would make the happiness subagent's bids for motor system access stronger. Alternatively, it might be framed as the overall system putting value on actually making scientific discoveries, and refusing to self-modify in a way which it predicted would hurt this goal. (You might note that this has some interesting similarities to things like the Cake or Death problem in AI alignment.)

In any case, integration is not always straightforward. Even if the system *does* detect a conflict between its subagents, it may have a reason to avoid doing so.

Having reviewed some potential barriers for integration, let us move on to different ways in which conflicts *can* be detected and integrated.

Ways to integrate conflicting subagents

Cognitive Behavioral Therapy

Scott Alexander has an old post where he quotes this excerpt from the cognitive behavioral therapy book *When Panic Attacks*:

> I asked Walter how he was thinking and feeling about the breakup with Paul. What was he telling himself? He said "I feel incredibly guilty and ashamed, and it seems like it must have been my fault. Maybe I wasn't skillful enough, attractive enough, or dynamic enough. Maybe I wasn't there for him emotionally. I feel like I must have screwed up. Sometimes I feel like a total fraud. Here I am, a marriage and family therapist, and my own relationship didn't even work out. I feel like a loser. A really, really big loser." [...]
>
> I thought the Double Standard Technique might help because Walter seemed to be a warm and compassionate individual. I asked what he'd say to a dear friend who'd been rejected by someone he'd been living with for eight years. I said "Would you tell him that there's something wrong with him, that he screwed up his life and flushed it down the toilet for good?"
>
> Walter looked shocked and said he'd never say something like that to a friend. I suggested we try a role-playing exercise so that he could tell me what he would say to a friend who was in the same predicament [...]
>
> **Therapist (role-playing patient's friend):** Walter, there's another angle I haven't told you about. What you don't understand is that I'm impossible to live with and be in a relationship with. That's the real reason I feel so bad, and that's why I'll be alone for the rest of my life.
>
> **Patient (role-playing as if therapist is his friend who just had a bad breakup):** Gosh, I'm surprised to hear you say that, because I've known you for a long time and never felt that way about you. In fact, you've always been warm and open, and a loyal friend. How in the world did you come to the conclusion that you were impossible to be in a relationship with?
>
> **Therapist (continuing role-play):** Well, my relationship with [my boyfriend] fell apart. Doesn't that prove I'm impossible to be in a relationship with?
>
> **Patient (continuing role-play):** In all honesty, what you're saying doesn't make a lot of sense. In the first place, your boyfriend was also involved in the relationship. It takes two to tango. And in the second place, you were involved in a reasonably successful relationship with him for eight years. So how can you claim that you're impossible to live with?
>
> **Therapist (continuing role-play:)** Let me make sure I've got this right. You're saying that I was in a reasonably successful relationship for eight years, so it doesn't make much sense to say that I'm impossible to live with or impossible to be in a relationship with?
>
> **Patient (continuing-role-play:)** You've got it. Crystal clear.
>
> At that point, Walter's face lit up, as if a lightbulb had suddenly turned on in his brain, and we both started laughing. His negative thoughts suddenly seemed absurd to him, and there was an immediate shift in his mood…after Walter put the lie to his negative thoughts, I asked him to rate how he was feeling again. His feeling of sadness fell all the way from 80% to 20%. His feelings of guilt, shame, and anxiety fell all the way to 10%, and his feelings of hopelessness dropped to

> 5%. The feelings of loneliness, embarrassment, frustration, and anger disappeared completely.

At the time, Scott expressed confusion about how just telling someone that their beliefs aren't rational, would be enough to transform the beliefs. But that wasn't really what happened. Walter was asked whether he'd say something harsh to a friend, and he said no, but that alone wasn't enough to improve his condition. What did help was putting him in a position where he had to really think through the arguments for why this is irrational in order to convince his friend, and then, after having formulated the arguments once himself, get convinced by them himself.

In terms of our framework, we might say that a part of Walter's mind contained a model which output a harsh judgment of himself, while another part contained a model which would output a much less harsher judgment of someone else who was in otherwise identical circumstances. Just bringing up the existence of this contradiction wasn't enough to change it: it caused the contradiction to be *noticed*, but didn't activate the relevant models extensively enough for their contents to be reprocessed.

But when Walter had to role-play a situation where he thought of himself as *actually* talking with a depressed friend, that required him to more fully activate the non-judgmental model and apply it to the relevant situation. This caused him to blend with the model, taking its perspective as the truth. When that perspective was then propagated to the self-critical model, the easiest way for the mind to resolve the conflict was simply to alter the model producing the self-critical thoughts.

Note that this kind of a result wasn't *guaranteed* to happen: Walter's self-critical model might have had a reason for why these cases were actually different, and pointing out that reason would have been another way for the contradiction to be resolved. In the example case, however, it seemed to work.

Mental contrasting

Another example of activating two conflicting mental models and forcing an update that way comes from the psychologist Gabriele Oettingen's book Rethinking Positive Thinking. Oettingen is a psychologist who has studied combining a mental imagery technique known as "mental contrasting" with trigger-action planning.

It is worth noting that this book has come under some heavy criticism and may be based on cherry-picked studies. However, in the book this particular example is just presented as an anecdote without even trying to cite any particular studies in its support. I present it because I've personally found the technique to be useful, and because it feels like a nice concise explanation of the kind of integration that often works:

> Try this exercise for yourself. Think about a fear you have about the future that is vexing you quite a bit and that you know is unjustified. Summarize your fear in three to four words. For instance, suppose you're a father who has gotten divorced and you share custody with your ex-wife, who has gotten remarried. For the sake of your daughter's happiness, you want to become friendly with her stepfather, but you find yourself stymied by your own emotions. Your fear might be "My daughter will become less attached to me and more attached to her stepfather." Now go on to imagine the worst possible outcome. In this case, it might be "I feel distanced from my daughter. When I see her she ignores me, but she eagerly

> spends time with her stepfather." Okay, now think of the positive reality that stands in the way of this fear coming true. What in your actual life suggests that your fear won't really come to pass? What's the single key element? In this case, it might be "The fact that my daughter is extremely attached to me and loves me, and it's obvious to anyone around us." Close your eyes and elaborate on this reality.
>
> Now take a step back. Did the exercise help? I think you'll find that by being reminded of the positive reality standing in the way, you will be less transfixed by the anxious fantasy. When I conducted this kind of mental contrasting with people in Germany, they reported that the experience was soothing, akin to taking a warm bath or getting a massage. "It just made me feel so much calmer and more secure," one woman told me. "I sense that I am more grounded and focused."
>
> Mental contrasting can produce results with both unjustified fears as well as overblown fears rooted in a kernel of truth. If as a child you suffered through a couple of painful visits to the dentist, you might today fear going to get a filling replaced, and this fear might become so terrorizing that you put off taking care of your dental needs until you just cannot avoid it. Mental contrasting will help you in this case to approach the task of going to the dentist. But if your fear is justified, then mental contrasting will confirm this, since there is nothing preventing your fear from coming true. The exercise will then help you to take preventive measures or avoid the impending danger altogether.

As in the CBT example, first one mental model (the one predicting losing the daughter's love) is activated and intentionally blended with, after which an opposing one is, forcing integration. And as in Walter's example, this is not *guaranteed* to resolve the conflict in a more reassuring way: the mind can also resolve the conflict by determining that actually the fear *is* justified.

Internal Double Crux / Internal Family Systems

On some occasions a single round of mental contrasting, or the Walter CBT technique, might be enough. In that case, there were two disagreeing models, and bringing the disagreement into consciousness was enough to reject the other one entirely. But it is not always so clear-cut; sometimes there are subagents which disagree, and both of them actually have some valid points.

For instance, someone might have a subagent which wants the person to do socially risky things, and another subagent which wants to play things safe. Neither is unambiguously wrong: on the other hand, some things *are* so risky that you should never try to do them. On the other hand, *never* doing anything which others might disapprove of is not going to lead to a particularly happy life, either.

In that case, one may need to actively facilitate a *dialogue* between the subagents, such as in the CFAR technique of Internal Double Crux (description, discussion and example, example as applied to dieting), iterating it for several rounds until both subagents come to agreement. The CBT and mental contrasting examples above might be considered special cases of an IDC session, where agreement was reached within a single round of discussion.

More broadly, IDC itself can be considered a special case of applying Internal Family Systems, which includes facilitating conversations between mutually opposing subagents as one of its techniques.

Self-concept editing

In the summer of 2017, I found Steve Andreas's book *Transforming Your Self*, and applied its techniques to fixing a number of issues in my self-concepts which had contributed to my depression and anxiety. Effects from this work which have lasted include no longer having generalized feelings of shame, no longer needing constant validation to avoid such feelings of shame, no longer being motivated by a desire to prove to myself that I'm a good person, and no longer having obsessive escapist fantasies, among other things.

I wrote an article at the time that described the work. The model in *Transforming Your Self* is that I might have a self-concept such as "I am kind". That self-concept is made up of memories of times when I either was kind (*examples* of the concept), or times when I was not (*counterexamples*). In a healthy self-concept, both examples and counterexamples are integrated together: you might have memories of how you are kind in general, but also memories of not being very kind at times when you were e.g. under a lot of stress. This allows you to both know your general tendency, as well as letting you prepare for situations where you know that you won't be very kind.

The book's model also holds that sometimes a person's counterexamples might be split off from their examples. This leads to an unstable self-concept: either your subconscious attention is focused on the examples and totally ignores the counterexamples, in which case you feel good and kind, or it swings to the counterexamples and totally ignores the examples, in which case you feel like a terrible horrible person with no redeeming qualities. You need a constant stream of external validation and evidence in order to keep your attention anchored on the examples; the moment it ceases, your attention risks swinging to the counterexamples again.

While I didn't have the concept back then, what I did could also be seen as integrating true but disagreeing perspectives between two subagents. There was one subagent which held memories of times when I had acted in what it thought of as a bad way, and was using feelings of shame to motivate me to make up for those actions. Another subagent was then reacting to it by making me do more and more things which I could use to prove to myself and others that I was indeed a good person. (This description roughly follows the framing and conceptualization of self-esteem and guilt/shame in the IFS book *Freedom from your Inner Critic*.)

Under the sociometer theory of self-esteem, self-esteem is an internal evaluation of one's worth as a partner to others. With this kind of an interpretation, it makes sense to have subagents acting in the ways that I described: if you have done things that your social group would judge you for, then it becomes important to do things which prove your worth and make them forgive you.

This then becomes a special case of an IFS exile/protector dynamic. Under that formulation, the splitting of the counterexamples and the lack of updating actually serves a purpose. The subagent holding the memories of doing shameful things doesn't want to stop generating the feelings of shame until it has received sufficient evidence that the "prove your worth" behavior has actually become unnecessary.

One of the techniques from *Transforming Your Self* that I used to fix my self-concept was integrating the examples by adding qualifiers to the counterexamples: “when I was a child, and my executive control wasn’t as developed, I didn’t always act as kindly as I could have”. Under the belief framing, this allowed my memories to be integrated in a way which showed that my selfishness as a child was no longer evidence of me being horrible in general. Under the subagent framing, this communicated to the shame-generating subagent that the things that I did as a child would no longer be held against me, and that it was safe to relax.

Another technique mentioned in *Transforming Your Self*, which I did not personally need to use, was translating the concerns of subagents into a common language. For instance, someone’s positive self-concept examples might be in the form of mental images, with their negative counterexamples being in the form of a voice which reminds them of their failures. In that case, they might translate the inner speech into mental imagery by visualizing what the voice is saying, turning both the examples and counterexamples into mental images that can then be combined. This brings us to…

Translating into a common language

Eliezer presents an example of two different framings eliciting conflicting behavior in his “Circular Altruism” post:

> Suppose that a disease, or a monster, or a war, or something, is killing people. And suppose you only have enough resources to implement one of the following two options:
>
> 1. Save 400 lives, with certainty.
>
> 2. Save 500 lives, with 90% probability; save no lives, 10% probability.
>
> Most people choose option 1. […] If you present the options this way:
>
> 1. 100 people die, with certainty.
>
> 2. 90% chance no one dies; 10% chance 500 people die.
>
> Then a majority choose option 2. *Even though it's the same gamble*. You see, just as a *certainty* of saving 400 lives seems to *feel* so much more comfortable than an unsure gain, so too, a certain loss *feels* worse than an uncertain one.

In my previous post, I presented a model where subagents which are most strongly activated by the situation are the ones that get access to the motor system. If you are hungry and have a meal in front of you, the possibility of eating is the most salient and valuable feature of the situation. As a result, subagents which want you to eat get the most decision-making power. On the other hand, if this is a restaurant in Jurassic Park and a velociraptor suddenly charges through the window, then the dangerous aspects of the situation become most salient. That lets the subagents which want you to flee to get the most decision-making power.

Eliezer’s explanation of the saving lives dilemma is that in the first framing, the certainty of saving 400 lives is salient, whereas in the second explanation the certainty of losing 100 lives is salient. We can interpret this in similar terms as the “eat or run” dilemma: the action which gets chosen, depends on which features are

the most salient and how those features activate different subagents (or how those features highlight different priorities, if we are not using the subagent frame).

Suppose that you are someone who was tempted to choose option 1 when you were presented with the first framing, and option 2 when you were presented with the second framing. It is now pointed out to you that these are actually exactly equivalent. You realize that it would be inconsistent to prefer one option over the other just depending on the framing. Furthermore, and maybe even more crucially, realizing this makes *both* the "certainty of saving 400 lives" *and* "certainty of losing 100 lives" features become equally salient. That puts the relevant subagents (priorities) on more equal terms, as they are both activated to the same extent.

What happens next depends on what the relative strengths of those subagents (priorities) are otherwise, and whether you happen to know about expected value. Maybe you consider the situation and one of the two subagents (priorities) happens to be stronger, so you decide to consistently save 400 or consistently lose 100 lives in both situations. Alternatively, the conflicting priorities may be resolved by introducing the rule that "when detecting this kind of a dilemma, convert both options into an expected value of lives saved, and pick the option with the higher value".

By converting the options to an expected value, one can get a basis by which two otherwise equal options can be evaluated and chosen between. Another way of looking at it is that this is bringing in a third kind of consideration/subagent (knowledge of the decision-theoretically optimal decision) in order to resolve the tie.

Urge propagation

CFAR and Harvard Effective Altruism is a video of a lecture given by former CFAR instructors Valentine Smith and Duncan Sabien. In Valentine's part of the lecture, he describes a few motivational techniques which work by mentally reframing the contents of an experience.

The first example involves having a $50 parking ticket, which - unless paid within 30 days - will accrue an additional $90 penalty. This kind of a thing tends to feel ughy to deal with, causing an inclination to avoid thinking about it - while also being aware of the need to do something about it. Something along the lines of two different subagents which are both trying to avoid pain using opposite methods - one by not thinking about unpleasant things, another by doing things which stop future unpleasantness.

Val's suggested approach involves noting that if you instead had a cheque for $90, which would expire in 30 days, then that would *not* cause such a disinclination. Rather, it would feel actively pleasant to cash it in and get the money.

The structure of the "parking ticket" and "cheque" scenarios are equivalent, in that both cases you can take an action to be $90 better off after 30 days. If you notice this, then it may be possible for you to re-interpret the action of paying off the parking ticket as something that *gains you money*, maybe by something like literally looking at it and imagining it as a cheque that you can cash in, until cashing it in starts feeling *actively pleasant*.

Val emphasizes that this is not just an arbitrary motivational hack: it's important that your reframe is *actually bringing in real facts from the world*. You don't want to just

imagine the parking ticket as a ticking time bomb, or as something else which it actually isn't. Rather, you want to do a reframe which integrates both perspectives, while also highlighting the features which will help fix the conflict.

One description of what happens here would be that once the pain-avoiding subagent notices that paying the parking ticket can feel like a net gain, and that it being a net gain is actually describing a real fact about the world, then it can drop its objection and you can proceed to take actions. The other way of looking at it is that like with expected value, you are introducing a common currency - the future impact on your finances - which allows the salient features from both subagents' perspectives to be integrated and then resolved.

Val's second example involves a case where he found himself not doing push-ups like he had intended to. When examining the reason why not, he noticed that the push-ups felt physically unpleasant: they involved sweating, panting, and a burning sensation, and this caused a feeling of aversion.

Part of how he solved the issue was by realizing that his original goal for getting exercise was to live longer and be in better health. The unpleasant physical sensations were a sign that *he was pushing his body hard enough that the push-ups would actually be useful for this goal*. He could then create a mental connection between the sensations and his goal of being healthier and living longer: the sensations started feeling like *something positive*, since they were an indication of progress.

Besides being an example of creating a common representation between the subagents, this can also be viewed as doing a round of Internal Double Crux, something like:

> **Exercise subagent:** We should exercise.
>
> **Optimizer subagent:** That feels unpleasant and costs a lot of energy, we would have the energy to do more things if we didn't exercise.
>
> **Exercise subagent:** That's true. But the feelings of unpleasantness are actually a sign of us getting more energy in the long term.
>
> **Optimizer subagent:** Oh, you're right! Then let's exercise, that furthers my goals too.

(There's also a bunch of other good stuff in the video that I didn't describe here, you may want to check it out if you haven't already done so.)

Exposure Therapy

So far, most of the examples have assumed that the person already has all the information necessary for solving the internal disagreement. But sometimes additional information might be required.

The prototypical use of *exposure therapy* is for phobias. Someone might have a phobia of dogs, while at the same time feeling that their fear is irrational, so they decide to get therapy for their phobia.

How the therapy typically proceeds is by exposing the person to their fear in increments that are as small as possible. For instance, a page by Anxiety Canada

offers this list of steps that someone might have for exposing themselves to dogs:

Step 1: Draw a dog on a piece of paper.

Step 2: Read about dogs.

Step 3: Look at photos of dogs.

Step 4: Look at videos of dogs.

Step 5: Look at dogs through a closed window.

Step 6: Then through a partly-opened window, then open it more and more.

Step 7: Look at them from a doorway.

Step 8: Move further out the doorway; then further etc.

Step 9: Have a helper bring a dog into a nearby room (on a leash).

Step 10: Have the helper bring the dog into the same room, still on a leash.

The ideal is that each step is enough to make you feel a little scared, but not so scared that it would serve to act retraumatize you or otherwise make you feel horrible about what happened.

In a sense, exposure therapy involves one part of the mind thinking that the situation is safe, and another part of the mind thinking that the situation is unsafe, and the contradiction being resolved by *testing* it. If someone feels nervous about looking at a photo of a dog, it implies that a part of their mind thinks that seeing a photo of a dog means they are potentially in danger. (In terms of the machine learning toy model from my IFS post, it means that a fear model is activated, which predicts the current state to be dangerous.)

By looking at photos sufficiently many times, and then afterwards noting that everything is okay, the nervous subagent gets information about having been wrong, and updates its model. Over time, and as the person goes forward in steps, the nervous subagent can eventually conclude that it had overgeneralized from the original trauma, and that dogs in general aren't that dangerous after all.

As in the CBT example, one can view this as activating conflicting models and the mind then fixing the conflict by updating the models. In this case, the conflict is between the frightened subagent's prediction that seeing the dog is a sign of danger, and another subagent's later assessment that everything turned out to be fine.

Conclusion to integration methods

I have considered here a number of ways of integrating subagent conflicts. Here are a few key principles that are used in them:

- **Selectively blending with subagents/beliefs to make disagreements between them more apparent.** Used in the Cognitive Behavioral Therapy and mental contrasting cases. Also used in a somewhat different form in exposure therapy, where you are partially blended with a subagent that thinks that the

situation is dangerous, while getting disagreeing information from the rest of the world.
- **Facilitating a dialogue between subagents “from the outside”.** Used in Internal Double Crux, Internal Family Systems. In a sense, the next bullet can also be viewed a special case of this.
    - **Combining aspects of the conflicting perspectives to a whole which allows for resolution.** Used in self-concept editing, Eliezer’s altruism example, and urge propagation.
- **Collecting additional information which allows for the disagreement to be resolved.** Used in exposure therapy.

I believe that we have evolved to use all of these spontaneously, without necessarily realizing what it is that we are doing.

For example, many people have the experience of it being useful to talk to a friend about your problems, weighting the pros and cons of different options. Frequently just getting to talk about it helps clarify the issue, even if the friend doesn’t say anything (or even if they are a rubber duck). Probably not coincidentally, if you are talking about the conflicting feelings that you have in your mind, then you are frequently doing something like an informal version of Internal Double Crux. You are representing all the sides of a dilemma until you have reached a conclusion and integrated the different perspectives.

To the extent that they are effective, various schools of therapy and self-improvement - ranging from CBT to IDC to IFS - are formalized methods for making such integration more effectively.

Subagents, neural Turing machines, thought selection, and blindspots

In my summary of Consciousness and the Brain (Dehaene, 2014), I briefly mentioned that one of the functions of consciousness is to carry out artificial serial operations; or in other words, implement a production system (equivalent to a Turing machine) in the brain.

While I did not go into very much detail about this model in the post, I've used it in later articles. For instance, in Building up to an Internal Family Systems model, I used a toy model where different subagents cast votes to modify the contents of consciousness. One may conceptualize this as equivalent to the production system model, where different subagents implement different production rules which compete to modify the contents of consciousness.

In this post, I will flesh out the model a bit more, as well as applying it to a few other examples, such as emotion suppression, internal conflict, and blind spots.

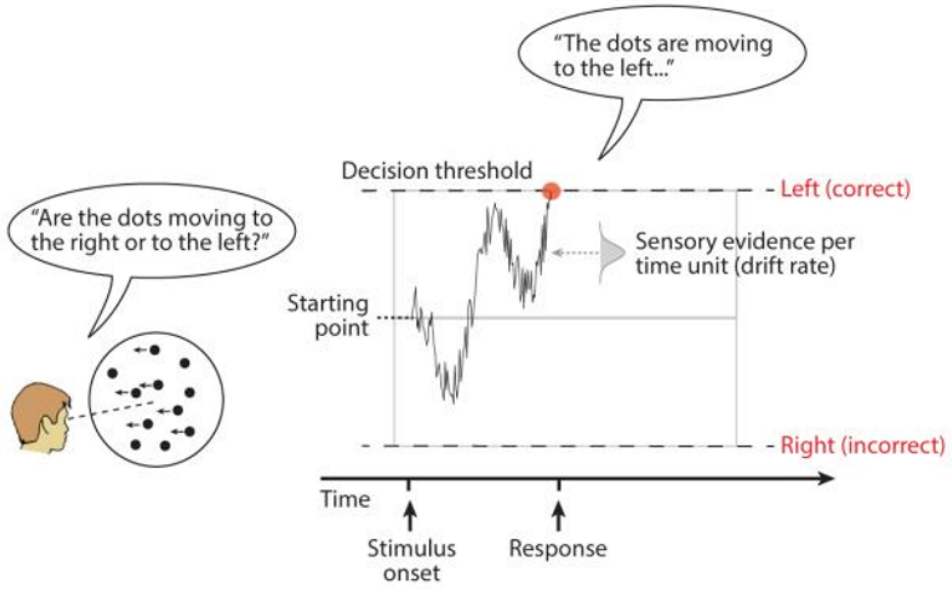
Evidence accumulation

Dehaene has outlined his model in a pair of papers (Zylberberg, Dehaene, Roelfsema, & Sigman, 2011; Dehaene & Sigman, 2012), though he is not the first one to propose this kind of a model. Daniel Dennett's *Consciousness Explained* (1991) also discusses consciousness as implementing a virtual Turing machine; both cite as examples earlier computational models of the mind, such as Soar and ACT, which work on the same principles.

An important building block in Dehane's model is based on what we know about evidence accumulation and decision-making in the brain, so let's start by taking a look at that.

Sequential sampling models (SSMs) are a family of models from mathematical psychology that have been developed since the 1960s (Forstmann, Ratcliff, & Wagenmakers, 2016). A particularly common SSM is the diffusion decision model (DDM) of decision-making, in which a decision-maker is assumed to noisily accumulate evidence towards a particular choice. Once the evidence in favor of a particular choice meets a decision threshold, that choice is taken.

For example, someone might be shown dots on a screen, some of which are moving in a certain direction. The task is to tell which direction the dots are moving in. After the person has seen enough dot movements, they will have sufficient confidence to make their judgment. The difficulty of the task can be precisely varied by changing the proportion of moving dots and their speed, making the movement easier or harder to detect. One can then measure how such changes affect the time needed for people to make a judgment.



A DDM is a simple model with just four parameters:

- *decision threshold*: a threshold for the amount of evidence in favor of one option which causes that option to be chosen
- *starting point bias*: a person may start biased towards one particular alternative, which can be modeled by them having some initial evidence putting them closer to one threshold than the other
- *drift rate*: the average amount of evidence accumulated per time unit
- *non-decision time*: when measuring e.g. reaction times, a delay introduced by factors such as perceptual processing which take time but are not involved in the decision process itself

These parameters can be measured from behavioral experiments, and the model manages to fit a wide variety of behavioral experiments and intuitive phenomena well (Forstmann et al., 2016; Ratcliff, Smith, Brown, & McKoon, 2016; Roberts & Hutcherson, 2019). For example, easier-to-perceive evidence in favor of a particular option is reflected in a faster drift rate towards the decision threshold, causing faster decisions. On the other hand, making mistakes or being falsely told that one's performance on a trial is below that of most other participants prompts caution, increasing people's decision thresholds and slowing down response times (Roberts & Hutcherson, 2019).

While the models have been studied the most in the context of binary decisions, one can easily extend the model to a choice between *n* alternatives by assuming the existence of multiple accumulators, each accumulating decision towards their own choice, possibly inhibiting the others in the process. Neuroscience studies have identified structures which seem to correspond to various parts of SSMs. For example, in random dot motion tasks, where participants have to indicate the direction that dots on a screen are moving in,

> the firing rates of direction selective neurons in the visual cortex (area MT/V5) exhibit a roughly linear increase (or decrease) as a function of the strength of motion in their preferred (or anti-preferred) direction. The average firing rate from a pool of neurons sharing similar direction preferences provides a time varying signal that can be compared to an average of another, opposing pool. This

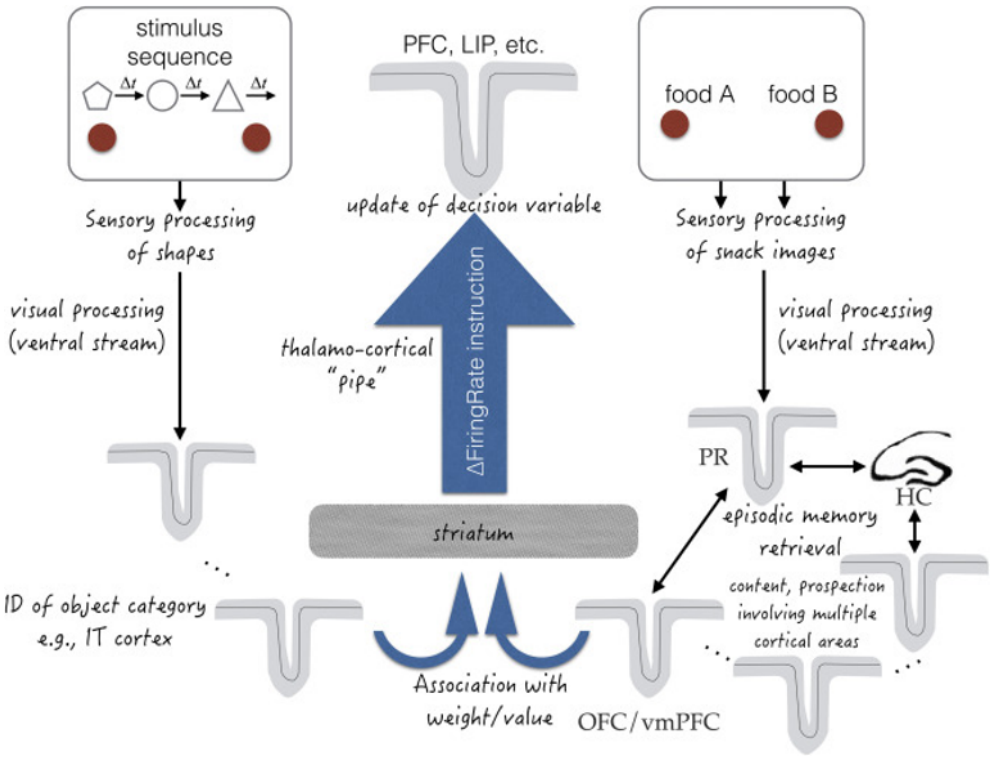
> difference can be positive or negative, reflecting the momentary evidence in favor of one direction and against the other. (Shadlen & Shohamy, 2016)

Shadlen & Shohamy (2016) note that experiments on more “real-world” decisions, such as decisions on which stock to pick or which snack to choose, also seem to be compatible with an SSM framework. However, this raises a few questions. For instance, it makes intuitive sense why people would take more time on a random motion task when they lose confidence: watching the movements for a longer time accumulates more evidence for the right answer, until the decision threshold is met. But what is the additional evidence that is being accumulated in the case of making a decision based on subjective value?

The authors make an analogy to a symbol task which has been studied in rhesus monkeys. The monkeys need to decide between two choices, one of which is correct. For this task, they are shown a series of symbols, each of which predicts one of the choices as being correct with some probability. Through experience, the monkeys come to learn the weight of evidence carried by each symbol. In effect, they are accumulating evidence not by motion discrimination but memory retrieval: retrieving some pre-learned association between a symbol and its assigned weight. This “leads to an incremental change in the firing rate of LIP neurons that represent the cumulative [likelihood ratio] in favor of the target".

The proposal is that humans make choices based on subjective value using a similar process: by perceiving a possible option and then retrieving memories which carry information about the value of that option. For instance, when deciding between an apple and a chocolate bar, someone might recall how apples and chocolate bars have tasted in the past, how they felt after eating them, what kinds of associations they have about the healthiness of apples vs. chocolate, any other emotional associations they might have (such as fond memories of their grandmother’s apple pie) and so on.

Shadlen & Shohamy further hypothesize that the reason why the decision process seems to take time is that different pieces of relevant information are found in physically disparate memory networks and neuronal sites. Access from the memory networks to the evidence accumulator neurons is physically bottlenecked by a limited number of “pipes”. Thus, a number of different memory networks need to take turns in accessing the pipe, causing a serial delay in the evidence accumulation process.



The biological Turing machine

In *Consciousness and the Brain,* Dehaene considers the example of doing arithmetic. Someone who is calculating something like 12 * 13 in their head, might first multiply 10 by 12, keep the result in memory, multiply 3 by 12, and then add the results together. Thus, if a circuit in the brain has learned to do multiplication, consciousness can be used to route its results to a temporary memory storage, with those results then being routed from the storage to a circuit that does addition.

Production systems in AI are composed of if-then rules (production rules) which modify the contents of memory: one might work by detecting the presence of an item like “10 * 12” and rewriting it as “120”. On a conceptual level, the brain is proposed to do something similar: various contents of consciousness activate neurons storing something like production rules, which compete to fire. The first one to fire gets to apply its production, changing the contents of consciousness.

If I understand Deheane’s model correctly, he proposes to apply the neural mechanisms discussed in the previous sections - such as neuron groups which accumulate evidence towards some kind of decision - at a slightly lower level. In the behavioral experiments, there are mechanisms which accumulate evidence towards which particular physical actions to take, but a person might still be distracted by unrelated thoughts while performing that task. Dehaene’s papers look at the kinds of mechanisms choosing *what thoughts to think*. That is, there are accumulator neurons which take “actions” to modify the contents of consciousness and working memory.

We can think of this as a two-stage process:

1. A process involving subconscious “decisions” about what thoughts to think, and what kind of content to maintain in consciousness. Evidence indicating the kind of conscious content is most suited for the situation is in part based on hardwired priorities, and in part stored associations about the kinds of thoughts that previously produced beneficial results.
2. A higher-level process involving decisions about what physical actions to take. While the inputs to this process do not necessarily need to go through consciousness, consciously perceived evidence has a much higher weight. Thus, the lower-level process has significant influence on which evidence gets to the accumulators on this level.

To be clear, this does not necessarily correspond to two clearly distinct levels: Zylberberg, Dehaene, Roelfsema, & Sigman (2011) do not talk about there being any levels, and they suggest that “triggering motor actions” is one of the possible decisions involved. But their paper seems to mostly be focused on actions - or, in their language, production rules - which manipulate the contents of consciousness.

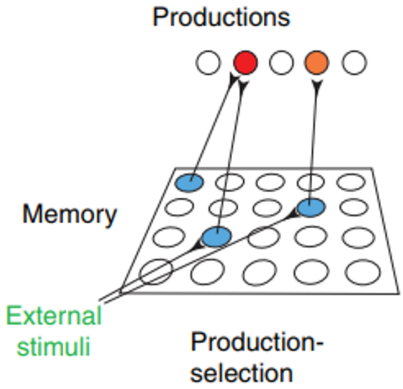
There seems to me to be a conceptual difference between the kinds of actions that change the contents of consciousness, and the kinds of actions which accumulate evidence over many items *in* consciousness (such as iterative memories of snacks). Zylberberg et al. talk about a “winner-take-all race” to trigger a production rule, which to me implies that the evidence accumulated in favor of each production rule is cleared each time that the contents of consciousness is changed. This is seemingly incompatible with accumulating evidence over many consciousness-moments, so postulating a two-level distinction between accumulators seems like a straightforward way of resolving the issue.

[EDIT: Hazard suggests that the two-level split is implemented by the basal ganglia carrying out evidence accumulation across changes in conscious content.]

As an aside, I am, as Dehaene is, treating consciousness and working memory as basically synonymous for the purposes of this discussion. This is not strictly correct; e.g. there may be items in working memory which are not currently conscious. However, since it’s generally thought that items in working memory need to be actively rehearsed *through* consciousness in order to avoid be maintained, I think that this equivocation is okay for these purposes.

Here’s a conceptual overview of the stages in the “biological Turing machine’s” operation (as Zylberberg et al. note, a production firing “is essentially equivalent to the action performed by a Turing machine in a single step”):

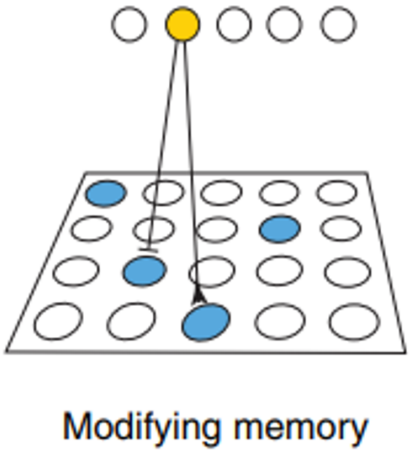
1. The production selection stage



At the beginning of a cognitive cycle, a person’s working memory contains a number of different items, some internally generated (e.g. memories, thoughts) and some external (e.g. the sight or sound of something in the environment). Each item in memory may activate (contribute evidence to) neurons which accumulate weight towards triggering a particular kind of production rule. When some accumulator neurons reach their decision threshold, they apply their associated production rule.

In the above image, the blue circles at the bottom represent active items in working memory. Two items are activating the same group of accumulator neurons (shown red) and one is activating an unrelated one (shown brown).

2. Production rule ignition



Once a group of accumulator neurons reach their decision threshold and fire a production rule, the model suggests that there are a number of things that the rule can do. In the above image, an active rule is modifying the contents of working memory: taking one of the blue circles, deleting it, and creating a new blue circle nearby. Hypothetically, this might be something like taking the mental objects holding “120” and “36”, adding them together, and storing the output of “156” in memory.

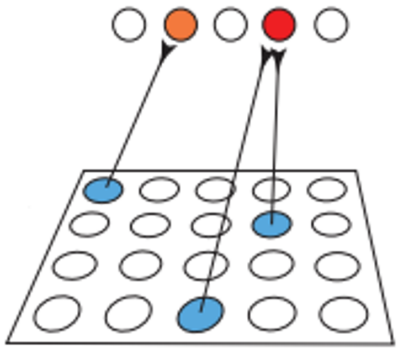
Obviously, since we are talking about brains, expressions like "writing into memory" or "deleting from memory" need to be understood in somewhat different terms than in

computers; something being “deleted from working memory” mostly just means that a neuronal group which was storing the item in its firing pattern stops doing so.

The authors suggest that among other things, production rules can:

- trigger motor actions (e.g. saying or doing something)
- change the contents of working memory to trigger a new processing step (e.g. saving the intermediate stage of an arithmetic operation, together with the intention to proceed with the next step)
- activate and broadcast information that is in a “latent” state (e.g. retrieving a memory and sending it to consciousness)
- activate peripheral processors capable of performing specific functions (e.g. changing the focus of attention)

3. New production selection



Production-selection

After the winning production rule has been applied, the production selection phase begins anew. At this stage or a future one, some kind of a credit assignment process likely modifies the decision weights involved in choosing production rules: if a particular rule was activated in particular circumstances and seemed to produce positive consequences, then the connections which caused those circumstances to be considered evidence for that rule are strengthened.

Practical relevance

Okay, so why do we care? What is the practical relevance of this model?

First, this helps make some of my previous posts more concrete. In Building up to an Internal Family Systems model, I proposed some sort of a process where different subagents were competing to change the contents of consciousness. For instance, “manager” subagents might be trying to manipulate the contents of consciousness so as to avoid unpleasant thoughts and to keep the person out of dangerous circumstances.

People who do IFS, or other kinds of “parts work”, will notice that different subagents are associated with different kinds of bodily sensations and flavors of consciousness. A

priori, there shouldn’t be any particular reason for this… except, perhaps, if the strength of such sensations correlated with the activation of a particular subagent, with those sensations then being internally used for credit assignment to identify and reward subagents which had been active in a given cognitive cycle. (This is mostly pure speculation, but supported by some observations to which I hope to return in a future post.)

In my original post, I mostly talked about exiles - neural patterns blocked from consciousness by other subagents - as being subagents related to a painful memory. But while it is not emphasized as much, IFS holds that other subagents can in principle be exiled too. For example, a subagent which tends to react using anger may frequently lead to harmful consequences, and then be blocked by other subagents. This can easily be modeled using the neural Turing machine framework: over time, the system learns decisions which modify consciousness so as to prevent the activation of a production rule that gives power to the angry subagent. As this helps avoid harmful consequences, this begins to happen more and more often.

Hazard has a nice recent post of this kind of a thing happening with emotions in general:

> So young me is upset that the grub master for our camping trip forgot half the food on the menu, and all we have for breakfast is milk. I couldn't "fix it" given that we were in the woods, so my next option was "stop feeling upset about it." So I reached around in the dark of my mind, and Oops, the "healthily process feelings" lever is right next to the "stop listening to my emotions" lever.
>
> The end result? "Wow, I decided to stop feeling upset, and then I stopped feeling upset. I'm so fucking good at emotional regulation!!!!!"
>
> My model now is that I substituted "is there a monologue of upsetness in my conscious mental loop?" for "am I feeling upset?". So from my perspective, it just felt like I was very in control of my feelings. Whenever I wanted to stop feeling something, I could. When I thought of ignoring/repressing emotions, I imagined trying to cover up something that was there, maybe with a story. Or I thought if you poked around ignored emotions there would be a response of anger or annoyance. I at least expected that if I was ignoring my emotions, that if I got very calm and then asked myself, "Is there anything that you're feeling?" I would get an answer.
>
> Again, the assumption was, "If it's in my mind, I should be able to notice if I look." This ignored what was actually happening, which was that I was cutting the phone lines so my emotions couldn't talk to me in the first place.

Feeling upset feels bad, ceasing to feel upset feels good. Brain notices that there is some operation which causes the feeling of upset to disappear from consciousness: carrying out this operation also produces a feeling of satisfaction in the form of “yay, I’m good at emotional regulation!”. As a result of being rewarded, it eventually becomes so automatic as to block even *hints* of undesired emotions, making the block in question impossible to notice.

Another observation is that in IFS as well as in Internal Double Crux, an important mental move seems to be “giving subagents a chance to finish talking”. For instance, subagent A might hold a consideration pointing in a particular direction, while subagent B holds a consideration in the opposite direction. When A starts presenting its points, B interrupts with its own point; in response, A interrupts with *its* point. It

seems to be possible to commit to not taking a decision before having heard both subagents, and having done that, ask them to take turns presenting their points and not interrupt each other. What exactly is going on here?

Suppose that a person is contemplating the decision, "should I trust my friend to have my back in a particular risky venture". Subagent A holds the consideration "allies are important, and we don't have any, we should really trust our friend so that we would have more allies". Subagent B holds the consideration "being betrayed would be really bad, and our friend seems untrustworthy, it's important that we don't sign up for this". Subagent A considers it *really important* to go on this venture together; subagent B considers it *really important* **not** to.

Recall that human decision-making happens by accumulating evidence towards different choices, until a decision threshold is met. If A were allowed to present its evidence in favor of signing up on the venture, that might sway the decision over the threshold before B was able to present the evidence against. Thus, there is a mechanism which allows B to "interrupt" A, in order to present its own evidence. Unfortunately, it is now A which risks B's evidence being sufficient to meet a decision threshold prematurely unless *B* is prevented from presenting its evidence, so A must interrupt.

Subjectively, this is experienced as intense internal conflict, with two extreme considerations pushing in opposite directions, allowing no decision to be made - unless there is a plausible commitment to *not* making a decision until both have been heard out. (To me, this feels like my attention being caught in a tug-of-war between one set of considerations versus another. Roberts & Hutcherson (2019) note that *A large body of work suggests that negative information draws focus through rapid detection [64–68] and attentional capture [69–71]. [....] Several studies now show that attending to a choice alternative or attribute increases its weighting in the evidence accumulation process [72–75]. To the extent that negative affect draws attention to a choice-relevant attribute or object, it should thus increase the weight it receives.*)

There's one more important consideration. Eliezer has written about cached thoughts - beliefs which we have once acquired, then never re-evaluated and just acted on them from that onwards. But this model suggests that things may be worse: it's not just that we are running on cached thoughts. Instead, *even the pre-conscious mechanisms deciding which thoughts are worth re-evaluating are running on cached values.*

Sometimes external evidence may be sufficient to force an update, but there can also be self-fulfilling blind spots. For instance, you may note that negative emotions never even surface into your consciousness. This observation then triggers a sense of satisfaction about being good at emotional regulation, so that thoughts about alternative - and less pleasant - hypotheses are never selected for consideration. In fact, evidence to the contrary may feel actively unpleasant to consider, triggering subagents which use feelings such as annoyance - or if annoyance would be too suspicious, just plain indifference - to push that evidence out of consciousness, before it can contribute to a decision.

And the older those flawed assumptions are, the more time there is for additional structures to build on top of them.

This post is part of research funded by the Foundational Research Institute. Thanks to Maija Haavisto for line editing an initial draft.

References

Dehaene, S. (2014). *Consciousness and the Brain: Deciphering How the Brain Codes Our Thoughts*. New York, New York: Viking.

Dehaene, S., & Sigman, M. (2012). From a single decision to a multi-step algorithm. *Current Opinion in Neurobiology*, *22*(6), 937–945.

Dennett, D. C. (1991). *Consciousness Explained* (1st edition). Boston: Little Brown & Co.

Forstmann, B. U., Ratcliff, R., & Wagenmakers, E.-J. (2016). Sequential Sampling Models in Cognitive Neuroscience: Advantages, Applications, and Extensions. *Annual Review of Psychology*, *67*, 641–666.

Ratcliff, R., Smith, P. L., Brown, S. D., & McKoon, G. (2016). Diffusion Decision Model: Current Issues and History. *Trends in Cognitive Sciences*, *20*(4), 260–281.

Roberts, I. D., & Hutcherson, C. A. (2019). Affect and Decision Making: Insights and Predictions from Computational Models. *Trends in Cognitive Sciences*, *23*(7), 602–614.

Shadlen, M. N., & Shohamy, D. (2016). Decision Making and Sequential Sampling from Memory. *Neuron*, *90*(5), 927–939.

Zylberberg, A., Dehaene, S., Roelfsema, P. R., & Sigman, M. (2011). The human Turing machine: a neural framework for mental programs. *Trends in Cognitive Sciences*, *15*(7), 293–300.

Subagents, trauma and rationality

[Content note: discussion of trauma, child and sexual abuse, sexual violence, lack of self-worth, dissociation, PTSD, flashbacks, DID, personality disorders; some mildly graphic examples of abuse and trauma mentioned in text form]

> I have spent over two years doing emotional support for people who had survived long-term childhood trauma, and in these cases spawning agents to deal with unbearable suffering while having no escape from it is basically a standard reaction that the brain/mind takes. The relevant psychiatric diagnosis is DID (formerly MPD, multiple personality disorder). In these cases the multiple agents often manifest very clearly and distinctly. It is tempting to write it off as a special case that does not apply in the mainstream, yet I have seen more than once the progression from someone suffering from CPTSD to a full-blown DID. The last thing that happens is that the person recognizes that they "switch" between personalities. Often way later than when others notice it, if they know what to look for. After gaining some experience chatting with those who survived severe prolonged trauma, I started recognizing subtler signs of "switching" in myself and others. This switching between agents (I would not call them sub-agents, as they are not necessarily less than the "main", and different "mains" often take over during different parts of the person's life) while a normal way to operate, as far as I can tell, almost never rises to the level of conscious awareness, as the brain carefully constructs the lie of single identity for as long as it can.

-- shminux

As the above comment suggests, the appearance of something like distinct subagents is particularly noticeable in people with heavy trauma, DID being the most extreme example.

This post will interpret the appearance of subagents as emerging from *unintegrated memory networks*, and argue that - as shminux suggests - the presence of these is a matter of degree. There's a continuous progression of fragmented (dissociated) memory networks giving arise to increasingly worse symptoms as the degree of fragmentation grows. The continuum goes from everyday procrastination and akrasia on the "normal" end, to disrupted and dysfunctional beliefs on the middle, and conditions like clinical PTSD, borderline personality disorder, and dissociative identity disorder on the severely traumatized end.

I will also argue that emotional work and exploring one's past traumas in order to heal them, is necessary for effective instrumental and epistemic rationality.

This post is largely based on what I understand to be relatively standard trauma theory (e.g. van der Kolk, 2014; Shapiro, 2017; Baldwin, 2013; Schauer & Elbert, 2010; Forgash & Copeley, 2007) and should not contain any particularly novel or original claims, except maybe for drawing some connections to topics and framings which I have been discussing previously in my sequence.

Emotional regulation as an approach-avoid tradeoff

In Building up to an Internal Family Systems model, I talked about "protector" subagents (subdivided into managers and firefighters), whose purpose was to keep negative emotions and memories ("exile" subagents) out of consciousness. If they predicted that entering some situation would trigger a negative memory, then they would try to prevent the person from going to that situation, because a situation triggering a negative memory correlates with such situations being dangerous. Thus, my explanation suggested that the only purpose of protectors was to keep a person out of concrete danger.

However, something like protectors is also a necessary component for emotional regulation. There are a number of difficult tradeoffs implied by the following facts:

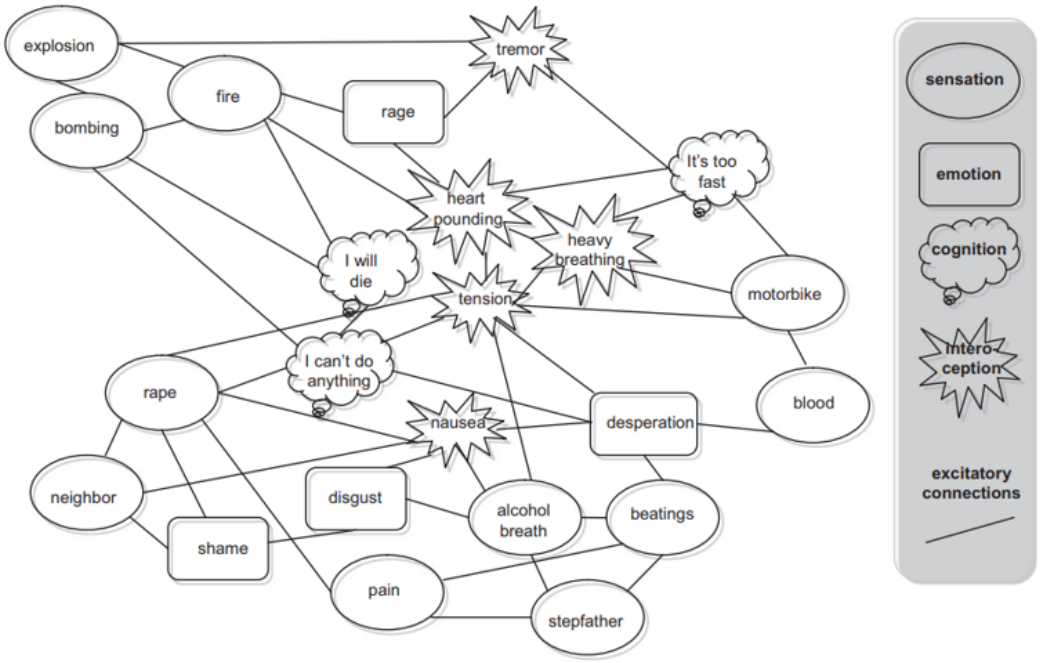
- In detecting possible threats, it is usually better to err on the side of too many false positives. Mistaking a tree branch for a snake is less costly than mistaking a snake for a tree branch.
- At the same time, false positives which trigger automatic fight-or-flight-or-freeze responses *do* also have a cost; once an alarm has been shown to be false, there need to be mechanisms for winding it down.
- Sometimes a situation *is* dangerous and unpleasant and costly, *and* going into it is what you need to do anyway. There has to be a mechanism for overriding the alarms when it is necessary.
- A special case of this is when trying to flee a dangerous situation is by itself dangerous. If you're wounded and there's an enemy nearby, you may have better odds trying to play dead than running away. You may also be someone's slave with poor chances of escaping. For those cases, you need a mechanism specifically for shutting down the fight-or-flight response.
- At the same time, an organism which had too easy of a time overriding its alarms would not take them sufficiently seriously, and would be killed or otherwise hurt by constantly going into dangerous situations and not trying to escape them.

In humans as well as other mammals, brain areas controlling evolutionarily ancient defense state responses become active when danger is detected. While the "higher" cognitive functions of the frontal cortex are to some extent capable of regulating these emotional responses when they are mild, the emotional brain can and will override the frontal cortex in dangerous situations. In situations of sufficient distress, rational thinking and the ability to regulate emotional responses shut down entirely.

Furthermore, sensory inputs are constantly scanned by the amygdala for patterns which have been associated with danger in the past. The time it takes for the amygdala to process inputs is shorter than the time it takes for them to reach consciousness, and the amygdala may trigger an emotional response before higher cognitive systems even become aware of the situation.

Thus, if the emotional brain detects something it considers threatening enough, it *will* react, regardless of whether or not the cognitive brain considers it a good idea. If particular cues tend to co-occur with particularly serious danger states reliably enough, then the brain will build up an associative network where detecting any of those cues will automatically trigger threat responses, with no opportunity for

cognitive overrides. This can become a serious problem for normal functioning, as any of the cues in the fear network may trigger an emergency response, even in completely harmless situations.



Example trauma network (Schauer & Elbert 2010).

Many of the elements of the pictured trauma network - neighbors, motorbikes, the smell of alcohol - are ones that would ordinarily have plenty of connections to other concepts in daily life. In order to prevent extreme responses, the associative network related to the trauma needs to be compartmentalized and isolated from the rest of a person's memory networks, or risk anything activating the network and triggering an emergency response.

Thus, the role of protectors is not just to avoid situations which would be actively dangerous: they also need to manipulate the contents of consciousness so as to avoid triggering the trauma network otherwise.

Even if someone's life circumstances have changed, and the original situation which created the trauma is no longer an active issue (such as if someone is an adult and has moved away from their abusive parents), the trauma network may remain too strongly charged to allow its contents to be reprocessed. Reprocessing would require the use of higher cognitive functions to put the experiences in a new context, but any activation of the network will trigger an immediate emergency response, shutting down those very cognitive functions. In such a situation, all that protectors can do is to try to bury the network and suppress all memories related to it. To rephrase that in less intentional language, the brain's Turing machine may come to learn that suppressing particular memories produces beneficial results, reinforcing the behavior.

> Memory loss has been reported in people who have experienced natural disasters, accidents, war trauma, kidnapping, torture, concentration camps, and physical

> and sexual abuse. Total memory loss is most common in childhood sexual abuse, with incidence ranging from 19 percent to 38 percent. This issue is not particularly controversial: As early as 1980 the DSM-III recognized the existence of memory loss for traumatic events in the diagnostic criteria for dissociative amnesia: “an inability to recall important personal information, usually of a traumatic or stressful nature, that is too extensive to be explained by normal forgetfulness.” Memory loss has been part of the criteria for PTSD since that diagnosis was first introduced.
>
> One of the most interesting studies of repressed memory was conducted by Dr. Linda Meyer Williams, which began when she was a graduate student in sociology at the University of Pennsylvania in the early 1970s. Williams interviewed 206 girls between the ages of ten and twelve who had been admitted to a hospital emergency room following sexual abuse. Their laboratory tests, as well as the interviews with the children and their parents, were kept in the hospital’s medical records. Seventeen years later Williams was able to track down 136 of the children, now adults, with whom she conducted extensive follow-up interviews. More than a third of the women (38 percent) did not recall the abuse that was documented in their medical records, while only fifteen women (12 percent) said that they had never been abused as children. More than two-thirds (68 percent) reported other incidents of childhood sexual abuse. Women who were younger at the time of the incident and those who were molested by someone they knew were more likely to have forgotten their abuse. (van der Kolk, 2014)

However, as I will soon discuss, a downside of this process is that the more that associative networks get disconnected from each other, the less consistent a person’s responses to different situations become. If a person is drawing on different memories in different situations, then in some situations they may seem like an entirely different person than in others. Another way of framing is that, if we define a subagent as decision-making entity that has access to its own set of beliefs and goals, then different subagents are in control at different times.

A few words on the “memory wars”

This post discusses suppressing traumatic memories, drawing on the theories of clinical practitioners, who have disagreements with clinical researchers about whether memory suppression is a thing (Patihis, Ho, Tingen, Lilienfeld, & Loftus, 2014).

Much of the criticism about repressed memories is aimed at a specific concept from Freudian theory, and/or on the question of how reliable therapeutically recovered memories are. Several of the critics (e.g. (Rofé, 2008)) acknowledge that people may suppress or intentionally forget painful memories, but argue that this is distinct from the Freudian concept of repression. However, memory suppression in the sense discussed in this post is not related to the Freudian concept, and also includes intentional attempts to forget or avoid thinking about something, as the examples will hopefully demonstrate.

In fact, the memories being *hard* to forget is exactly the problem, which is something that many critics of the standard Freudian paradigm are keen to point out - traumatic memories are often particularly powerful and long-lasting.

I do make the assumption that conscious attempts to forget something may eventually become sufficiently automated so as to become impossible for the person

themselves to notice; but this seems like a straightforward inference from the observation that skills and habits in general can become automated enough so as to happen without the person realizing what they are doing. A recent experiment (unreplicated, but I have a reasonably high prior for cognitive psychology experiments replicating) also showed that once people are trained to intentionally forget words that are associated with a particular cue, the cue will reduce recall of words even when it is paired with them in a form that is too short to consciously register (Salvador et al. 2018).

I make no strong claims about the reliability of memories recovered in therapy. It has been clearly demonstrated that it is possible for therapists to accidentally or intentionally implant false memories, but there have also been cases of people recovering memories which have then been confirmed from other sources. Probably some recovered memories are genuine (though possibly distorted) and some are not.

Mild disconnection: unintegrated considerations and ugh fields

In my previous post, I mentioned the hypothesis (Shadlen & Shohamy, 2016) that in choosing between several options, we sample memories related to our past experiences with those options. This sampling may draw from several memory networks, located in physically distinct areas in the brain. I also mentioned how particularly negative memories or consideration may draw one's attention, so as to make it hard to integrate those concerns with the concerns in other networks. And in Integrating disagreeing subagents, I talked about how akrasia might be interpreted in terms of conflicting and unintegrated beliefs pulling in opposite directions.

In the very mildest form, there isn't necessarily any trauma at all: just two different associative networks pointing in opposite directions, with concerns that have not been integrated. This might cause different kinds of behavior in different situations, depending on which network happens to become activated. However, once the discrepancy is detected, integration may happen automatically or assisted in a relatively straightforward fashion with techniques such as IDC.

The next level is when a concern is not serious, but still feels somewhat unpleasant to think about - the territory of everyday ugh fields:

> For example, suppose that you started off in life with a wandering mind and were punished a few times for failing to respond to official letters. Your TDL algorithm began to propagate the pain back to the moment you looked at an official letter or bill. As a result, you would be less effective than average at responding, so you got punished a few more times. Henceforth, when you received a bill, you got the pain before you even opened it, and it laid unpaid on the mantelpiece until a Big Bad Red late payment notice with an $25 fine arrived. More negative conditioning. Now even thinking about a bill, form or letter invokes the flinch response, and your lizard brain has fully cut you out out. You find yourself spending time on internet time-wasters, comfort food, TV, computer games, etc. Your life may not obviously be a disaster, but this is only because you can't see the alternative paths that it could have taken if you had been able to take advantage of the opportunities that came as letters and forms with deadlines.

> The subtlety with the Ugh Field is that the flinch occurs before you start to consciously think about how to deal with the Unhappy Thing, meaning that you never deal with it, and you don't even have the option of dealing with it in the normal run of things.

As the post notes, your brain is automatically trying to avoid thinking about things which would trigger the "ugh". (I've certainly noticed in myself a tendency to just conveniently forget about lots of mildly unpleasant things I should get around doing.) Plans which would involve engaging the ugh are ranked low in your preference ordering so are never even generated as options. Except that *some* of your networks *do* know that you need to engage with the ugh eventually, so they keep annoyingly reminding you about it whenever something happens to activate them.

Assuming that the ughs are something that you do need to deal with, that is. They could also be something else, such as past upsets and embarrassments that you would prefer to forget, and then keep pushing away until nothing reminds you of them anymore. In that case, the memory network might be successfully locked away and compartmentalized - but still exerting a subtle influence in the form of the now-automated mental motions aimed at making sure that it remains hidden. Those same motions would also prevent the beliefs associated with it from coming up and being re-evaluated.

A relevant question here is: if a negative memory is suppressed from consciousness, to what extent does it influence decision-making? After all, one might think that if it remains suppressed, then it should not have any effect. But it seems likely that even if the memory itself does not influence decision-making, the hoops that the brain has learned to jump through to keep it suppressed do affect one's decisions. (Under a subagent framing, those hoops in question would be thought of as a protector.)

To take a minor example not involving memory suppression, for some reason I did not learn to tie my shoelaces as a child. Instead I just used shoes which did not have laces. After a while, it became kind of embarrassing to not know how to do that, so I developed an identity and preference for using shoes without laces. This lasted well into adulthood, despite the fact it would have been pretty trivial for me to just finally learn how to do it. Yet the thought of asking someone to teach me would have felt embarrassing, so my brain continued using the structures it had developed for avoiding shoelaces. Even if I had somehow suppressed the knowledge of these structures existing because of my embarrassment - something of which I was on some level aware all the time - I very much doubt that that alone would have changed the structures, or reduced my resistance to using shoelaces. It might plausibly even have made things worse, as I would no longer even realize that my identity was a justification constructed to guard against embarrassment. (Eventually I looked up a YouTube video called something like "easiest way of teaching your child to tie their shoelaces" and figured it out.)

Eliezer says that "If you once tell a lie, the truth is ever after your enemy" - maintaining one incorrect justification for your behavior may require your brain to contort itself to quite a lot of weird shapes. We all probably know examples of people who seem reasonable and rational on most issues, but then on some they are oddly forceful about their positions, or otherwise do not quite seem to be thinking clearly. (But that's just *other* people being irrational, of course. *We* would never do such a thing.)

> A “big T” trauma (Criterion A event necessary to diagnose PTSD), such as rape, sexual molestation, or combat experience, clearly has an impact on its victims in terms of how they behave, think, and feel about themselves, and in their susceptibility to pronounced symptoms, such as nightmares, flashbacks, and intrusive thoughts. These victims will have self-attributions such as “I’m powerless,” “I’m worthless,” or “I’m not in control.” Of course, clients who have not experienced such traumas may also have dominant negative self-attributions, such as “I’m worthless,” “I’m powerless,” or “I’m going to be abandoned.” Many of these clients seem to have derived their negative self-statements from early childhood experiences. […] Like “big T” trauma victims, they see the event, feel it, and are profoundly affected by it.
>
> Such clients were not, of course, blown up in a minefield or molested by a parent. Nevertheless, a memory of something that was said or that happened to them is locked in their brain and seems to have an effect similar to that of a traumatic experience. In fact, by dictionary definition, any event that has had a lasting negative effect on the self or psyche is by its nature “traumatic.” Consequently, these ubiquitous adverse life events have long been referred to in EMDR practice as “small t” traumas to keep in mind the nature of their impact […]. A wide range of adverse life experiences can be the basis of pathology, because of their emotional impact. For instance, while being humiliated in grade school cannot be designated a “trauma” for the diagnosis of PTSD, on an emotional level, such an event can be considered the evolutionary equivalent of being cut out of the herd. The impact can be affectively devastating, with long-lasting effects. (Shapiro, 2017)

Moderate disconnection: unintegrated core beliefs

Sometimes a belief network is in some sense too central to just be pushed away: life circumstances force it to be active despite being negatively laden, but that negativity also prevents it from being integrated with other networks. Beliefs about ourselves are a particularly common candidate for this category.

> Imagine that a little girl is walking beside her father and reaches up for his hand. At that moment the father deliberately or inadvertently swings his arm back and hits the child in the face. The child experiences intense negative affect, which might be verbalized as “I can’t get what I want; there is something wrong with me.” […] The affect, perhaps intense feelings of worthlessness and powerlessness, and the images, sounds, and the pain of the blow are stored in the child’s nervous system. This experience becomes a touchstone, a primary self-defining event in her life; in the Adaptive Information Processing model we call it a node. […] the next event that represents a similar rejection is likely to link up with the node in the ongoing creation of a neuro network that will be pivotal to the girl’s definition of her self-worth. Subsequent experiences of rejection by mother, siblings, friends, and others may all link up with the node in channels of associated information. Even before language is adequately developed, all the different childhood experiences containing similar feelings of powerlessness, despair, and inadequacy are stored as information linking into a memory network organized around the node of the earlier touchstone experience. Positive experiences are not assimilated into the network because the node is defined by the negative affect.

> When there is sufficient language to formulate a self-concept, such as “I can’t get what I want; there is something wrong with me,” verbalization is linked associatively with the network by the affect that the meaning of those words engenders. In essence, once the affect-laden verbal conceptualization is established in the neuro network, it can be viewed as generalizing to each of the subsequent experiences stored as information in the network. The process continues in adolescence, such as when, for instance, the girl in our example experiences a rejection by a teacher or a boyfriend. Thus, all subsequent related events may link to the same node point and take on the attributions of the initial experience. Therefore, the assessment associated with such an event is not limited to a function-specific statement (e.g., “I can’t get what I want in this instance”), but is linked to the dysfunctional generalized statement “I can’t get what I want; there is something wrong with me.”
>
> What happens when the girl reaches adulthood and something happens that seems like—or even threatens to become—a rejection? This new information is assimilated into the neuro network, and the concept “I can’t get what I want; there is something wrong with me” and its affect generalize and become associated with it. Over time, the accumulated related events produce a self-fulfilling prophecy; thus, any hint or chance of rejection can trigger the neuro network with its dominant cognition of “There is something wrong with me.” This person’s consequent behavior and attributions in the present are dysfunctional because what motivates and fuels them is the intense affect, fear, pain, and powerlessness of that first experience, now compounded by all of the subsequent experiences. (Shapiro, 2017)

Despite the prevalence of the negative affect, there is still a strong need to push it away, in order to be able to function normally and experience positive feelings. This leads to something like unstable fluctuation of at least two distinct memory networks taking turns being active. One is loaded with all the negative examples, while another might contain all the positive ones which have not been successfully integrated into the negative ones.

Also, the person in question might understand on an *intellectual* level that there is nothing actually wrong with them, that this is a negative cognition created by trauma, and so on… but this belief also resides in its own network, separate from the one which is causing the negative experience.

Dissociation

The examples so far have largely assumed that one is capable of somehow - in principle at least - avoiding the unpleasant situation or trigger. But what if one is not?

While “fight or flight” are the most commonly known defensive reactions, there are actually several defensive states. Exactly how many and how they should be classified is somewhat disputed, but one model has the following in increasing order of threat imminence: freeze-alert (stopping still and paying attention when noticing something potentially threatening), fight, flight, freeze-fright (if the fight and flight options are unviable), and collapse (feigning death):

> A simple thought experiment illustrates these five defensive options. Imagine that you surprise a large bear while alone in the wilderness. Your immediate stillness is the freeze-alert state. If the bear moves off, you can return home with an exciting

> story for your family. If the bear approaches, your danger deepens. Neither flight nor fight offers a viable option in this and many other cases of extreme threat, where active defenses increase the risk of death. The best option here is freeze-fright, although your chances are slim unless a hunter is nearby. Finally, when the bear has you in its mouth, you are out of options. You go limp in a state of collapse […] Collapse reduces the likelihood of continued violence, while preparing the individual for injury or death (release of endogenous opioids decreases pain; […]). Immobility is the most effective response during attack because quiescence eliminates auditory and visual cues that elicit or maintain aggression. All of these defense states survive in us from our evolutionary past because each has enhanced the odds of survival. (Baldwin 2013)

"Dissociation" is a broad term with several meanings; in reference to states of consciousness in particular, it refers to states such as ones in which a person feels detached from the world, up to the point of their experiences feeling unreal, them being unable to see or hear anything, or feeling like they are someone else and watching events which are happening to some stranger.

Schauer & Elbert (2010) conceptualize dissociative states of consciousness as ones which are related to the freeze and collapse stages of the defensive progression. In global workspace terms, one might frame it as a workspace state which helps - as a part of a wider physiological shutdown - prevent the kinds of more active defensive responses which would put the person in more danger. Many rape and abuse victims describe having had dissociative episodes during their experience.

The milder, non-pathological forms of dissociative states include e.g. daydreaming, which may help keep the person still in situations such as boring lectures with obligatory attendance, suppressing desires to move elsewhere. The more serious forms may kick in during e.g. repeated abuse a person does not have a chance to escape from, and to protect against memories of such incidents. (Schauer & Elbert also hypothesize that forms of self-harm, such as cutting, may act as ways to self-regulate by escalating one's defensive state to one of shutdown, calming down stressful memories and responses.)

If someone is forced to live with their abuser, dissociation may help them remain functional in the abuser's presence rather than trying to uselessly flee; as dissociative states frequently involve memory deficits, this may further lead to fragmented memory networks.

Extreme disconnection: PTSD, personality disorders, DID

An extreme case of incomplete memory suppression is PTSD, where the trauma network may be so intense as to completely overwhelm the person whenever it is activated. As a result, cognitive analysis may shut down whenever the network is triggered. The person may become completely flooded with the memory, to the point of reliving it as if they were experiencing the event again.

> Because the overwhelm is so strong, they may afterwards have little memory of what even happened - the cognitive shutdown may suppress the ability to form new memories of the event or to express it in language, leaving the experience completely in a world of its own.

> The overwhelming experience is split off and fragmented, so that the emotions, sounds, images, thoughts, and physical sensations related to the trauma take on a life of their own. The sensory fragments of memory intrude into the present, where they are literally relived. As long as the trauma is not resolved, the stress hormones that the body secretes to protect itself keep circulating, and the defensive movements and emotional responses keep getting replayed. […]
>
> … many people may not be aware of the connection between their "crazy" feelings and reactions and the traumatic events that are being replayed. They have no idea why they respond to some minor irritation as if they were about to be annihilated. Flashbacks and reliving are in some ways worse that the trauma itself. A traumatic event has a beginning and an end—at some point it is over. But for people with PTSD a flashback can occur at any time, whether they are awake or asleep. There is no way of knowing when it's going to occur again or how long it will last. People who suffer from flashbacks often organize their lives around trying to protect against them. They may compulsively go to the gym to pump iron (but finding that they are never strong enough), numb themselves with drugs, or try to cultivate an illusory sense of control in highly dangerous situations (like motorcycle racing, bungee jumping, or working as an ambulance driver). Constantly fighting unseen dangers is exhausting and leaves them fatigued, depressed, and weary. (van der Kolk 2014)

There is debate about whether or not DID is a real phenomenon. I do not have the expertise to have a strong opinion on this, but to the extent that it is, it could be considered a more extreme version of the dissociative responses described in the previous sections. Forced to live for an extended time in extreme circumstances with contradictory demands - such as being required to be happy and obedient while also being the subject of regular extreme abuse - a child may develop extreme amnesiac walls between different memory networks, to the point of having an entirely different personality depending on which network happens to be active. This allows for feelings of fear or panic that would otherwise arise in the company of an abusive parent, to be kept away as the memories associated with the abuse cannot be accessed.

Several clinicians also consider borderline personality disorder to be a PTSD-like response to trauma, with symptoms such as extreme fears of abandonment, alternation between idealization and devaluation, and poor emotional regulation being caused by the kinds of mechanisms that I have been describing.

In cases of sufficient trauma, an extreme form of dissociation seems to happen, where protectors extensively shut down access to bodily sensations and emotional awareness, turning off any systems which might cause emotional reactions that would be too strong for the system to handle:

> While Sherry dutifully came to every appointment and answered my questions with great sincerity, I did not feel we were making the sort of vital connection that is necessary for therapy to work. Struck by how frozen and uptight she was, I suggested that she see Liz, a massage therapist I had worked with previously. During their first meeting Liz positioned Sherry on the massage table, then moved to the end of the table and gently held Sherry's feet. Lying there with her eyes closed, Sherry suddenly yelled in a panic: "Where are you?" Somehow Sherry had lost track of Liz, even though Liz was right there, with her hands on Sherry's feet.
>
> Sherry was one of the first patients who taught me about the extreme disconnection from the body that so many people with histories of trauma and

> neglect experience. [...] Once I was alerted to this, I was amazed to discover how many of my patients told me they could not feel whole areas of their bodies. Sometimes I'd ask them to close their eyes and tell me what I had put into their outstretched hands. Whether it was a car key, a quarter, or a can opener, they often could not even guess what they were holding—their sensory perceptions simply weren't working. [...]
>
> In response to the trauma itself, and in coping with the dread that persisted long afterward, these patients had learned to shut down the brain areas that transmit the visceral feelings and emotions that accompany and define terror. Yet in everyday life, those same brain areas are responsible for registering the entire range of emotions and sensations that form the foundation of our self-awareness, our sense of who we are. What we witnessed here was a tragic adaptation: In an effort to shut off terrifying sensations, they also deadened their capacity to feel fully alive. (van der Kolk, 2014)

Takeaway: emotional healing as a prerequisite for rationality

In this post, I have covered ways in which painful experiences - anything from "big-T Trauma" to mildly unpleasant thoughts - seem to shape our thinking. Everyone has a built-in desire to avoid painful thoughts and experiences, which reinforces cognitive patterns aimed at keeping those kinds of thoughts hidden and buried. Often this is functional, as keeping them suppressed allows us to remain more functional in situations where it would not be useful for old and non-relevant memories to come up, causing fear and avoidance responses when we need to be doing something else.

At the same, these kinds of processes control that which we can think; and as they become automated, they nudge our reasoning to take weird contortions, keeping our belief networks fragmented and our behavior less than coherent, operating in ways that keep themselves hidden. They also limit us with regard to instrumental rationality, as options which we could otherwise have taken - anything from wearing particular kinds of shoes to taking up new careers which challenge our chosen identities - are judged as categorically unacceptable.

To actually change your mind, you need to be able to dig up and address your past traumas. Fortunately, there are various tools available for that purpose.

References

Baldwin, D. V. (2013). Primitive mechanisms of trauma response: an evolutionary perspective on trauma-related disorders. *Neuroscience and Biobehavioral Reviews*, *37*(8), 1549–1566.

Forgash, C., & Copeley, M. (Eds.). (2007). *Healing the Heart of Trauma and Dissociation with EMDR and Ego State Therapy* (1 edition). Springer Publishing Company.

Patihis, L., Ho, L. Y., Tingen, I. W., Lilienfeld, S. O., & Loftus, E. F. (2014). Are the "Memory Wars" Over? A Scientist-Practitioner Gap in Beliefs About Repressed Memory.

Psychological Science, *25*(2), 519–530.

Rofé, Y. (2008). Does Repression Exist? Memory, Pathogenic, Unconscious and Clinical Evidence. *Review of General Psychology: Journal of Division 1, of the American Psychological Association*, *12*(1), 63–85.

Salvador, A., Berkovitch, L., Vinckier, F., Cohen, L., Naccache, L., Dehaene, S., & Gaillard, R. (2018). Unconscious memory suppression. *Cognition*, *180*, 191–199.

Schauer, M., & Elbert, T. (2010). Dissociation Following Traumatic Stress. *Swiss Journal of Psychology: Official Publication of the Swiss Psychological Society Schweizerische Zeitschrift Fur Psychologie = Revue Suisse de Psychologie*, *218*(2), 109–127.

Shadlen, M. N., & Shohamy, D. (2016). Decision Making and Sequential Sampling from Memory. *Neuron*, *90*(5), 927–939.

Shapiro, F. (2017). *Eye Movement Desensitization and Reprocessing (EMDR) Therapy, Third Edition: Basic Principles, Protocols, and Procedures* (Third edition). The Guilford Press.

van der Kolk, B. (2014). *The Body Keeps the Score: Brain, Mind, and Body in the Healing of Trauma* (1 edition). Viking.

System 2 as working-memory augmented System 1 reasoning

The terms System 1 and System 2 were originally coined by the psychologist Keith Stanovich and then popularized by Daniel Kahneman in his book *Thinking, Fast and Slow.* Stanovich noted that a number of fields within psychology had been developing various kinds of theories distinguishing between fast/intuitive on the one hand and slow/deliberative thinking on the other. Often these fields were not aware of each other. The S1/S2 model was offered as a general version of these specific theories, highlighting features of the two modes of thought that tended to appear in all the theories.

Since then, academics have continued to discuss the models. Among other developments, *Stanovich and other authors have discontinued the use of the System 1/System 2 terminology as misleading*, choosing to instead talk about Type 1 and Type 2 processing. In this post, I will build on some of that discussion to argue that Type 2 processing is a *particular way of chaining together the outputs of various subagents using working memory*. Some of the processes involved in this chaining are themselves implemented by particular kinds of subagents.

This post has three purposes:

- Summarize some of the discussion about the dual process model that has taken place in recent years; in particular, the move to abandon the System 1/System 2 terminology.
- Connect the framework of thought that I have been developing in my multi-agent minds sequence with dual-process models.
- Push back on some popular interpretations of S1/S2 theory which I have been seeing on LW and other places, such as ones in which the two systems are viewed as entirely distinct, S1 is viewed as biased and S2 as logical, and ones in which it makes sense to identify more as one system or the other.

Let's start with looking at some criticism of the S1/S2 model endorsed by the person who coined the terms.

What type 1/type 2 processing is not

The terms "System 1 and System 2" suggest just that: two distinct, clearly defined systems with their own distinctive properties and modes of operation. However, there's no single "System 1": rather, a wide variety of different processes and systems are lumped together under this term. It is also unclear whether there is any single System 2, either. As a result, a number of researchers including Stanovich himself have switched to talking about "Type 1" and "Type 2" processing instead (Evans, 2012; Evans & Stanovich, 2013; Pennycook, Neys, Evans, Stanovich, & Thompson, 2018).

What exactly defines Type 1 and Type 2 processing?

A variety of attributes have been commonly attributed to either Type 1 or Type 2 processing. However, one criticism is that there is no empirical or theoretical support

for such attributes to *only* occur with one type of processing. For instance, Melnikoff & Bargh (2018) note that one set of characteristics which has been attributed to Type 1 processing is “efficient, unintentional, uncontrollable, and unconscious”, whereas Type 2 processing has been said to be “inefficient, intentional, controllable and conscious”.

(Before you read on, you might want to take a moment to consider the extent to which this characterization matches your intuition of Type 1 and Type 2 processing. If it does match to some degree, you can try to think of examples which are well-characterized by these types, as well as examples which are not.)

.

.

.

.

.

.

.

.

.

.

They note that this correlation has never been empirically examined, and that there are also various processes in which attributes from both sets co-occur. For example:

- **Unconscious (T1) and Intentional (T2).** A skilled typist can write sentences without needing to consciously monitor their typing, “but will never start plucking away at their keys without intending to type something in the first place.” Many other skills also remain intentional activities even as one gets enough practice to be able to carry them out without conscious control: driving and playing piano are some examples. Also, speaking involves plenty of unconscious processes, as we normally have very little awareness of the various language-production rules that go into our speech. Yet we generally only speak when we intend to.
- **Unconscious (T1) and Inefficient (T2).** Unconscious learning can be less efficient than conscious learning. For example, some tasks can be learned quickly using a verbal rule which describes the solution, or slowly using implicit learning so that we figure out how to do the task but cannot give an explicit rule for it.
- **Uncontrollable (T1) and Intentional (T2).** Consider the bat-and-ball problem: "A bat and a ball cost $1.10 in total. The bat costs $1 more than the ball. How much does the ball cost?" Unless they have heard the problem before, people nearly always generate an initial (incorrect) answer of 10 cents. This initial response is uncontrollable: no experimental manipulation has been found that would cause people to produce any other initial answer, such as 8 cents to 13 cents. At the same time, the process which causes this initial answer to be produced is intentional: "it is not initiated directly by an external stimulus (the

> question itself), but by an internal goal (to answer the question, a goal activated by the experimental task instructions). In other words, reading or hearing the bat-and-ball problem does not elicit the 10 cents output unless one intends to solve the problem."

Regarding the last example, Melnikoff & Bargh note:

> Ironically, this mixture of intentionality and uncontrollability characterizes many of the biases documented in Tversky and Kahneman's classic research program, which is frequently used to justify the classic dual-process typology. Take, for example, the availability heuristic, which involves estimating frequency by the ease with which information comes to mind. In the classic demonstration, individuals estimate that more words begin with the letter K than have K in the third position (despite the fact that the reverse is true) because examples of the former more easily come to mind [107]. This bias is difficult to control – we can hardly resist concluding that more letters start with K than have K in the third position – but again, all of the available evidence suggests that it only occurs in the presence of an intention to make a judgment. The process of generating examples of the two kinds of words is not activated directly by an external stimulus, but by an internal intention to estimate the relative frequencies of the words. Likewise for many judgments and decisions.

They also give examples of what they consider uncontrollable (T1) but inefficient (T2), unintentional (T1) but inefficient (T2), as well as unintentional (T1) but controllable (T2). Further, they discuss each of the four attributes themselves and point out that they all contain various subdimensions. For example, people whose decisions are influenced by unconscious primes are conscious of their decision but not of the influence from the prime, meaning that the process has both conscious and unconscious aspects.

Type 1/Type 2 processing as working memory use

Rather than following the "list of necessary attributes" definition, Evans & Stanovich (2013) distinguish between *defining features* and *typical correlates*. In previous papers, Evans has generally defined Type 2 processing in terms of requiring working memory resources and being able to think hypothetically. On the other hand, Stanovich has focused on what he calls cognitive decoupling, which his work shows is highly correlated with fluid intelligence as the defining feature.

Cognitive decoupling can be defined as the ability to create copies of our mental representations of things, so that the copies can be used in simulations without affecting the original representations. For example, if I see an apple in a tree, my mind has a representation of the apple. If I then imagine various strategies of getting the apple - such as throwing a stone at the tree to knock the apple down - I can mentally simulate what would happen to the apple as a result of my actions. But even as I imagine the apple falling down from the tree, I never end up thinking that I can get the real apple down simply by an act of imagination. This because the mental object representing the real apple is *decoupled* from the apple in my hypothetical scenario. I can manipulate the apple in the hypothetical without those manipulations being passed on to the mental object representing the original apple.

In their joint paper, Evans & Stanovich propose to combine their models and define Type 2 processes as those which use working memory resources (closely connected with fluid intelligence) in order to carry out hypothetical reasoning and cognitive decoupling. In contrast, Type 1 reasoning is anything which does not do that. Various features of thought - such as being automatic and the other controlled - may tend to correlate more with one or the other type, but these are only correlates, not necessary features.

Type 1 process (intuitive)	Type 2 process (reflective)
Defining features	
Does not require working memory	*Requires working memory*
Autonomous	*Cognitive decoupling; mental simulation*
Typical correlates	
Fast	Slow
High capacity	Capacity limited
Parallel	Serial
Nonconscious	Conscious
Biased responses	Normative responses
Contextualized	Abstract
Automatic	Controlled
Associative	Rule-based
Experience-based decision making	Consequential decision making
Independent of cognitive ability	Correlated with cognitive ability

Type 2 processing as composed of Type 1 components

In previous posts of my multi-agent minds sequence, I have been building up a model of mind that is composed of interacting components. How does it fit together with the proposed Type 1/Type 2 model?

Kahneman in *Thinking Fast and Slow* mentions that giving the answer to 2 + 2 = ? is a System (Type) 1 task, whereas calculating 17 * 24 is a System (Type) 2 task. This might be starting to sound familiar. In my post on subagents and neural Turing machines, I discussed Stanislas Dehane's model where you do complex arithmetic by breaking up a calculation into subcomponents which can be done automatically, and then routing the intermediate results through working memory. You could consider this to also involve cognitive decoupling: for instance, if part of how you calculate 17 * 24 is by first noting that you can calculate 10 * 24, you need to keep the original representation of 17 * 24 intact in order to figure out what other steps you need to take.

To me, the calculation of 10 * 24 = 240 happens mostly automatically; like 2 + 2 = 4, it feels like a Type 1 operation rather than a Type 2 one. But what this implies, then, is that we carry out Type 2 arithmetic by *chaining together Type 1 operations through Type 2 working memory.*

I do not think that this is just a special case relating to arithmetic. Rather it seems like an implication of the Evans & Stanovich definition which they do not mention explicitly, but which is nonetheless relatively straightforward to draw: that *Type 2 reasoning is largely built up of Type 1 components.*

Under this interpretation, there are some components which are specifically dedicated to Type 2 processes: things like working memory storages and systems for manipulating their contents. But those components cannot do anything alone. The original input to be stored in working memory originates *from* Type 1 processes (and the act of copying it to working memory decouples it from the original process which produced it), and working memory alone could not do anything without those Type 1 inputs.

Likewise, there may be something like a component which is Type 2 in nature, in that it holds rules for how the contents of working memory should be transformed in different situations - but many of those transformations happen by firing various Type 1 processes which then operate on the contents of the memory. Thus, the rules are about *choosing which Type 1 process to trigger*, and could again do little without those processes. (My post on neural Turing machines explicitly discussed such rules.)

Looking through Kahneman's examples

At this point, you might reasonably suspect that arithmetic reasoning is an example that I cherry-picked to support my argument. To avoid this impression, I'll take the first ten examples of System 2 operations that Kahneman lists in the first chapter of *Thinking, Fast and Slow* and suggest how they could be broken down into Type 1 and Type 2 components.

Kahneman defines System 2 in a slightly different way than we have defined Type 2 operations - he talks about System 2 operations requiring attention - but as attention and working memory are closely related, this still remains compatible with our model. Most of these examples involve somehow focusing attention, and manipulating attention can be understood as manipulating the contents of working memory to ensure that a particular mental object remains in working memory. Modifying the contents of working memory was an important type of production rule discussed in my earlier post.

Starting with the first example in Kahneman's list:

> Brace for the starter gun in a race.

One tries to keep their body in such a position that it will be ready to run when the gun sounds; recognizing the feel of the correct position is a Type 1 operation. Type 2 rules are operating to focus attention on the output of the system which outputs proprioceptive data, allowing Type 1 processes to notice mismatches with the required body position and correct them. Additionally, Type 2 rules are focusing attention on the sound of the gun, so as to more quickly identify the sound when the gun fires (a Type 1 operation), causing the person to start running (also a Type 1 operation).

> Focus attention on the clowns in the circus.

This involves Type 2 rules which focus attention on a particular sensory output, as well as keeping one's eyes physically oriented towards the clowns. This requires detecting when one's attention/eyes are on something else than the clowns and then applying an internal (in the case of attention) or external (in the case of eye position) correction. As Kahneman offers "orient to the source of a sudden sound", "detect hostility in a voice", "read words on large billboards", and "understand simple sentences" as Type 1 operations, we can probably say that recognizing something as a clown or not-clown and moving one's gaze accordingly are Type 1 operations.

Focus on the voice of a particular person in a crowded and noisy room.

As above, Type 2 rules check whether attention is on the voice of that person (a comparison implemented using a Type 1 process), and then adjust focus accordingly.

Look for a woman with white hair.

Similar to the clown example.

Search memory to identify a surprising sound.

It's unclear to me exactly what is going on here. But introspectively, this seems to involve something like keeping the sound in attention so as to feed it to memory processes, and then applying the rule of "whenever the memory system returns results, compare them against the sound and adjust the search based on how relevant they seem". The comparison feels like it is done by something like a Type 1 process.

Maintain a faster walking speed than is natural to you.

Monitor the appropriateness of your behavior in a social situation.

Walking: Similar to the "brace for the starter gun" example, Type 2 rules keep calling for a comparison of your current walking speed with the desired one (a Type 1 operation), passing any corrections resulting from that comparison to the Type 1 system controlling your walking speed.

Social behavior: maintain attention on a conscious representation of what you are doing, checking it against various Type 1 processes which contain rules about appropriate and inappropriate behavior. Adjust or block accordingly.

Count the occurrences of the letter *a* in a page of text.

Focus attention on the letters of a text; when a Type 1 comparison detects the letter "a", increment a working memory counter by one.

Tell someone your phone number.

After a retrieval of the phone number from memory has been initiated, Type 2 rules use Type 1 processes to monitor that it is said in full.

Park in a narrow space (for most people except garage attendants).

Keeping attention focused on what you are doing to allow a series of evaluations, mental simulations, and cached (Type 1) procedural operations determining how to act in response to a particular situation in the parking process.

A general pattern in these examples is that Type 2 processing can maintain attention on something as well as hold the intention to invoke comparisons to use as the basis for behavioral adjustments. As comparisons involve Type 1 processes, Type 2 processing is fundamentally reliant on Type 1 processing to be able to do anything.

Consciousness and dual process theory

Alert readers might have noticed that focusing one's attention on something involves keeping it in consciousness, whereas the previous Evans & Stanovich definition noted that consciousness is not a defining part of the Type 1/Type 2 classification. Is this a contradiction? Probably not, since as remarked previously, different aspects of the same process may be conscious and unconscious at the same time.

For example, if one intends to say something, one may be conscious of the intention while the actual speech production happens unconsciously; once they say it and they hear their own words, an evaluation process can run unconsciously but output its results into consciousness. With "conscious" being so multidimensional, it doesn't seem like a good defining characteristic to use, even if some aspects of it did very strongly correlate with Type 2 processing.

Evans (2012) writes in a manner which seems to me compatible with the notion of there being many different kinds of Type 2 processing, with different processing resources being combined according to different rules as the situation warrants:

> The evidence suggests that there is not even a single type 2 system for reasoning, as different reasoning tasks recruit a wide variety of brain regions, according to the exact demands of the task [...].
>
> I think of type 2 systems as ad hoc committees that are put together to deal with a particular problem and then disbanded when the task is completed. Reasoning with abstract and belief-laden syllogisms, for example, recruits different resources, as the neural imaging data indicate: Only the latter involve semantic processing regions of the brain. It is also a fallacy to think of "System 2" as a conscious mind that is choosing its own applications. The ad hoc committee must be put together by some rapid and preconscious process—any feeling that "we" are willing and choosing the course of our thoughts and actions is an illusion [...]. I therefore also take issue with dual-process theorists [...] who assign to System 2 not only the capacity for rule-based reasoning but also an overall executive role that allows it to decide whether to intervene upon or overrule a System 1 intuition. In fact, recent evidence suggests that while people's brains detect conflict in dual-process paradigms, the conscious person does not.

If you read my neural Turing machines post, you may recall that I noted that the rules which choose what becomes conscious operate below the level of conscious awareness. We may have the subjective experience of being able to choose what thoughts we think, but this is a post-hoc interpretation rather than a fact about the process.

Type 1/Type 2 and bias

People sometimes refer to Type 1 reasoning as biased, and to Type 2 reasoning as unbiased. But as this discussion should suggest, there is nothing that makes one of the two types intrinsically more or less biased than the other. The bias-correction power of Type 2 processing emerges from the fact that *if* Type 1 operations are known to be erroneous *and* a rule-based procedure for correcting them exists, a Type 2 operation can be learned which implements that rule.

For example, someone familiar with the substitution principle may know that their initial answer to a question like "how popular will the president be six months from now?" comes from a Type 1 process which *actually* answered the question of "how popular is the president right now?".

They may then have a Type 2 rule saying something like "when you notice that the question you were asked is subject to substitution effects, replace the initial answer with one derived from a particular procedure". But this still requires a) a Type 1 process recognizing the situation as one where the rule should be applied b) knowing a procedure which provides a better answer c) the cue-procedure rule having been installed previously, itself a process requiring a number of Type 1 evaluations (about e.g. how rewarding it would be to have such a rule in place).

There is nothing to say that somebody couldn't learn an outright wrong Type 2 rule, such as "whenever you think of 2+2 = 4, substitute your initial answer of '4' with a '5'".

Often, it is also unclear of what the better Type 2 rule even *should* be. For instance, another common substitution effect is that when someone is asked "How happy are you with your life these days?", they actually answer the question of "What is my mood right now?". But what *is* the objectively correct procedure for evaluating your current happiness with life?

On the topic of Type 1/2 and bias, I give the final word to Evans (2012):

> One of the most important fallacies to have arisen in dual-process research is the belief that the normativity of an answer [...] is diagnostic of the type of processing. Given the history of the dual-process theory of reasoning, one can easily see how this came about. In earlier writing, heuristic or type 1 processes were always the "bad guys," responsible for cognitive biases [...]. In belief bias research, authors often talked about the conflict between "logic" and "belief," which are actually dual sources, rather than dual processes. Evans and Over [...] defined "rationality2" as a form of well-justified and explicit rule-based reasoning that could only be achieved by type 2 processes. Stanovich [...] in his earlier reviews of his psychometric research program emphasized the association between high cognitive ability, type 2 processing and normative responding. Similarly, Kahneman and Frederick [...] associate the heuristics of Tversky and Kahneman with System 1 and successful reasoning to achieve normatively correct solutions to the intervention of System 2.
>
> The problem is that a normative system is an externally imposed, philosophical criterion that can have no direct role in the psychological definition of a type 2 process. [...] if type 2 processes are those that manipulate explicit representations through working memory, why should such reasoning necessarily be normatively correct? People may apply the wrong rules or make errors in their application. And why should type 1 processes that operate automatically and without reflection necessarily be wrong? In fact, there is much evidence that expert decision making

> can often be well served by intuitive rather than reflective thinking [...] and that sometimes explicit efforts to reason can result in worse performance [...].
>
> Reasoning research somewhat loads the dice in favor of type 2 processing by focusing on abstract, novel problems presented to participants without relevant expertise. If a sports fan with much experience of following games is asked to predict results, he or she may be able to do so quite well without need for reflective reasoning. However, a participant in a reasoning experiment is generally asked to do novel things, like assuming some dubious propositions to be true and deciding whether a conclusion necessarily follows from them. In these circumstances, explicit type 2 reasoning is usually necessary for correct solution, but certainly not sufficient. Arguably, however, when prior experience provides appropriate pragmatic cues, even an intractable problem like the Wason selection task becomes easy to solve [...], as this can be done with type 1 processes [...]. It is when normative performance requires the deliberate suppression of unhelpful pragmatic cues that higher ability participants perform better under strict deductive reasoning instructions [...].
>
> Hence, [the fallacy that type 1 processes are responsible for cognitive biases and type 2 processes for normatively correct reasoning] is with us for some fairly precise historical reasons. In the traditional paradigms, researchers presented participants with hard, novel problems for which they lacked experience (students of logic being traditionally excluded), and also with cues that prompted type 1 processes to compete or conflict with these correct answers. So in these paradigms, it does seem that type 2 processing is at least necessary to solve the problems, and that type 1 processes are often responsible for cognitive biases. But this perspective is far too narrow, as has recently been recognized. In recent writing, I have attributed responsibility for a range of cognitive biases roughly equally between type 1 and type 2 processing [...]. Stanovich [...] similarly identifies a number of reasons for error other than a failure to intervene with type 2 reasoning; for example, people may reason in a quick and sloppy (but type 2) manner or lack the necessary “mindware” for successful reasoning.

Summary and connection to the multiagent models of mind sequence

In this post, I have summarized some recent-ish academic discussion on dual-process models of thought, or what used to be called System 1 and System 2. I noted that the popular conception of them as two entirely distinct systems with very different properties is mistaken. While there is a defining difference between them - namely, the use of working memory resources to support hypothetical thinking and cognitive decoupling - they seem to rather refer to differences in two *types* of thought, either of which may use very different kinds of systems.

It is worth noting at this point that there are many different dual-process models in different parts of psychology. The Evans & Stanovich model which I have been discussing here is intended as a generalized model of them, but as they themselves (2013) write:

> … we defend our view that the Type 1 and 2 distinction is supported by a wide range of converging evidence. However, we emphasize that not all dual-process

> theories are the same, and we will not act as universal apologists on each one's behalf. Even within our specialized domain of reasoning and decision making, there are important distinctions between accounts. S. A. Sloman [...], for example, proposed an architecture that has a parallel-competitive form. That is, Sloman's theories and others of similar structure [...] assume that Type 1 and 2 processing proceed in parallel, each having their say with conflict resolved if necessary. In contrast, our own theories [...] are default-interventionist in structure [...]. Default-interventionist theories assume that fast Type 1 processing generates intuitive default responses on which subsequent reflective Type 2 processing may or may not intervene.

In previous posts of the multi-agent models of mind sequence, I have been building up a model of the mind being built up of a variety of subsystems (which might in some contexts be called subagents).

In my discussion of Consciousness and the Brain, I summarized some of its conclusions as saying that:

- The brain has multiple subagents doing different things; many of the subagents do unconscious processing of information. When a mental object becomes conscious, many subagents will synchronize their processing around analyzing and manipulating that mental object.
- The collective of subagents can only have their joint attention focused on one mental object at a time.
- The brain can be compared to a production system, with a large number of subagents carrying out various tasks when they see the kinds of mental objects that they care about. E.g. when doing mental arithmetic, applying the right sequence of mental operations for achieving the main goal.

In Building up to an Internal Family Systems model, I used this foundation to discuss the IFS model of how various subagents manipulate consciousness in order to achieve various kinds of behavior. In Subagents, neural Turing machines, thought selection, and blindspots, I talked about the mechanistic underpinnings of this model and how processes like thought selection and firing of production rules might actually be implemented.

What had been lacking so far was a connection between these models and the Type 1/Type 2 typology. However, if we take something like the Evans & Stanovich model of Type 1/Type 2 processing to be true, then it turns out that our discussion has been connected with their model all along. Already in "Consciousness and the Brain", I mentioned the "neural Turing machine" passing on results from one subsystem to another through working memory. That, it turns out, is the defining characteristic of Type 2 processing - with Type 1 processing simply being any process which does *not* do that.

Under this model, then, Type 2 processing is a *particular way of chaining together the outputs of various Type 1 subagents using working memory*. Some of the processes involved in this chaining are themselves implemented by particular kinds of subagents.

References

Evans, J. S. B. T. (2012). Dual process theories of deductive reasoning: facts and fallacies. *The Oxford Handbook of Thinking and Reasoning*, 115–133.

Evans, J. S. B. T., & Stanovich, K. E. (2013). Dual-Process Theories of Higher Cognition: Advancing the Debate. *Perspectives on Psychological Science: A Journal of the Association for Psychological Science*, *8*(3), 223–241.

Melnikoff, D. E., & Bargh, J. A. (2018). The Mythical Number Two. *Trends in Cognitive Sciences*, *22*(4), 280–293.

Pennycook, G., Neys, W. D., Evans, J. S. B. T., Stanovich, K. E., & Thompson, V. A. (2018). The Mythical Dual-Process Typology. *Trends in Cognitive Sciences*, *22*(8), 667–668.

Book summary: Unlocking the Emotional Brain

If the thesis in *Unlocking the Emotional Brain* (UtEB) is even half-right, it may be one of the most important books that I have read. Written by the psychotherapists Bruce Ecker, Robin Ticic and Laurel Hulley, it claims to offer a neuroscience-grounded, comprehensive model of how effective therapy works. In so doing, it also happens to formulate its theory in terms of belief updating, helping explain how the brain models the world and what kinds of techniques allow us to actually change our minds. Furthermore, if UtEB is correct, it also explains why rationalist techniques such as Internal Double Crux [1 2 3] work.

UtEB's premise is that much if not most of our behavior is driven by emotional learning. Intense emotions generate unconscious predictive models of how the world functions and what caused those emotions to occur. The brain then uses those models to guide our future behavior. Emotional issues and seemingly irrational behaviors are generated from implicit world-models (schemas) which have been formed in response to various external challenges. Each schema contains memories relating to times when the challenge has been encountered and mental structures describing both the problem and a solution to it.

According to the authors, the key for updating such schemas involves a process of memory reconsolidation, originally identified in neuroscience. The emotional brain's learnings are usually locked and not modifiable. However, once an emotional schema is activated, it is possible to simultaneously bring into awareness knowledge contradicting the active schema. When this happens, the information contained in the schema can be overwritten by the new knowledge.

While I am not convinced that the authors are *entirely* right, many of the book's claims definitely feel like they are pointing in the right direction. I will discuss some of my caveats and reservations after summarizing some of the book's claims in general. I also consider its model in the light of an issue of a psychology/cognitive science journal devoted to discussing a very similar hypothesis.

Emotional learning

In UtEB's model, emotional learning forms the foundation of much of our behavior. It sets our basic understanding about what situations are safe or unsafe, desirable or undesirable. The authors do not quite say it explicitly, but the general feeling I get is that the subcortical emotional processes set many of the priorities for what we want to achieve, with higher cognitive functions then trying to figure out *how* to achieve it - often remaining unaware of what exactly they are doing.

UtEB's first detailed example of an emotional schema comes from the case study of a man in his thirties they call Richard. He had been consistently successful and admired at work, but still suffered from serious self-doubt and low confidence at his job. On occasions such as daily technical meetings, when he considered saying something, he experienced thoughts including "Who am I to think I know what's right?", "This could be wrong" and "Watch out - don't go out on a limb". These prevented him from expressing any opinions.

From the point of view of the authors, these thoughts have a definite cause - Richard has “emotional learnings according to which it is adaptively *necessary* to go into negative thoughts and feelings towards [himself].” The self-doubts are a *strategy* which his emotional brain has generated for solving some particular problem.

Richard’s therapist guided Richard to imagine what it would feel like if he was at one of his work meetings, made useful comments, and felt confident in his knowledge while doing so. This was intended to elicit information about what Richard’s emotional brain predicted would happen if it failed to maintain the strategy of self-doubt. The book includes the following transcript of what happened after Richard started imagining the scene as instructed:

> **Richard:** Now I’m feeling really uncomfortable, but-it’s in a different way.
>
> **Therapist:** OK, let yourself feel it - this different discomfort. *[Pause.]* See if any words come along with this uncomfortable feeling.
>
> **Richard:** *[Pause.]* Now they hate me.
>
> **Therapist:** “Now they hate me.” Good. Keep going: See if this really uncomfortable feeling can also tell you *why* they hate you now.
>
> **Richard:** *[Pause.]* Hnh. Wow. It’s because… now I’m… an arrogant asshole… like my father… a totally self-centered, totally insensitive know-it-all.
>
> **Therapist:** Do you mean that having a feeling of confidence as you speak turns you into an arrogant asshole, like Dad?
>
> **Richard:** Yeah, exactly. Wow.
>
> **Therapist:** And how do you *feel* about being like him in this way?
>
> **Richard:** It’s horrible! It’s what I’ve always vowed *not* to be!

Richard had experienced his father as being assertive as well as obnoxious and hated. His emotional brain had identified this as a failure mode to be avoided: if you are assertive, then you are obnoxious and will be hated. The solution was to generate feelings of doubt so as to stop him from being too confident. This caused him suffering, but the prediction of his emotional brain was that acting otherwise would produce even *worse* suffering, as being hated would be a terrible fate.

UtEB describes Richard as having had the following kind of unconscious schema:

> ***Perceptual, emotional and somatic memory of original experiences:*** *his suffering from his father’s heavily dominating, hyper-confident self-expression, plus related suffering from unmet needs for fatherly expressions of love, acceptance, understanding, validation. (This is the “raw data”; matching features in current situations are triggers of the whole schema.)*
>
> ***A mental model or set of linked, learned constructs operating as living knowledge of a problem and a solution:***
>
> ***The problem: knowledge of a vulnerability to a specific suffering.*** *Confident assertiveness in any degree inflicts crushing oppression on others and is hated by them. I would be horrible like Dad and hated by others, as he is, if I*

asserted my own knowledge or wishes confidently. (This is a model of how the world is, and current situations that appear relevant to this model are triggers of the whole schema.)

The solution: knowledge of an urgent broad strategy and concrete tactic(s) for avoiding that suffering. *Never express any confident assertiveness, to avoid being horrible and hated (general strategy and pro-symptom purpose), by vigilantly noticing any definite knowledge or opinions forming in myself and blocking them from expression by generating potently self-doubting, self-invalidating thoughts (concrete tactic and manifested symptom).*

Emotional schemas can be brought to light during a variety of ways, including Focusing, IFS, and imagining yourself doing something and seeing what you expect to happen as a result.

But suppose that you do manage to bring up a schema which seems wrong to you. What do you do then?

Memory reconsolidation: updating the emotional learning

The formation of memory traces involves *consolidation*, when the memory is first laid out in the brain; *deconsolidation*, when an established memory is "opened" and becomes available for changes; and *reconsolidation*, when a deconsolidated memory (along with possible changes) is stored and becomes frozen again. The term "reconsolidation" is also used to refer to the general process from deconsolidation to reconsolidation; UtEB generally applies the term to mean the entire process. Unless the context indicates otherwise, I do the same.

UtEB reviews some of the history of memory research. Until 1997, neuroscientists believed that past emotional learning became permanently locked in the brain, so that memories could only consolidate, never de- or reconsolidate. More recent research has indicated that once a memory becomes activated, it is temporarily unlocked, allowing it to be changed or erased.

Starting from 2004, new studies suggested that activation alone is not sufficient to deconsolidate the memory. The memories are used to predict that things will occur in a similar fashion as they did previously. Besides just activation, there has to be a significant *mismatch* between what one experiences and what the memory suggests is about to happen. The violation of expectation can be qualitative (the predicted outcome not occurring at all) or quantitative (the magnitude of the outcome not being fully predicted). In either case, it is this prediction error which triggers the deconsolidation and subsequent reconsolidation.

The memory erasure seems to be specific to the interpretation from which the prediction was produced. For example, someone who has had an experience of being disliked may later experience being liked. This may erase the emotional generalization "I am inherently dislikable", but it will not erase the memory of the person also having been disliked.

Applied reconsolidation: an example of the schema update process

So, assuming that the model outlined above is correct, how does one apply it in practice?

From what we have discussed so far, the essential steps of erasing a learned belief (including an emotional schema) involves identifying it, activating it, and then finding a mismatch between its prediction and reality.

The first difficulty is that the beliefs involved with the schema are not necessarily consciously available at first. Richard knew that he suffered from a lack of self-esteem, but he was not aware of its reason. The process started from him describing in concrete details how this manifested: as skeptical self-talk during a daily meeting.

As he was guided to imagine what would happen if he didn't have those thoughts and acted confidently, his therapist was seeking to retrieve the implicit schema and bring it into consciousness so that its contents would become available for access. Once it had been retrieved, the therapist and Richard worked together to express the belief in the schema in maximally emotional language:

> "Feeling *any* confidence means I'm arrogant, self-centered, and totally insensitive like Dad, and people will hate me for it, so I've got to *never* feel confident, ever."

The authors have developed a therapeutic approach called Coherence Therapy, whose steps closely follow the steps of the memory reconsolidation process. The example of Richard is from this school of therapy.

In Coherence Therapy (as well as related approaches, such as Internal Family Systems), one initially avoids any impulse to argue with or disprove the retrieved schema. This would risk it being pushed away *before* it has become sufficiently activated to allow for reconsolidation.

Instead, one stays with it. Richard was given a card with the above phrase and instructed to review it every day until the next therapy session, just feeling the extent to which it felt true to him. This served to further integrate access to the schema in question, making it better consciously available.

Two weeks later, Richard had frequently noticed his self-doubt, used it as a prompt for reading the card, and experienced its description ringing true as a reason for his thoughts. When speaking with his therapist, he mentioned a particular event which had stuck in his memory. In a recent meeting, he had thought of a solution to a particular problem, but then kept quiet about it. A moment later, another person had spoken up and suggested the same solution in a confident manner. Looking around, Richard had seen the person's solution and confidence being received positively by the others. Richard had been struck by how that reaction differed from what his schema predicted would happen if *he* had made the same suggestion in that tone.

Because Richard had made the implicit assumptions in his schema explicit, he was able to consciously notice a situation which seemed to violate those assumptions: a prediction mismatch. His therapist recognized this as a piece of contradictory knowledge which could be used to update the old schema. The therapist then guided

Richard through a process intended to activate the old schema while bringing the contradictory information into awareness, triggering a reconsolidation process.

The therapist first instructed Richard to mentally bring himself back to the situation where he had just thought of the solution, but held it back. To properly activate the schema, the therapist guided Richard's attention to the purpose behind his reluctance and Richard's certainty of any confidence making him disliked. Next, the therapist told Richard to re-live what happened next: the other person making the same suggestion, and the other people in the room looking pleased rather than angry.

The book then has a transcript of the therapist guiding Richard to repeat this juxtaposition of the old schema and the disconfirming experience (italicized brackets in the original):

> **Therapist:** Stay with that. Stay with being surprised at what you're seeing—surprised because in your life, you've had such a definite knowing that saying something confidently to people will always come across like Dad, like an obnoxious know-it-all, and people will hate that. That's what you know, yet at the same time, here you're seeing that saying something confidently *isn't* always like Dad, and then people are fine with it. And it's quite a surprise to know that. *[That was an explicit prompting of another side-by-side experiencing of the two incompatible knowings, with the therapist expressing empathy for both, with no indication of any favoring of one knowing over the other. The therapist paused for several seconds, then asked:]* Does it feel true to describe it like that? Your old knowing right alongside this other new knowing that's so different?
>
> **Richard:** *[Quietly, seeming absorbed in the experience.]* Yeah.
>
> **Therapist:** *[Softly.]* All along, it seemed to you that saying something confidently could be done only in Dad's dominating way of doing it, and now suddenly you're seeing that saying something confidently can be done very differently, and it feels fine to people. *[This was another deliberate repetition of the same juxtaposition experience.]*
>
> **Richard:** Yeah.
>
> **Therapist:** Mm-hm. *[Silence for about 20 seconds.]* So, how is it for you be in touch with both of these knowings, the old one telling you that anything said with confidence means being like Dad, and the new one that knows you can be confident in a way that feels okay to people? *[Asking this question repeated the juxtaposition experience yet again, and, in addition, the "how is it" portion of the question prompted Richard to view the experience with mindful or metacognitive awareness, while remaining in the experience.]*
>
> **Richard:** It's sort of weird. It's like there's this part of the world that I didn't notice before, even though it's been right there.
>
> **Therapist:** I'm intrigued by how you put that. Sounds like a significant shift for you.
>
> **Richard:** Yeah, it is. Huh.
>
> **Therapist:** You're seeing both now, the old part of the world and this other part of the world that's new, even though it was right there all along. *[That cued the juxtaposition experience for a fourth time, followed by silence for about 30*

seconds.] So, keep seeing both, the old part and the new part, when you open your eyes in a few seconds and come back into the room with me. *[Richard soon opens his eyes and blinks a few times.]* Can you keep seeing both?

Richard: Yeah.

Therapist: What's it like to see both and feel both now? *[With the transformation sequence complete, this question begins the next step of verification— Step V— because it probes for whether the target learning still exists as an emotional experience.]*

Richard: *[Pause, then sudden, gleeful laughter.]* It's kind of funny! Like, what? How could I think that? *[This is an initial marker indicating that the pro-symptom schema may have been successfully disconfirmed, depotentiated, and dissolved by the transformation sequence.]*

Therapist: Do you mean, how could you think that simply saying what you know, or mentioning some good idea that you've had, would make you seem arrogant, insensitive and dominating like Dad and be hated for it?

Richard: *[Laughing again.]* Yeah!

Afterwards, the therapist and Richard wrote a new card together, which Richard was told to review daily:

> All along it's been so clear that if I confidently say what I know, I will always come across as arrogant, insensitive, and dominating like Dad, and be hated for it. And it's so weird, looking around the room and seeing that it doesn't come across like that.

The purpose of the card was to provide additional juxtaposition experiences between the old schema and the new knowledge. While the original transformation sequence might have been enough to eliminate the old schema, the schema might also have been stored in the context of many different situations and contained in several memory systems. In such a situation, further juxtapositions would have helped deal with it.

In a follow-up meeting, Richard reported having lost the feelings of self-doubt, and that speaking up no longer felt like it was any big deal. To verify that the old schema really had lost its power, the therapist tried deliberately provoking his old fears again:

> Dropping his voice to a quieter tone, the therapist added, "But tell me, when you have something to say and just say it, what about the danger of coming across as a know-it-all, like Dad, and being hated for that? What about your fear of that and how urgent it is to protect yourself from that?" […]
>
> Richard took in the question, gazed at the therapist in silence for a few seconds, and then replied, "Well, I don't know what to tell you. All I can say is, that doesn't trouble me any more. And hearing you say it, it seems a little strange that it ever did—like, what was my problem?"

Applied reconsolidation: the schema update process in general

Now that we have looked at a specific example, we can look at a more general version of the process.

Accessing sequence

In Coherence Therapy, the *accessing sequence* is the preliminary phase of making both a person's implicit schema and some disconfirming knowledge accessible, so that they can be used in the juxtaposition process:

1. *Symptom identification.* Establishing which specific symptoms the person regards as problematic, and when and where they manifest. In Richard's case, the general symptom was a lack of confidence, which specifically manifested as negative self-talk in meetings.
2. *Retrieval of target learning.* Bringing into explicit awareness the purpose behind the symptoms. This can then be used to guide the search for disconfirming knowledge, as well as accessing the original schema in order to reconsolidate it. In Richard's case, the purpose was to avoid expressing confidence in a way that would make people hate him.
3. *Identification of disconfirming knowledge.* Identifying some past or present experience which directly contradicts the original learning. This knowledge does not necessarily need to feel "better" or "more positive" than the old one, just as long as it is mutually exclusive with the old one. In Richard's case, the disconfirming knowledge was the experience of his co-worker confidently proposing a solution and being well-received.

Erasure sequence

Once both the target schema and the disconfirming knowledge are known, the erasure steps can be applied to update the learning:

1. *Reactivation of the target schema.* Tapping into the felt truth of the original learning, experiencing it as vividly as possible.
2. *Activation of disconfirming knowledge, mismatching the target schema.* Activating, at the same time, the contradictory belief and having the experience of simultaneously believing in two different things which cannot both be true.
3. *Repetitions of the target-disconfirmation pairing.*

Something that the authors emphasize is that when the target schema is activated, there should be no attempt to explicitly argue against it or disprove it, as this risks pushing it down. Rather, the belief update happens when one experiences their old schema as vividly true, while *also* experiencing an entirely opposite belief as vividly true. It is the juxtaposition of believing X and not-X at the same time, which triggers an inbuilt contradiction-detection mechanism in the brain and forces a restructuring of one's belief system to eliminate the inconsistency.

The book notes that this distinguishes Coherence Therapy from approaches such as Cognitive Behavioral Therapy, which is premised on treating some beliefs as intrinsically irrational and then seeking to disprove them. While UtEB does not go further into the comparison, I note that this is a common complaint that I have heard of CBT: that by defaulting to negative emotions being caused by belief distortions, CBT risks belittling those negative emotions which are actually produced by *correct* evaluations of the world.

I would personally add that not only does treating all of your beliefs - including emotional ones - as provisionally valid seem to be a requirement for actually updating them, this approach is also good rationality practice. After all, you can only seek evidence to test a theory, not confirm it.

If you notice different parts of your mind having conflicting models of how the world works, the correct epistemic stance should be that you are trying to figure out which one is true - not privileging one of them as "more rational" and trying to disprove the other. Otherwise it will be unavoidable that your preconception will cause you to dismiss as false beliefs which are actually true. (Of course, you can still reasonably *anticipate* the belief update going a particular way - but you need to take seriously at least the *possibility* that you will be shown wrong.)

This can actually be a relief. Trying to stack the deck towards receiving favorable evidence would just also sabotage the brain's belief update process. So you might as well give up trying to do so, relax, and just let the evidence come in.

I speculate that this limitation might also be in place in part to help avoid the error where you decide which one of two models is more correct, and then discard the other model entirely. Simultaneously running two contradictory schemas at the same time allows both of them to be properly evaluated and merged rather than one of them being thrown away outright. I suspect that in Richard's case, the resulting process didn't cause him to entirely discard the notion that some behaviors will make him hated like his dad was - it just removed the overgeneralization which had been produced by having too little training data as the basis of the schema.

Of course, this means that there does need to be some contradictory information available which *could* be used to disprove the original schema. One might have a schema for which no disconfirmation is available because it is correct, or a schema which might or might not be correct but which is making things worse and cannot easily be disconfirmed. UtEB mentions the example of a man, "Tómas", who had a desire to be understood and validated by someone important in his life. Tómas remarked that a professional therapist who was being paid for his empathy could never fulfill that role. The update contradicting the schema that nobody in his life really understood him, would have to come from someone actually in his life.

Another issue that may pop up with the erasure sequence is that there is another schema which predicts that, for whatever reason, running this transformation may produce adverse effects. In that case, one needs to address the objecting schema first, essentially be carrying out the entire process on it before returning to the original steps. (This is similar to the phenomenon in e.g. Internal Family Systems, where objecting parts may show up and have their concerns addressed before work on the original part can proceed.)

Verification step

Finally, after the erasure sequence has been run, one seeks to verify that lasting change has indeed happened and that the target schema has been transformed. UtEB offers the following behavioral markers as signs that a learning which has previously generated emotional responses has in fact been erased:

- "A specific emotional reaction abruptly can no longer by reactivated by cues and triggers that formerly did so or by other stressful situations."

- “Symptoms of behavior, emotion, somatics, or thought that were expressions of that emotional reaction also disappear permanently.”
- “Non-recurrence of the emotional reaction and symptoms continues effortlessly and without counteractive or preventive measures of any kind.”

The authors interpret current neuroscience to say that only memory reconsolidation can produce these kinds of markers. They cannot be produced by counteractive or competitive processes, such as trying to learn an opposite habit to replace a neurotic behavior. Counteractive processes are generally fragile and susceptible to relapse. When these markers are observed in clinical work, UtEB argues that one may infer that reconsolidation has led to the original learning being replaced.

Further examples

For additional examples of the schema update process, I recommend reading the book, which contains several more case studies of issues which were dealt with using this approach. Here’s a brief summary of the most detailed ones (note that some of these examples are actually more detailed and include additional complications, such as more than one symptom-producing schema; I have only summarized the most prominent ones to give a taste of them):

- “Charlotte”. *Issue*: obsessive attachment to a former lover. *Schema:* “It would be much better if I was merged with my lover”. *Contradictory knowledge:* The harm caused by not having boundaries.
- “Ted”. *Issue:* an inability to hold a steady job and a general lack of success in life. *Schema: “*If my life is a mess, my father will be forced to admit how badly he screwed up as a parent.” *Contradictory knowledge:* Realizing that Ted’s father would never admit failure, no matter what.
- “Brenda”. *Issue:* stage fright when having a leading role in an upcoming play. *Schema:* being on the stage in front of an audience means being unable to get off, causing helplessness similar to when Brenda was in a car with her alcoholic father and couldn’t get off. *Contradictory knowledge:* re-imagining the scene and the way how Brenda *could* actually have gotten out of the car.
- “Travis”. *Issue:* inability to experience intimate emotional closeness in relationships. *Schema:* “Nobody will pay attention to how I feel or give me understanding for how I’m hurting. I don’t matter, and I’m all on my own.” *Contradictory knowledge:* the therapist’s empathic presence and listening.
- “Regina”. *Issue:* strong anxiety and panic during/after interacting with other people. *Schema:* “I’m acceptable and lovable only if I do everything *perfectly*.” *Contradictory knowledge:* Regina’s Uncle Theo loves her regardless of what imperfections she might have.
- “Carol”. *Issue:* wanting to avoid sex with her husband despite feeling emotionally close to him. *Schema:* engaging in any sexuality means being overtly sexual and harming Carol’s daughter, in the way that Carol was harmed by her mother’s overt sexuality. *Contradictory knowledge:* once the schema was made conscious, it activated the brain’s spontaneous mismatch detection mechanisms and started to feel silly.

As the last item suggests, sometimes just making a schema explicit is enough to start to dismantle it. The authors suggest that the brain has a built-in detection system which compares any consciously experienced beliefs for inconsistencies with other things that a person knows, and can spontaneously create juxtaposition experiences by bringing up such inconsistent information. They suggest that therapies which are

based on digging up previously unconscious material, but which do not have an explicit juxtaposition step, work to the extent that the uncovered material happens to trigger this spontaneous mismatch detection. (We already saw this happening with Richard - once his underlying schema had been made conscious, he was startled to later notice what seemed like a contradiction.)

One may note the connection to the model in *Consciousness and the Brain* that when some subsystem in the brain manages to elevate a mental object into the content of consciousness, multiple subsystems will synchronize their processing around that object. If the object is an explicit belief, then any subsystem which is paying attention to that object may presumably detect inconsistencies with that subsystem's own models.

Besides these case studies from Coherence Therapy, the authors also analyze published case studies from Accelerated Experiental Dynamic Psychotherapy, Emotion-Focused Therapy, Eye-Movement Desensitization and Reprocessing, and Interpersonal Neurobiology. They try to show how these cases also carried out a juxtaposition process, even if the theoretical frameworks of those therapies did not explicitly realize it. It is the claim of the authors that any therapy which causes lasting emotional change does it through reconsolidation. Finally, the book contains four essays from other therapists (using Coherence Therapy and EMDR), who analyze some of their own case studies.

Evaluating the book's plausibility

Now that we have looked at the book's claims, let's look at whether we should believe in them.

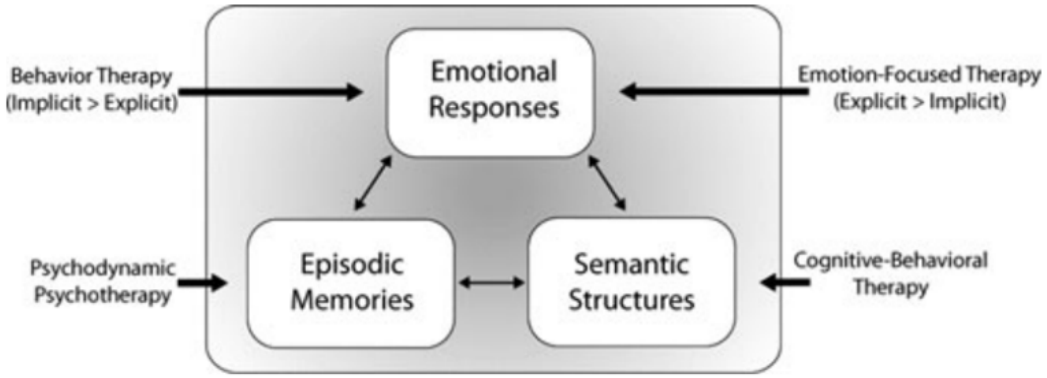
It is unclear to me how reliable the neuroscience results are; the authors cite a number of studies, but each individual claim only references a relatively small number of them.

On a brief look, I could not find any reviews or papers that would have directly made a critical assessment of the book's model. However, I found something that might be even better.

Behavioral and Brain Sciences is a respected journal covering subject areas across the cognitive sciences. BBS publishes "target articles" which present some kind of a thesis or review about a particular topic, together with tens of brief commentaries which respond to the target article, and a final response by the target article's authors to the commentaries.

In 2015, four prestigious (with a total of 500 published research articles between them) psychologists published a BBS target article, *Memory reconsolidation, emotional arousal, and the process of change in psychotherapy: New insights from brain science* (Lane et al. 2015). While the exact model that they outline has a number of differences from the UtEB model, the core idea is the same: that therapeutic change from a wide variety of therapeutic approaches, "including behavioral therapy, cognitive-behavioral therapy, emotion-focused therapy, and psychodynamic psychotherapy, results from the updating of prior emotional memories through a process of reconsolidation that incorporates new emotional experiences."

One interesting difference was that Lane et al. describe emotional schemas somewhat differently. In their model, the schemas form memory structures with three mutually integrated components: emotional responses, episodic/autobiographical memories, and semantic structures (e.g. abstract beliefs which generalize over the various incidents, such as the claim that “people are untrustworthy”). Any of these components can be used as an entry point to the memory structure, and can potentially update the other components through reconsolidation. They hypothesize that different forms of therapy work by accessing different types of components: e.g. behavior therapy and emotion-focused therapy access emotional responses, conventional psychoanalysis uses access to biographical memories, and cognitive behavioral therapy accesses semantic structures.



I read the target article, all the commentaries, and the responses. Given the similarities between Lane et al.’s model and the UtEB model, I think we can consider the responses to Lane et al. to generally offer a useful evaluation of the UtEB model as well.

One significant difference which needs to be noted is that Lane et al.’s model of memory reconsolidation does not mention the requirement for a prediction mismatch before reconsolidation can happen. This was remarked on in the response from UtEB’s authors. In their counter-response, Lane et al. noted UtEB’s model to be highly compatible with theirs, and remarked that further research is needed to nail down the conditions which make reconsolidation the most effective.

The other responses to Lane et al. were mostly from psychologists, psychiatrists, and neuroscientists, but also included the occasional economist, philosopher, philologist and folklorist. Several of the responses were generally positive and mostly wanted to contribute additional details or point out future research directions.

However, there were also a number of skeptical responses. A common theme which emerged from several concerned the limitations of the current neuroscience research on memory reconsolidation. In particular, most of the studies so far have been carried out on rats, and specifically testing the elimination of a fear response to electric shocks. As one of the responses points out, “neither the stimuli nor the subjects are generalizable to the kind of rich autobiographical memories involved in therapy.” A number also raised the question whether all therapeutic change really involves reconsolidation, as opposed to some related mechanism, such as creating new memory structures which compete with the original as opposed to replacing it.

My non-expert reading is that the critical responses are right in that a gap remains between the clinical and behavioral findings on the other hand, and the neuroscience

findings on the other. There are various patterns which can be derived from psychological research and clinical therapy experience, and a small number of neuroscience findings which establish the existence of something that *could* explain those patterns. However, the neuroscience findings have only been established in a rather narrow and limited context; the connection between them and the higher-level patterns is a plausible link, but it remains speculative nonetheless.

Personally I consider the book's model tentatively promising, because it seems to explain many observations which I had independently arrived at *before* reading it. For example, I had noticed an interesting thing with anxieties, where I let e.g. a sensation of social anxiety stay active in my mind, neither accepting it as truth *nor* pushing it away while I went to do social things. This would then cause the anxiety to update, making me feel less anxious if it was indeed the case that the social interaction was harmless. This fits nicely together with the framework of an activated memory structure becoming open to reconsolidation and then being updated by a prediction mismatch (the situation not being as bad as expected).

Likewise, in my post Integrating disagreeing subagents, I reviewed a variety of rationality and therapeutic techniques, and suggested that they mostly worked either by merging or combining two existing models that a person's brain already had, or augmenting the existing models by collecting additional information.

In particular, I considered an example from cognitive behavioral therapy, where a man named Walter was feeling like he was impossible to be in a relationship with after he had broken up with his boyfriend. At the same time, he did not think that someone *else* breaking up with their partner was an indication of them being impossible to be in a relationship with. He and his therapist role-played an interaction where the therapist pretended to be a friend who had recently broken up, and Walter explained why this did not make the friend a relationship failure. In the process of doing so, Walter suddenly realized that he wasn't a failure, either.

I commented:

> Walter was asked whether he'd say something harsh to a friend, and he said no, but that alone wasn't enough to improve his condition. What did help was putting him in a position where he had to really think through the arguments for why this is irrational in order to convince his friend, and then, after having formulated the arguments once himself, get convinced by them himself.
>
> In terms of our framework, we might say that a part of Walter's mind contained a model which output a harsh judgment of himself, while another part contained a model which would output a much less harsher judgment of someone else who was in otherwise identical circumstances. Just bringing up the existence of this contradiction wasn't enough to change it: it caused the contradiction to be noticed, but didn't activate the relevant models extensively enough for their contents to be reprocessed.
>
> But when Walter had to role-play a situation where he thought of himself as actually talking with a depressed friend, that required him to more fully activate the non-judgmental model and apply it to the relevant situation. This caused him to blend with the model, taking its perspective as the truth. When that perspective was then propagated to the self-critical model, the easiest way for the mind to resolve the conflict was simply to alter the model producing the self-critical thoughts.

This seems like a straightforward instance of belief juxtaposition, and one where I ended up independently deriving something like UtEB's memory reconsolidation model: I too noted that the relevant belief structures need to be simultaneously activated in the right way to allow for the brain to revise one of them after noticing the contradiction. In general, UtEB's model of how things work rings true in my experience, making me inclined to believe that its description of how therapy works is correct, and that its model of how it is connected to neuroscience might also be.

UtEB and the subagent model

As many readers know, I have been writing a sequence of posts on multi-agent models of mind. In Building up to an Internal Family Systems model, I suggested that the human mind might contain something like subagents which try to ensure that past catastrophes do not repeat. In subagents, coherence, and akrasia in humans, I suggested that behaviors such as procrastination, indecision, and seemingly inconsistent behavior result from different subagents having disagreements over what to do.

As I already mentioned, my post on integrating disagreeing subagents took the model in the direction of interpreting disagreeing subagents as conflicting beliefs or models within a person's brain. Subagents, trauma and rationality further suggested that the appearance of drastically different personalities within a single person might result from unintegrated memory networks, which resist integration due to various traumatic experiences.

This post has discussed UtEB's model of conflicting emotional schemas in a way which further equates "subagents" with beliefs - in this case, the various schemas seem closely related to what e.g. Internal Family Systems calls "parts". In many situations, it is probably fair to say that this is what subagents are.

That said, I think that while this covers a very important subset of subagents, not *everything* which I have been referring to as a subagent falls straightforwardly under the belief-schema model. In subagents and neural Turing machines as well as Against "System 1" and "System 2", I also covered subagents in a more general way, as also including e.g. the kinds of subsystems which carry out object recognition and are used to carry out tasks like arithmetic. This was also the lens through which I looked at subagents in my summary of Consciousness and the Brain. Which kind of view is the most useful, depends on exactly what phenomenon we are trying to understand.

A mechanistic model of meditation

Meditation has been claimed to have all kinds of transformative effects on the psyche, such as improving concentration ability, healing trauma, cleaning up delusions, allowing one to track their subconscious strategies, and making one's nervous system more efficient. However, an explanation for why and how exactly this would happen has typically been lacking. This makes people reasonably skeptical of such claims.

In this post, I want to offer an explanation for one kind of a mechanism: meditation increasing the degree of a person's introspective awareness, and thus leading to increasing psychological unity as internal conflicts are detected and resolved.

Note that this post does *not* discuss "enlightenment". That is a related but separate topic. It is possible to pursue meditation mainly for its ordinary psychological benefits while being uninterested in enlightenment, and vice versa.

What is introspective awareness?

In an earlier post on introspective awareness, I distinguished between being *aware of something*, and *being aware of having been aware of something*. My example involved that of a robot whose consciousness contains one mental object at a time, and which is aware of different things at different times:

> Robot's thought at time 1: It's raining outside
>
> Robot's thought at time 2: Battery low
>
> Robot's thought at time 3: Technological unemployment protestors are outside
>
> Robot's thought at time 4: Battery low
>
> Robot's thought at time 5: I'm now recharging my battery

At times 2-5, the robot has no awareness of the fact that it was thinking about rain at time 1. As soon as something else captures its attention, it has no idea of this earlier conscious content - *unless* a particular subsystem happens to record the fact, and can later re-present the content in an appropriately tagged form:

> Time 6: At time 1, there was the thought that [It's raining outside]

I said that at time 6, the robot had a *moment of introspective awareness*: a mental object containing a summary of its previous thoughts, which can then be separately examined and acted upon.

Humans are not robots. But I previously summarized the neuroscience book *Consciousness and the Brain*, and its global neuronal workspace (GNW) model of consciousness. According to this model, the contents of consciousness correspond to what is being represented in a particular network of neurons - the global workspace - that connects different parts of the brain. Different systems are constantly competing to get their contents into the global workspace, which can only hold one piece of content at a time. Thus, like robots, we too are only aware of one thing at a time, and tend to lose awareness of our earlier thoughts - unless something reminds us of them.

In what follows, I will suggest that like robots, humans also have a type of conscious content that we might call *introspective awareness*, which allows us to be more aware of our previous mental activity. (I am borrowing the term from the meditation book *The Mind Illuminated,* which distinguishes between *introspective attention*, *introspective awareness*, and *metacognitive introspective awareness*. I am eliding these differences for the sake of simplicity.)

I will also explore the idea that introspective awareness is a sensory channel in a similar sense as vision and sound are. The experience of sight or sound is produced by subsystems which send information to consciousness; likewise, introspective awareness is produced by a subsystem which captures information in the brain and then sends it (back) to consciousness.

We can train our other senses to become more accurate and detailed. Gilbert, Sigman & Crist (2001), reviewing the neuroscience on sensory training, list a number of ways in which discrimination can be increased in a variety of sensory modalities: among other things, "visual acuity, somatosensory spatial resolution, discrimination of hue, estimation of weight, and discrimination of acoustical pitch all show improvement with practice"; even the spatial resolution of the visual system can be deliberately increased by training.

If introspective awareness is a sensory channel, can it also be practiced to improve the number of details it will pick up on? One may feel that I am stretching the metaphor here. But in fact, *Consciousness and the Brain* suggests that *all* sensory training is in a sense training in introspection. The additional information that we get by training our senses has always been collected by our brain, but that information has remained isolated at lower levels of processing. To make it conscious, one needs to grow new neural circuits which extract the lower-level information and re-encode it in a format which can be sent to consciousness.

Thus, the brain *already* has the ability to take normally unavailable subconscious information and make it consciously available by practice. What is needed is a way to point that learning process at the kind of information that we would normally consider "introspective", rather than on an external information source.

From *Consciousness and the Brain:*

> ... a fourth way in which neural information can remain unconscious, according to workspace theory, is to be diluted into a complex pattern of firing. To take a concrete example, consider a visual grating that is so finely spaced, or that flickers so fast (50 hertz and above), that you cannot see it. Although you perceive only a uniform gray, experiments show that the grating is actually encoded inside your brain: distinct groups of visual neurons fire for different orientations of the grating. Why can’t this pattern of neuronal activity be brought to consciousness? Probably because it makes use of an extremely tangled spatiotemporal pattern of firing in the primary visual area, a neural cipher too complex to be explicitly recognized by global workspace neurons higher up in the cortex. Although we do not yet fully understand the neural code, we believe that, in order to become conscious, a piece of information first has to be re-encoded in an explicit form by a compact assembly of neurons. The anterior regions of the visual cortex must dedicate specific neurons to meaningful visual inputs, before their own activity can be amplified and cause a global workspace ignition that brings the information into awareness. If the information remains diluted in the firing of myriad unrelated neurons, then it cannot be made conscious.

> Any face that we see, any word that we hear, begins in this unconscious manner, as an absurdly contorted spatiotemporal train of spikes in millions of neurons, each sensing only a minuscule part of the overall scene. Each of these input patterns contains virtually infinite amounts of information about the speaker, message, emotion, room size . . . if only we could decode it—but we can't. We become aware of this latent information only once our higher-level brain areas categorize it into meaningful bins. Making the message explicit is an essential role of the hierarchical pyramid of sensory neurons that successively extract increasingly abstract features of our sensations. Sensory training makes us aware of faint sights or sounds because, at all levels, neurons reorient their properties to amplify these sensory messages. Prior to learning, a neuronal message was already present in our sensory areas, but only implicitly, in the form of a diluted firing pattern inaccessible to our awareness.

Richard's therapy session

We saw an example of introspective awareness in my post on the book *Unlocking the Emotional Brain*. In the transcript, a man named Richard has been suffering from severe self-doubt, and is asked to imagine how it would feel like if he made confident comments in a work meeting. The following conversation follows:

> Richard: Now I'm feeling really uncomfortable, but-it's in a different way.
>
> Therapist: OK, let yourself feel it - this different discomfort. [Pause.] See if any words come along with this uncomfortable feeling.
>
> Richard: [Pause.] Now they hate me.

The therapist is asking Richard to focus his attention on the feeling of discomfort, generating moments of introspective awareness *about* the discomfort. Notice that Richard becomes more thoughtful and less reactive to the anxiety as he does so. My guess of what is happening is something like this:

When Richard is feeling anxious, this means that a mental object encoding something like "the feeling of anxiety" is being represented in the workspace. This activates neural rules which trigger the kinds of responses that anxiety has evolved to produce. For example, a system may be triggered which attempts to plan how to escape the situation causing the anxiety. This system's intentions are then injected into the workspace, producing a state of mind where the feeling of anxiety alternates with thoughts of how to get away.

Introspective awareness is its own type of mental object, produced by a different subsystem which takes inputs from the global workspace, re-encodes them in a format which highlights particular aspects of that data, and outputs that back into the workspace. When a representation of an anxious state of mind is created, that representation does not by itself trigger the same rules as the original anxiety did.

As a result, as representations of the anxiety begin to alternate *together with* the anxiety, there are proportionately less moments of anxiety. This in turn triggers fewer of the subsystems attempting to escape the situation, making it easier to reflect on the anxiety without being bothered by it.

When Richard's therapist asks him to feel the anxiety and to see if any words come along with it, the subsystem for introspective awareness was primed to look for any content that could be re-presented in verbal form. As Richard's anxiety had been produced by an emotional schema including a prediction that being confident makes you hated, some of that information had passed through the workspace and been available for the awareness subsystem to capture. This brought up the verbalization of what the schema predicted would happen if Richard was confident - "now they hate me".

> Therapist: "Now they hate me." Good. Keep going: See if this really uncomfortable feeling can also tell you why they hate you now.

According to the GNW model, when a particular piece of content is maintained as the center of attention, it strengthens the activation of any structures associated with it. As Richard's therapist guides him to focus on the verbal content, more information related to it is broadcast into the workspace. The further prompt guides the awareness subsystem to look for patterns that feel like the *reason* for the hate.

> Richard: [Pause.] Hnh. Wow. It's because… now I'm… an arrogant asshole… like my father… a totally self-centered, totally insensitive know-it-all.

The therapist then takes a pattern which Richard has brought up and helps crystallize it further, and throws it back to Richard for verification.

> Therapist: Do you mean that having a feeling of confidence as you speak turns you into an arrogant asshole, like Dad?
>
> Richard: Yeah, exactly. Wow.

In this example, we saw that having more moments of introspective awareness was beneficial for Richard. As aspects of his moment-to-moment consciousness were made available for other subsystems to examine, the emotional schema causing the anxiety was identified and its contents extracted into a format which could be fed into other subsystems. Later on, when Richard's co-worker displayed confidence which others approved of, a contradiction-detection mechanism noticed a discrepancy between reality and the prediction that confidence makes you hated, allowing the prediction to be revised.

Under this model, the system which produces moments of introspective awareness is a subsystem like any other in the brain. This means that it will be activated when the right cues trigger it, and its outputs compete with the outputs of other systems submitting content to consciousness. The circumstances under which the system triggers, and its probability of successfully making its contents conscious, are modified by reinforcement learning. Just as practicing a skill such as arithmetic eventually causes various subsystems to manipulate the content of consciousness in the right order, practicing a skill which benefits from introspective awareness will cause the subsystem generating introspective awareness to activate more often.

Meditation as a technique for generating moments of introspective awareness

Just as there are different forms and styles of therapy, there are also different forms and styles of meditation. All of them involve introspective awareness to at least some degree, but they differ in what that awareness is then used for.

In the example with Richard, his therapist asked him to imagine being confident and to then bring his awareness to *why* that felt uncomfortable. In contrast, a more behaviorally oriented therapist might not have examined the reason behind the discomfort. Rather, they might have taught Richard to notice his reaction to the discomfort, and then use that as a cue for implementing an opposite reaction. Both kinds of therapists would ask their clients to generate *some* introspective awareness, but aiming that awareness at different kinds of features, and using the awareness to trigger different kinds of strategies. The results would correspondingly be very different.

Likewise, systems of meditation differ in how much introspective awareness they produce, what kinds of features the awareness-producing subsystem is trained to extract, and what that awareness is then used for. For this article, I have chosen to use the example of the system in *The Mind Illuminated* (TMI), as it is clearly explained and explicitly phrased in these terms. (Again, TMI has a more precise distinction between introspective attention and introspective awareness, which I am eliding for the sake of simplicity.)

In TMI's system, as in many others, you start with trying to keep your attention on your breath. In terms of our model, this means that you want to keep sensory outputs corresponding to your breath as the main thing in your consciousness.

The problem with this goal is that there is no subsystem which can just unilaterally decide what to maintain as the center of attention. At any given moment, many different subsystems are competing to make their content conscious. So one system might have the intention to follow the breath, and you do it for a while, but then a planning system kicks in with its intention to think about dinner. Such planning has tended to feel rewarding, so it wins out and the intent to meditate is forgotten until five minutes later, when you decide what you want for dinner and then suddenly remember the thing about following your breath.

TMI calls this *mind-wandering from forgetting*, and the first step of practice is just to notice it whenever it happens, congratulate yourself for having noticed it, and then return to the breath. Being able to notice forgetting requires having a moment of introspective awareness which points out the fact that you had not been following your breath. When you take satisfaction in having noticed this, your awareness-producing subsystem gets assigned a reward and becomes slightly more likely to activate in the future. "Have I remembered to follow my breath or not?" acts a feedback mechanism that you can explicitly train on.

As the awareness-producing system starts to activate more often and ping you if you have forgotten to meditate, periods of mind-wandering grow shorter.

Now, even if you stop getting entirely lost in thought, you still have *distraction:* content from other subsystems that is in consciousness together with the sensations of the breath and the intention to focus on the breath. For example, you might be having stray thoughts, hearing sounds from your environment, and experiencing sensations from your body.

To more exclusively focus on the breath, you are instructed to maintain the intent to both attend to it and also to be aware of any distractions. The subsystems which

output mental content can, and normally do, operate independently of each other. This means that the following may happen:

Subsystem 1: I'm meditating well!

Subsystem 2: Hmm, what's that smell.

Subsystem 1: I'm meditating well! No distractions.

Subsystem 2: Smells kinda like cookies.

Subsystem 2: Mmm, cookies.

Subsystem 1: Continuing to meditate well!

Subsystem 2: Say, what's for dinner?

That is, a system which tracks the breath can continue to repeatedly find the breath, and report that your meditation is proceeding well and with no distractions… all the while the content of your consciousness continues to alternate with distracted thoughts, which the breath-tracking subsystem is failing to notice (because it is tracking the breath, not the presence of other thoughts). Worse, since you may find it rewarding to just *think that you are meditating well*, that *thought* may start to become rewarded, and you may find yourself just *thinking* that you are meditating well… even as that thought has become self-sustaining and no longer connected to whether you are following the breath or not!

There are all kinds of subtle traps like this, and reducing the amount of distraction requires you to first have better awareness *of* the distraction. This means more moments of introspective awareness which are tracking what's *actually* happening in your mind:

Subsystem 1: I'm meditating well!

Subsystem 2: Hmm, what's that smell.

Subsystem 1: I'm meditating well! No distractions.

Subsystem 2: Smells kinda like cookies.

Subsystem 2: Mmm, cookies.

A*wareness subsystem:* Wait, one train of thought keeps saying that it's meditating well, but another is totally getting into the thought of food.

Subsystem 1: Oh. Better refocus that attention on the breath, and spend less time thinking about the *concept* of following the breath.

This kind of a process also teaches you to pay attention to patterns of cause and effect in your mind. In this example, the smell of cookies caused you to think of cookies, which in turn made you think of dinner, which could have ultimately led to forgetting and mind-wandering.

Catching the train of thought after "mmm, cookies" meant that three "processing steps" had passed before you noticed it. If you practice tracing back trains of thought in your mind, you seem to teach your awareness-system to collect and store data

from a longer period, even when it is not actively outputting it. This means that at the “mmm, cookies” stage, you can query your awareness to get a trace of the immediately preceding thought chain.

You notice that you started to get distracted starting from the smell of the cookie and can then use this as further input to your awareness system. You are essentially taking the re-presented smell of the cookie which the system output, and feeding it back in, asking it to pay more attention to detecting “things like this”. The next time that you notice a smell, your introspective awareness may flag it right away, letting you catch the distraction at the very first stage and before it turns into an extended train of thought.

Note that there is nothing particularly mysterious or unusual about any of this. You are employing essentially the same process used in learning *any* skill. In learning to ride a bike, for example, attempting to keep the bike balanced involves adjusting your movements in response to feedback. When you do so, your brain becomes better at detecting things like “tilting towards the right” in the sense data, increasing your ability to apply the right correction. After you have learned to identify tilting-a-lot-but-not-quite-falling, your brain learns to backtrace to the preceding state of tilting-a-little-less, and apply the right correction there. Once its precision has been honed to identify that state, you can further detect an even subtler tilt, until you automatically apply the right corrections to keep you balanced.

Essentially the kind of a learning algorithm is being applied here. Increased sensory precision leads to improvements in skill which allow for increased sensory precision. (See also this article, which goes into more detail about TMI as a form of deliberate practice.)

Uses for moments of introspective awareness

I should again emphasize that the preceding explanation is only looking at one particular meditation system. There are other systems which work very differently, but they all use or develop introspective awareness to some extent. For example:

- **In Shinzen Young's formulation of “do nothing” practice,** you have just two basic instructions: *let whatever happens, happen* and *when you notice an intention to control your attention, drop that intention.* This trains introspective awareness to notice when one is trying to control their attention… but it is also a very different system, since maintaining an intention to notice when that happens would also be an attempt to control attention! Thus, one is instructed to drop intentions if one spontaneously notices them, but not to actively look for them.
- **In noting practice,** you are trying to consciously name or notice everything that happens in your consciousness. Introspective awareness is trained to very rapidly distinguish between everything that happens, but is not trained to maintain attention on any particular thing.
- **In visualization practice,** you might create a visual image in your mind, then use introspective awareness to examine the mental object that you’ve created and compare it to what a real image would look like. This gives the subsystem creating the visualization feedback, and helps slowly develop a more realistic image.

Going back to TMI-style introspective awareness, once you get it trained up, you can use it for various purposes. In particular, once you learn to maintain it during your daily life - and not just on the meditation couch - it will bring up more assumptions in your various schemas and mental models. Think of Richard paying attention to the assumptions behind his unwanted reactions and making them explicit, but as something that happens on a regular basis as the reactions come up.

Romeo Stevens described what he called “the core loop of Buddhism”:

> So, what is the core loop?
>
> It's basically cognitive behavioral therapy, supercharged with a mental state more intense than most pharmaceuticals.
>
> There are two categories of practice, one for cultivating the useful mental state, the other uses that mental state to investigate the causal linkages between various parts of your perception (physical sensations, emotional tones, and mental reactions) which leads to clearing out of old linkages that weren't constructed well.
>
> You have physical sensations in the course of life. Your nervous system reacts to these sensations with high or low valence (positive, negative, neutral) and arousal (sympathetic and parasympathetic nervous system activation), your mind reacts to these now-emotion-laden sensations with activity (mental image, mental talk) out of which you then build stories to make sense of your situation.
>
> The key insight that drives everything is the knowledge (and later, direct experience) that this system isn't wired up efficiently. Importantly: I don't mean this in a normative way. Like you should wire it the way I say just because, but in the 'this type of circuit only needs 20 nand gates, why are there 60 and why is it shunting excess voltage into the anger circuits over there that have nothing to do with this computation?' way. Regardless of possible arguments over an ultimately 'correct' way to wire everything, there are very low hanging fruit in terms of improvements that will help you effectively pursue *any* other goal you set your mind to.

Again, we saw an example of this with Richard. He had experienced his father as acting confident and as causing suffering to Richard and others; sensations which his mind has classified as negative. In order to avoid them, a model (story) was constructed saying that confidence is horrible, and behaviors (e.g. negative self-talk) were created to avoid appearing horrible.

Now, this caused problems down the line, making him motivated to try to appear more confident… meaning that there was now a mechanism in his brain trying to prevent him from appearing confident, and another which considered this a problem and tried to make him more confident, in opposition to the first system. See what Romeo means when talking about circuits that only need 20 gates but are implemented using 60?

The article “tune your motor cortex” makes the following claims about muscle movement:

> Your motor cortex automatically learns to execute complex movements by putting together simpler ones, all the way down to control of individual muscles.

> Because the process of learning happens organically, the resulting architecture of neural connections (you can think of them as "hidden layers" in machine learning terms) is not always perfectly suited to the task.
>
> Some local optima of those neural configurations are hard to get out of, and constantly reinforced by using them.
>
> There is some pressure for muscle control to be efficient, and the motor cortex is doing a "good enough" job at it, but tends to stop a fair bit from perfection.
>
> By repeating certain movements and positions over and over again (e.g. during sitting work), we involuntarily strengthen connections between movements and muscles that don't make much sense lumped together.
>
> E.g. control of shoulders might become spuriously wired together with control of thighs (both are often tense during sitting).

There are various mental motions which are learned in basically the same way as physical motions are:

- You learn to calculate 12*13 by a technique such as first multiplying 10*13, keeping the result in your memory, calculating 2*13, and then adding the intermediate results together.
- You learn that a particular memory makes you feel slightly unpleasant, and that flinching away from anything that would remind you of it takes the pain away.
- You learn that this also works on uncomfortable chores, teaching you to keep pushing the thought of them away.
- You learn that your father's behavior is painful to you, and that any confidence reminds you of that, so you learn negative self-talk which blocks you from acting confident.
- You learn that saying "no" to people reminds you of being punished for saying "no" to your parents, but that saying "yes" too often means that you are constantly fulfilling promises to other people - so you learn to avoid situations where you would be asked anything.
- You learn that there's something you can do in your mind to stop feeling upset, so you start ignoring your emotions and any information they might have.
- You learn that if you feel bad about not getting the respect you want, thinking "if only I was good enough at persuasion, I would get what I want" gives you a sense of control - even though this pattern also makes you feel personally at fault when you *don't* get what you want.
- You learn that it's rewarding to punish people who have wronged, so you *always* want to punish someone when something goes wrong - even if there is nobody but reality to punish.
- You learn that it feels good to mentally punish someone who is munching too loud, but actually complaining about it would feel petty, and you've learned that pettiness is frowned upon. So you also learn to block the impulse to say anything out loud, but continue to get increasingly angry about the sound, causing an escalating circle of both the annoyance and the blocking ramping up in intensity.

As with physical movements, these can form local optima that are hard to get out of. Many of them are learned in childhood, when your understanding of the world is limited. But new behaviors continue to build on top of them, so you will eventually end up with a system which could use a lot of optimization.

If you have more introspective awareness of the exact processes that are happening in your mind, you can make more implicit assumptions conscious, causing your brain's built-in contradiction detector to notice when they contradict your later learning. Also, getting more feedback about what exactly is happening in your mind allows you to notice more wasted motion in general.

One particular effect is that, as *Unlocking the Emotional Brain* notes, the mind often makes trade-offs where it causes itself some minor suffering in order to avoid a perceived greater suffering. For example, someone may feel guilt in order to motivate themselves, or experience self-doubt to avoid appearing too confident. By employing greater introspective awareness, one may find ways to achieve their goals *without* needing to experience any suffering in order to do so.

Of course, Buddhist meditation is not the only way to achieve this. Various therapies and techniques such as Focusing, Internal Family Systems, Internal Double Crux, and so on, are also methods which use introspective awareness to reveal and refactor various assumptions. Increased introspective awareness from meditation tends to also boost the effectiveness of related techniques, as well as reveal more situations where they can be employed.

If introspective awareness is so great, why don't we have it naturally?

As with anything, there are tradeoffs involved. Having more introspective awareness can help fix a lot of issues… but it also comes with risks, which I assume is the reason why we have *not* evolved to have a lot of it all the time.

First, it's worth noting that even for experienced meditators, intense emotional reactions tend to shut down introspective awareness. If one of the functions of e.g. fear and anxiety is to cause a rapid response, then excessive amounts of introspective awareness would slow down that response by reducing cognitive fusion. Many emotions seem to inhibit many competing processes from accessing consciousness, so that you can deal with the situation at hand.

Another consideration involves traumatic memories. In the beginning of the article, I suggested that anxiety is a special kind of mental object which activates particular behaviors. In general, different emotional states have specific kinds of behaviors and activities associated with them - meaning that if you have some memories which are *really* painful, they can become overwhelming, making it necessary to block them in order to carry on with your normal life. Meditation can be helpful for working through your trauma, but it can also bring it up before you are ready for it, to the point of requiring professional psychotherapy to get through. If you are better at noticing all kinds of subtle details in your mind, it also becomes easier to notice anything that would remind you of things you don't want to remember. A *decrease* in introspective awareness seems to be a common trauma symptom, as this helps block the unpleasant memories from being too easily triggered.

I have also heard advanced meditators mention that increased introspective awareness makes it difficult to push away pangs of conscience that they would otherwise have ignored, causing practical problems. For example, people have said that they are no longer able to eat animal products or tell white lies.

On the other hand, extended concentration practice can also make it easier to *block* things which you would be better off not blocking.

So far, this article has mostly focused on using introspective awareness to notice the *content* of your thoughts. But you can also use it to notice the *structure* of the higher-level processes generating your thoughts. Part of how you develop concentration ability is by maintaining introspective awareness of the fact that being able to concentrate on just one thing feels more pleasant than having your attention jump between many different things. This can give you an improved ability to choose what you are concentrating on… but also to selectively exclude anything unpleasant from your mind.

For example, there was an occasion when I needed to do some work, but also had intense anxiety about not wanting to; intense enough that it would normally have made it impossible for me to focus on it. So then I *tried* to work, and let my introspective awareness observe the feeling of head-splitting agony from my attention alternating between the work and the desire not to… and to also notice that whenever my attention was on the work, I felt temporarily better.

After a while of this, the anxiety started to get excluded from my consciousness, until it suddenly dropped away completely - as if some deeper process had judged it useless and revoked its access to consciousness. And while this allowed me to do the work that I needed to, it also felt internally violent, and like it would be too easy to repress any unpleasant thoughts using it. I still use this kind of technique on occasion when I need to concentrate on something, but I try to be cautious about it.

The negative side of being able to get better feedback about your mental processes, is that you can also get better feedback on exactly how pleasant wireheading feels. If you like to imagine pleasant things, you can get better and better at imagining pleasant things, and excluding any worries about it from your consciousness. Meditation teacher Daniel Ingram warns:

> Strong insight and concentration practice, even when that practice wasn't dedicated to the powers, can make people go temporarily or permanently (or for the rest of that lifetime) psychotic. The more the practice involves creating experiences that diverge significantly from what I will crudely term "consensus reality", and the longer one engages in these practices, the more likely prolonged difficulties are. It is of note that a significant number of the primary propagators of the Western magickal traditions became moderately nuts towards the ends of their lives.
>
> As one Burmese man said to Kenneth, "My brother does concentration practice. You know, sometimes they go a little mad!" He was talking about what can sometimes happen when people get into the powers. [...]
>
> I remember a letter from a friend who was on a long retreat in Burma and was supposed to be doing insight practices but had slipped into playing with these sorts of experiences. He was now fascinated by his ability to see spirit animals and other supernormal beings and was having regular conversations with some sort of low-level god that kept telling him that he was making excellent progress in his insight practice—that is, exactly what he wanted to hear. However, the fact that he was having stable visionary experiences and was buying into their content made it abundantly clear that he wasn't doing insight practices at all, but was lost in and being fooled by these.

Now, it should be pointed out that “being able to exclude anything unpleasant from your consciousness” is only going to be a worry for advanced practitioners who spend a lot of time on the kind of practice that inclines you towards these kinds of risks. Before you get to the point of something like this being a risk, you will get to resolve a lot of internal conflicts and old issues first.

Here is Culadasa, the author of *The Mind Illuminated,* being interviewed about this kind of a “first you resolve a lot of issues, but then you can get the ability to push down the rest” dynamic:

> Michael Taft: … and you’re using the meditation practice to help work with your stuff. But what about the other case that we both know of where people have reached very high levels of meditative capacity, they’ve got a lot of insight, maybe they’re at some level of awakening, and they seem to have, in a way, missed a whole pocket of material, or several pockets of material. It’s like they think they’re doing fine, but maybe everyone around them is aware that they’ve got these behavior patterns that do not seem awake at all. And yet the meditation has somehow missed that.
>
> Culadasa: Yes, yes. […] ... there seems to be a certain level of the stuff that we’re talking about that it’s necessary to deal with to achieve awakening, but it’s sort of a minimal level. […] What I think that is indicative of is that if that hasn’t been sufficiently dealt with earlier, it has to get dealt with in one way or another at that point. That doesn’t necessarily mean that it’s going to get resolved; it may just get reburied a little more deeply.
>
> Michael Taft: Pushed out of the way.
>
> Culadasa: Yeah, pushed out of the way, or bypassed in some way. That allows a person to go ahead and [progress] and it’s unrealistic to think that everything has been resolved. […] a lot of the things that change […] actually help to push these things aside, to bypass them in one way or another, whereas before somebody has [made as much progress] these would have been sufficiently problematic in their life that, in one way or another, they would be aware of them, whether or not they did anything about them or were at a place of just taking for granted that I have these, quote, “personality characteristics” that are a bit difficult.

I used to be very enthusiastic about TMI’s meditation system. I still consider it important and useful to make progress on, but am slightly more guarded after some of my own experiences, hearing about the experience of a friend who reached a high level in it, reading some critiques of its tendency to emphasize awareness of positive experiences [1 2], and considering both the interview quoted above and Culadasa’s subsequent actions. (That said, the focus on positive experiences can be a useful counterbalance for people who start off with an overall negative stance towards life.)

I continue to practice it, and would generally find it safe until you get to around the sixth or so of its ten stages, at which point I would suggest starting to exercise some caution. Off the couch, I mostly don’t do much concentration practice (except in a context where I would need to concentrate anyway). Rather I try to focus my introspective awareness towards just observing my mind without actively interfering with it, Internal Family Systems -style practice, and other activities that do not seem to risk excluding too much unpleasant material.

Finally, developing too much awareness into your mind may cause you to start noticing contradictions between how you thought it worked, and how it actually works.

I suspect that a part of how our brains have evolved to operate, *relies on* those differences going unnoticed. This gets us to the topic of enlightenment, which I have not yet discussed, but will do in my next post.

Thanks to Maija Haavisto, Lumi Pakkanen and Romeo Stevens for comments on an earlier draft.

A non-mystical explanation of insight meditation and the three characteristics of existence: introduction and preamble

Introduction

Insight meditation, enlightenment, what's that all about?

The sequence of posts starting from this one is my personal attempt at answering that question. It grew out of me being annoyed about so much of this material seeming to be straightforwardly explainable in non-mysterious terms, but me also being unable to find any book or article that would do this to my satisfaction. In particular, I wanted something that would:

- Explain what kinds of implicit assumptions build up our default understanding of reality and how those assumptions are subtly flawed. It would then point out aspects from our experience whose repeated observation will update those assumptions, and explain how this may cause psychological change in someone who meditates.
- It would also explain how the so-called "three characteristics of existence" of Buddhism - impermanence, no-self and unsatisfactoriness - are all interrelated and connected with each other in a way your average Western science-minded, allergic-to-mysticism reader can understand.

I failed to find a resource that would do this in the way I had in mind, so then I wrote one myself.

From the onset, I want to note that I am calling this *a* non-mystical take on the three characteristics, rather than *the* non-mystical take on the three characteristics. This is an attempt to explain what I personally think is going on, and to sketch out an explanation of how various experiences and Buddhist teachings *could* be understandable in straightforward terms. I don't expect this to be anything like a complete or perfect explanation, but rather one particular model that might be useful.

The main intent of this series is summarized by a comment written by Vanessa Kosoy, justifiably skeptical of grandiose claims about enlightenment that are made without further elaboration on the actual mechanisms of it:

> I think that the only coherent way to convince us that Enlightenment is real is to provide a model from a 3rd party perspective. [...] The model doesn't have to be fully mathematically rigorous: as always, it can be a little fuzzy and informal. However, it must be precise enough in order to (i) correctly capture the essentials and (ii) be interpretable more or less unambiguously by the sufficiently educated reader.
>
> Now, having such a model doesn't mean you can actually reproduce Enlightenment itself. [...] However, producing such a model would give us the

> enormous advantages of (i) being able to come up with experimental tests for the model (ii) understanding what sort of advantages we would gain by reaching Enlightenment (iii) being sure that your are talking about something that is at least a coherent possible world even if we are still unsure whether you are describing the actual world.

I hope to at least put together a starting point for a model that would fulfill those criteria.

Note that these articles are *not* saying “you should meditate”. Getting deep in meditation requires a huge investment of time and effort - though smaller investments are also likely to produce benefits - and is associated with its own risks [1 2 3 4]. My intent is merely to discuss some of the mechanisms involved in meditation and the mind. Whether one should get direct acquaintance with them is a separate question that goes beyond the scope of this discussion.

Briefly on the mechanisms of meditation

In a previous article, A Mechanistic Model of Meditation, I argued that it is possible in principle for meditation to give people an improved understanding of the way their mind operates.

To briefly recap my argument: we know it is possible for people to train their senses, such as learning to notice more details or make more fine-grained sensory discriminations. One theory is that those details have always been processed in the brain, but the information has not made it to the higher stages of the processing hierarchy. As you repeatedly focus your attention to a particular kind of pattern in your consciousness, neurons re-orient to strengthen that pattern and build connections to the lower-level circuits from which it emerges. This re-encodes the information in those circuits in a format which can be represented in consciousness.

This means at least some kinds of sensory training are *training in introspection* - learning to better access information which already exists in your brain. This implies you can also learn to strengthen *other* patterns in your consciousness, especially if you have some source of feedback that you can use to guide the training.

I gave an example of experiential forms of therapy doing exactly this, and then described how a particular style of meditation used one’s awareness of the breath as an objective feedback signal for developing increased “introspective awareness” of one’s own mind.

That post was mostly describing the ways in which meditation can be used to become more aware of the *content* of your thoughts. However, in observing the content, it is hard to avoid noticing at least some of the *structure* of the thought process as well.

For example, you might try to follow your breath and think you are doing a good job. In this case, there are at least two kinds of content in your mind: the actual sensory experience of the breath, and thoughts about how badly or well you are doing. The latter might take the form of e.g. mental dialogue that says things like "I'm still managing to follow my breath". Now, since you may find it rewarding to just think that you are meditating well, *that thought* may start to become rewarded, and you may find yourself repeatedly *thinking* that you are successfully following the breath... even

as the thought of “I am meditating well” has become self-sustaining and no longer connected to whether you are following the breath or not.

Eventually you will realize that you have actually been *thinking about following the breath* rather than *actually following it.* This is a minor insight into the way that your thought processes are structured, revealing it is possible for sensations and thoughts about sensations to become mixed up.

It is also possible to practice meditation in a way which explicitly focuses on investigating structure. We can make an analogy to looking at a painting. (Thanks to Alexei Andreev for suggesting this analogy.) Seen from some distance, a painting has "content": it depicts things like people, buildings, boats and so forth. But when you get closer to it and look carefully, you can see that all the content is composed of things like brush strokes, individual colored shapes, paint of varying thickness, and so on. This is "structure". While all types of meditation are going to reveal *something* about structure, there are also types of meditation which are specifically aimed at exploring it. Meditation which focuses on investigating structure is commonly called *insight* meditation.

Investigating the mind vs. investigating reality

Now, it is worth noting that these practices are not always framed in terms of "investigating the structure of the mind", nor does the actual experience of doing them necessarily feel like that. Rather, the framing and experience is commonly that of investigating the nature of *reality*.

For example, in an earlier article trying to explain insight meditation, I mentioned I had once had the thought that I could never be happy. When I paid closer attention to why I thought that, I noticed that my mental image of a happy person included strong extraversion, which conflicted with the self-image that I had of myself as an introvert. After I noticed the happiness-extraversion connection, it became apparent that I could be happy even as an introvert, and the original thought disappeared. (Although I didn't know it at the time, it is common for emotional beliefs to change when they become explicit enough for the brain to notice them being erroneous.)

Essentially, I had originally believed “I can never be happy”, and this belief about me didn’t feel like a “belief”. It felt like *a basic truth of what I was*, the kind of truth that you just know - in the same way that you might look at an apple and *just know* you are having the experience of seeing an apple. But when I investigated the details of that experience, I realized that this wasn’t actually a fact about me. Rather it was just a belief that I had.

In a similar way, there are many aspects of our subjective experience that feel like facts about reality, but upon doing insight practices and investigating them closer, we can come to see that they are not so.

The philosopher Daniel Dennett has coined the term "heterophenomenology" to refer to a particular approach to the study of consciousness. In this approach, we assume that people are correctly describing how things *seem* to them and treat this as something that needs to be explained. However, the actual mechanism of why things seem like that to them, may be different from what they assume.

If I see an apple, it typically feels to me like I am seeing reality as it is. From a scientific point of view, this is mistaken: the sight of an apple is actually a complex interpretation my brain has created. Likewise, if I have the experience that I can never be happy, then this also feels like a raw fact while actually being an interpretation. In either case, if I manage to do practices which reveal my interpretation to be flawed, they will subjectively feel like I am investigating reality... while from a third-person perspective, we would rather say that I am investigating the way my mind builds up reality.

It is valid to stick to just the first-person experience of investigating reality directly. Many of these practices are framed solely in those terms, because a stance of curiosity and having as few assumptions as possible is the best mindset for actually doing the practices. But if one says that meditation investigates the nature of reality, then it becomes hard to test the claim from a third-person perspective. A common criticism is that meditation certainly *changes* how people experience the world, but it might just as well be *loosening* their grasp on reality.

On the other hand, if we provisionally assume that meditation works by revealing how the mind structures its model of reality, then we can check whether the kinds of insights that people report are compatible with what science tells us about the brain. If it turns out that meditators doing insight practices are coming up with experiences that match our understanding of actual brain mechanisms, then the practices might actually provide insight rather than delusion. In cases where no scientific evidence is yet available, it should at least be possible to construct a model that *could* be true and compatible with the third-person evidence.

In previous posts, I have explored some scientifically-informed models of the brain, which I think are naturally linked to the kinds of discoveries made in insight meditation. This article will more explicitly connect concepts from the theory of meditation to those kinds of models.

It is also worth noting that I think *both* claims about meditative insights are true: some things you can do with meditation *do* give you a better insight into reality, while some other things *do* just break your brain and reduce your contact with reality. (A fact responsible meditation teachers also warn about.) This makes it important to have third-person models of what could be a genuine insight and what is probably delusion, to help avoid the dangerous territory.

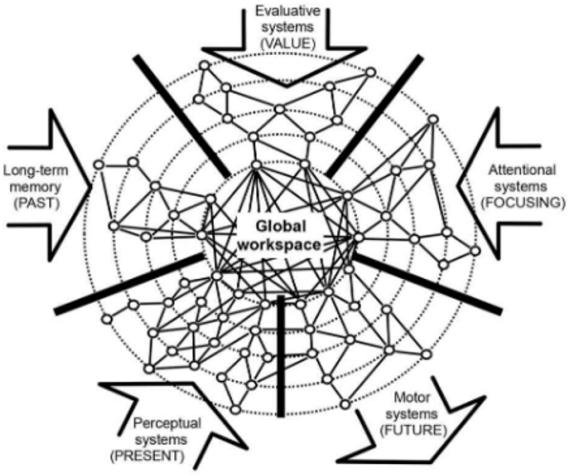
My multiagent model of mind

I have been calling my interpretation of those models a "multiagent model of mind". What follows is a highly abridged version of it; see the linked index of posts for much more extensive discussion, including the sources that I have been drawing on for my synthesis.

One of the main ideas of the multiagent model is that the brain contains a number of different subsystems operating in parallel, each focusing on their own responsibilities. They share information on a subconscious level, but also through conscious thought. The content of consciousness roughly corresponds to information which is being processed in a "global workspace" - a "brain web" of long-distance neurons, which link multiple areas of the brain together into a densely interconnected network.

The global workspace can only hold a single piece of information at a time. At any given time, multiple different subsystems are trying to send information into the workspace, or otherwise modify its contents. Experiments show that a visual stimuli needs to be shown for about 50 milliseconds for it to be consciously registered, suggesting that the contents of consciousness might be updated at least 20 times per second. Whatever information makes it into consciousness will then be broadcast widely throughout the brain, allowing many subsystems to synchronize their processing around it.

The exact process by which this happens is not completely understood, but involves a combination of top-down mechanisms (e.g. attentional subsystems trying to strengthen particular signals and keep those in the workspace) as well as bottom-up ones (e.g. emotional content getting a priority). For example, if you are listening to someone talk in a noisy restaurant, both their words and the noise are bottom-up information within the workspace, while a top-down process tries to pick up on the words in particular. If a drunk person then suddenly collides with you, you are likely to become startled, which is a bottom-up signal strong enough to grab your attention (dominate the workspace), momentarily pushing away everything else.



There is also a constant learning process going on, where the brain learns which subsystems should be given access in which circumstances, while the subsystems themselves also undergo learning about what kind of information to send to consciousness.

When I talk about “subsystems” sending content into consciousness, I mean this as a very generic term, which includes all of the following:

- Literal subsystems, e.g. information from the visual, auditory, and other sensory systems
- Subpatterns within larger subsystems, e.g. a particular neuronal pattern encoding a specific memory or habit
- Emotional schemas which trigger in particular situations and contain an interpretation of that situation and a response
- Working memory buffers associated with type 2 (“System 2”) reasoning, helping chain the outputs of several different subsystems together

In some cases, I might talk about there being two separate subsystems, when one could argue that this would be better described as something like two separate pieces of data within a single subsystem. For example, I might talk about two different memories as two different subsystems, when one could reasonably argue that they are both contained within the same memory subsystem. Drawing these kinds of distinctions within the brain seems tricky, so rather than trying to figure out what term to use when, I will just talk about subsystems all the time.

Epistemic status

> Buddhist theories of the mind are based on textual traditions that purport to record the remembered word of the Buddha, on religious and philosophical interpretations of those texts, and on Buddhist practices of mental cultivation. The theories aren't formulated as scientific hypotheses and they aren't scientifically testable. Buddhist insights into the mind aren't scientific discoveries. They haven't resulted from an open-ended empirical inquiry free from the claims of tradition and the force of doctrinal and sectarian rhetoric. They're stated in the language of Buddhist metaphysics, not in an independent conceptual framework to which Buddhist and non-Buddhist thinkers can agree. Buddhist meditative texts are saturated with religious imagery and language. Buddhist meditation isn't controlled experimentation. It guides people to have certain kinds of experiences and to interpret them in ways that conform to and confirm Buddhist doctrine. The claims that people make from having these experiences aren't subject to independent peer review; they're subject to assessment within the agreed-upon and unquestioned framework of the Buddhist soteriological path. […]
>
> I'm not saying that Buddhist meditative techniques haven't been experientially tested in any sense. Meditation is a kind of skill, and it's experientially testable in the way that skills are, namely, through repeated practice and expert evaluation. I have no doubt that Buddhist contemplatives down through the ages have tested meditation in this sense. I'm also not saying that meditation doesn't produce discoveries in the sense of personal insights. (Psychoanalysis can also lead to insights.) Rather, my point is that the experiential tests aren't experimental tests. They don't test scientific hypotheses. They don't provide a unique set of predictions for which there aren't other explanations. The insights they produce aren't scientific discoveries. […]
>
> I'm also not trying to devalue meditation. On the contrary, I'm trying to make room for its value by showing how likening it to science distorts it. Meditation isn't controlled experimentation. Attention and mindfulness aren't instruments that reveal the mind without affecting it. Meditation provides insight into the mind (and body) in the way that body practices like dance, yoga, and martial arts provide insight into the body (and mind). Such mind-body practices—meditation included—have their own rigor and precision. They test and validate things experientially, but not by comparing the results obtained against controls.
>
> -- Evan Thompson, *Why I Am Not A Buddhist*

I think it is reasonable to believe that meditation can give us genuine insights into the way the mind functions. The meditative techniques and practices which I am drawing upon in this series have been developed within Buddhist traditions, and I make frequent references to the theory developed within those traditions.

At the same time, while I am drawing upon theories developed within these traditions, I am treating those as a source of inspiration to be critically examined, rather than as sources of authority.

For one, there are many different Buddhist theories and schools that disagree with each other, many of them claiming to teach what the Buddha *really* meant. And as e.g. Evan Thompson's book discusses, one cannot cleanly separate Buddhist meditative techniques from Buddhist religious teaching. People who meditate using those techniques - myself included - do so while being guided by an existing conceptual framework, framing their experiences in light of their framework. Practitioners who use different kinds of techniques and frameworks end up drawing different conclusions: e.g. some frameworks end up at the conclusion that no selves exist, while others end up believing that everything is self. (The extent to which this difference in framing actually leads to a different *experience* is unclear.) Many of these frameworks also draw upon supernatural elements, such as claims of rebirth and remembering past lives.

Still, many meditation teachers also say things along the lines of "you should not take any of this on faith, just try it out and see for yourself". Personally I started out skeptical of many claims, dismissing them as pre-scientific folk-psychological speculation, before gradually coming to believe in them - sometimes as a result of meditation which hadn't even been aimed at investigating those claims in particular, but where I thought I was doing something completely different. And it seems to me that many of the meditative techniques actively require you to *suspend* your expectations in order to work properly, requiring you to look at what's present rather than at the thing you expect to see.

So, like many others, I simultaneously believe that i) meditative techniques point at genuine insights and also produce them in the minds of people who meditate and also that ii) we should not put excess faith in the claims of the existing meditation traditions. As many teachers encourage exactly this line of thought - as in the comment of taking nothing on faith - this feels like an appropriate spirit for approaching these matters.

Rather than trying to be authentically Buddhist, this article is concerned with building a model of the neural and psychological mechanisms I think the three characteristics are pointing at, even if that model ends up sharply deviating from the original theories. I heavily draw on my own experiences and the experiences and theories of other people whose reports I have reason to trust. I proceed from the assumption that regardless of whether the original frameworks are true or false, they do systematically produce similar effects and insights in the minds of the people following them, and that is an observation which needs to be explained.

In fact, I am happy to mix and match examples, exercises, interpretations and results drawn from all of the contemplative traditions that I happen to have any familiarity with, with current-day Western psychology and psychotherapy thrown in for good measure. They may have different approaches, but to the extent that they share commonalities, those commonalities tell us something about what human minds might have in common. And to the extent that they differ, one tradition might be pointing out aspects about the human mind that the others have neglected and vice versa, as in the fable of the blind men and the elephant.

Current articles in this series:

- Introduction and preamble (you are here)
- A non-mystical explanation of "no-self"
- Craving, suffering, and predictive processing
- From self to craving
- On the construction of the self
- Impermanence

Thank you to Alexei Andreev, David Chapman, Eliot Re, Jacob Spence, James Hogan, Magnus Vinding, Max Daniel, Matthew Graves, Michael Ashcroft, Romeo Stevens, Santtu Heikkinen, and Vojtěch Kovařík for valuable comments. Additional special thanks to Maija Haavisto. None of the people in question necessarily agree with all the content in this or the upcoming posts; much of the content has also been rewritten after the drafts that most of them saw.

A non-mystical explanation of "no-self" (three characteristics series)

This is the second post of the "a non-mystical explanation of insight meditation and the three characteristics of existence" series. You can read the first post, explaining my general intent and approach, here.

On the three characteristics

So, just what are the three characteristics of existence?

My take is that they are *a rough way of clustering the kinds of insights that you may get from insight meditation*: in one way or another, most insights about the structure of your mind can be said to be about no-self, impermanence, unsatisfactoriness, or some combination of them. As upcoming posts should hopefully make obvious, this is not a very clear-cut distinction: the three are deeply intertwined with each other, and you can't fully explain one without explaining the others. I am starting with a discussion of no-self, then moving to unsatisfactoriness, then coming back to no-self, moving between the different characteristics in a way that seems most clear.

I think that what's called "enlightenment" refers to the gradual accumulation of these kinds of insights, combined with practices aimed at exploiting an understanding of them. There are many different insights and ways of exploring them, as well as many general approaches for making use of them. Different traditions also seem to have different enlightenments [1, 2]. Thus, rather than providing any definitive explanation of "*this* is enlightenment", I attempt to focus on exploring how various cognitive mechanisms behind different enlightenments work. My intent is to cover enough of different things to give a taste of what's out there and what kinds of outcomes might be possible, while acknowledging that there's also a lot that I have no clue of yet.

So this is not trying to be anything like "a definitive and complete explanation of the three characteristics"; I don't think anyone could write such a thing, as nobody can have explored all the aspects of all the three. Rather, this is more of a sketch of those aspects of the three characteristics which I think I have some understanding of.

In particular, this explanation strongly emphasizes no-self and unsatisfactoriness, which I feel I have a better understanding of. Impermanence, which some approaches consider the very core characteristic, ends up relatively neglected. Apologies to any impermanence fans - maybe some day I'll come back to write more about it.

But let's get started with talking about no-self.

No-self

No-self is a confusing term, since it can easily be interpreted as the claim that, well, there is no self. But at least on one interpretation of Buddhism, the claim is much more subtle. Here's an excerpt from the article No-self or Not-self?, by the Buddhist monk Thanissaro Bhikkhu:

> In fact, the one place where the Buddha was asked point-blank whether or not there was a self, he refused to answer. When later asked why, he said that to hold either that there is a self or that there is no self is to fall into extreme forms of wrong view that make the path of Buddhist practice impossible. Thus the question should be put aside. [...]
>
> So, instead of answering "no" to the question of whether or not there is a self — interconnected or separate, eternal or not — the Buddha felt that the question was misguided to begin with.

Now, there are many interpretations of Buddha's teaching, and what the Buddha even really said in the first place; other people will offer a different kind of an account. But let's suppose that this particular interpretation *is* correct. What might it mean?

Well, clearly people *feel* that something like a "self exists". But rather than arguing about whether or not a self exists, one should investigate the mechanisms by which the experience of having a self is constructed. Once those cognitive algorithms are understood, one knows what creates a *feeling* of having a self - and then there is nothing more to explain.

Of course, in Buddha's day, they did not have cognitive science and a theory of neural networks, so he was unable to express his position in those terms. They did, however, have well-developed meditative techniques. And those techniques could be used to investigate how the experience of having a self was developed.

Now, one might reasonably ask, if the question of "does the self exist" is misleading, then why is this often phrased in the form of the claim that the self does *not* exist?

"The self does not exist in the way you think"

In my daily experience, it generally feels like there exists a distinct "me". There is *someone*, an "I" who sees what I see, hears what I hear, feels what I feel. It feels like I can generally make choices, consider information, act according to my best judgment. It feels that there's a meaningful sense in which the same me existed yesterday, and will continue to exist tomorrow. If you were to make a copy of me that was atom-to-atom identical, I might intuitively feel that there would exist a distinct difference between the original me and the copy. We might be exactly identical and act exactly the same, but there would still be a different *experiencer*.

But I also know about the scientific "multi-agent" models of mind, described briefly in my last post and more extensively in earlier ones, where different subsystems within the brain are responsible for my actions. In those models, there is no privileged subsystem in charge of making decisions. Different subsystems take charge at different times, based on a preconscious selection process which is not under the control of any particular subsystem. There is also no particular subsystem which could be singled out as *the* one experiencing things. Rather, anything which makes it to consciousness is broadcast into many different subsystems, each of which can do different things with that information.

So experientially, I feel like I have a self which works in a particular way. Science suggests that my mind actually works in a different way: e.g. decisions are made by a distributed collection of semi-independent subsystems, rather than by a distinct "deciding self". So some might claim that the self *as I intuitively experience it* does not exist, as the intuitive conception does not match reality.

For example, *The Manual of Insight* is a treatise on meditation by the Theravada Buddhist monk Mahasi Sayadaw, who had a significant impact on spreading insight meditation in the West. The book quotes the Theravada scripture of Paṭisambhidāmagga as saying, in its elaboration of no-self, that:

> There is no self that is able to control, to own, to feel, to give orders, to behave according to one's will, no self that is everlasting, or that is the agent of going, seeing, and so on. [...]
>
> [As a result of meditation practice, one comes to see the mind as] empty of self (suññato). Here "self" (atta) means an entity that is the owner of the body, permanently residing in the body, the agent of going, seeing, and so on, the agent who feels pleasant and unpleasant feelings, able to give any orders, and able to exercise mastery. Such an entity, which is [a product of one's] speculation, belief, or obsession, may be called being, soul, ego, or self.

Likewise, Daniel Ingram, meditation teacher and author of the widely-read book *Mastering the Core Teachings of the Buddha,* writes that (emphasis mine):

> The original Pali term, anatta, means literally "not-self". This same term is also rendered by other authors in other ways, some of which can be extremely problematic, such as egolessness, a terribly problematic term, since ego as understood in the Western psychological sense is not the referent of the conception of "self" targeted in Buddhism. Another problematic rendering of this term is "emptiness". Emptiness, for all its mysterious-sounding connotations, means that **reality is empty of, devoid of, or lacking a permanent, separate, independent, acausal, autonomous self**. It doesn't mean that reality is not there, but that reality is not there in the way it may appear to us to be. [...]
>
> It's not that the constellation labeled "me", or "you", a grouping of physical and mental components, does not exist and function in some ordinary sense. **It's that none of those components exist independently or acausally, which is how ignorance conceives of them.** Ultimate unfindability of the components of reality in no way precludes their conventional existence!

Intellectually many people do not think that their self has an acausal existence, independent of the laws of physics. But the kind of understanding one can get from meditation is different. As I will discuss, the ways that specific subsystems react to various situations is linked to their model of the self. Normally, even if you intellectually understand that you do not have an acausally acting self, your mind cannot directly *see* the actual causality. Many of the subconscious models driving your behavior will only update if they are forced to directly witness evidence contradicting their old assumptions. (For a previous discussion of this in the context of psychotherapy and emotional beliefs, see my review of *Unlocking the Emotional Brain*.)

Early insights into no-self

Recall again the model that the content of consciousness roughly corresponds to a "global workspace" which contains information submitted by different subsystems. In normal circumstances, there are some objects in the stream of experience (global workspace) which the overall system treats as being more "me" than the rest. For

example, many people experience themselves as inhabiting a space somewhere behind their eyes, looking at the world from that location.

Suppose that I now do some kind of practice where I examine this experience in more detail. Here is a simple one:

1. Look at an object in front of you. Spend a moment simply examining its features.
2. Become aware of the sensation of being someone who is looking at this object. While letting your attention rest on the object, try to notice what this sensation of being someone who is looking at the object feels like. Does it have a location, shape, or feel?

You may wish to take a moment to do this right now, before reading about my results.

.

.

.

.

.

.

When I do this kind of exercise, a result that I may get is that there is the sight of the object, and then a pattern of tension behind my eyes. *Something* about the pattern of tension feels like “me” - when I feel that “I am looking at a plant in front of me”, this could be broken down to “there is a tension in my consciousness, it feels like the tension is what’s looking at the plant, and that tension feels like me”.

Your result may be different from this. You may find yourself identifying with another sensation, or you might not be able to hone down on any particular sensation on the first try… but if you are like most people, you probably still have *some* kind of a feeling of looking out at the world.

My guess is that this sensation is a tag coming from some subsystem whose task is to keep track of one’s spatial location relative to their surroundings. We know that there are multiple such systems in the brain, and that these systems getting out of sync - one system indicating a particular location and another indicating a differing location - can create the feeling of an out-of-body experience. In computer terms, sensory data comes in, and then some subsystem parses that sensory data and indicates where one’s “I” is located, passing this tag for other subsystems to use. Going by the previous example of me feeling a tension around my eyes that feels like me looking at the plant, we might think that something like the following is happening:

- Subsystem 1 sends the sight of a plant into the global workspace
- Subsystem 2 sends the feeling of tension around the eyes into the global workspace
- Subsystem 3 tags the tension as my current location, and binds all of these percepts together as an experience of “I am seeing the plant”, which is also sent to the global workspace

An interesting thing is that the subsystems in the brain seem to take the tag as an ontological fact. Suppose that someone hands you a map of your surroundings, and has helpfully marked your current location with a red tag saying “YOU ARE HERE”.

But suppose that you now get a little confused. Rather than taking the spot with red ink as *indicating* your location in your physical world, you take the red spot on the map to *be* your physical location. That is, you think that you *are* the “YOU ARE HERE” tag, looking at the rest of the map *from* the red ink itself.

But of course, the fact that you are seeing the above picture, means that you cannot be looking *from* the red ink in the picture. The map includes the red ink, meaning that the person who is looking at it is actually *outside* the map.

Likewise, people tend to have a sensation of looking at the world from behind their eyes; but they are actually aware *of* the sensation, as opposed to being aware *from* it. It is a computational representation of a location, rather than being the location itself. Still, once this representation is fed into other subsystems in the brain, those subsystems will treat the tagged location as the one that they are “looking at the sense data from”, as if they had been fed a physical map of their surroundings with their current location marked.

But a particular tag in the sense data is not actually where they are looking at it from; for one, the visual cortex is located in the back of the head, rather than right behind the eyes. Furthermore, any visual information is in principle just a piece of data that has been fed into a program running in the brain. If we think of cognitive programs as analogous to computer programs, then a computer program that is fed a piece of data isn't really "looking at" the data "from" any spatial direction.

In vipassana-style meditation, you train your attention to dissect components of your experience into smaller pieces. (Vipassana is commonly translated as insight meditation, but here I treat it as a particular subcategory of insight meditation.) In third-person terms, this probably trains up pattern-detectors which can monitor the content of the global workspace in extreme detail. Eventually, there’s sufficient clarity about the sense of location for low-level schemas to pick up on the inherent contradiction involved in looking *at* something which the system is supposedly looking *out from*.

The opposite strategy is commonly associated with what are so-called nondual techniques. Instead of training an analytical, attention-controlled part of the mind to examine the sense of self, the nondual route is to nudge the mind into a state where

those analytical parts of the brain become less active. As those parts also produce the sense of ‘the observer’ in the first place, attenuating their activity can offer a glimpse into a state of consciousness where that sensation is lacking. Some versions of this approach seem to be tapping into some of the same machinery which causes people to experience a state of flow, as flow states also seem to involve a downregulation in both analytical thought and the sense of self.

Frequently, the sense of self being diminished in this way is a sufficiently interesting experience that the analytical subsystems kick back online to make sense of it - but over time, one can train oneself to experience more such glimpses, until there is a broader shift.

It is not clear to me to what extent these routes lead to exactly the same result. It seems to me that both eventually end up at a state where the sensations tagging one’s physical location still continue to be produced, and can be used as an aid for spatial reasoning, but the system no longer intrinsically identifies with them. Rather, the sensations are seen as being constructed by a machinery which is independent of the actual stream of sensory input.

But there seem to be some differences in how you reach that place. For the sake of analogy, let's pretend that the machinery is a hologram projector, painting a realistic image of a person in the middle of a room. The vipassana path would correspond to looking very closely at all the details of the hologram, until you noticed discrepancies in how it was created. That would give you a detailed insight into how exactly the projector used light to draw the image, but would be rather slow. In contrast, the nondual route involves just turning the projector off for a moment - making it very obvious that the hologram was in fact a hologram, but telling you much less of how it was built.

Another difference is the no-self versus all-self interpretation. Some schools say that this kind of practice leads you to realizing that there is no self; other schools, generally more associated with Hinduism than Buddhism, say that they lead you to realizing that all is self. (Western philosophy has the corresponding concepts of closed, open and empty individualism.)

Some of the end results from both paths are described in a similar way, however. For example, a common metaphor about the result of some varieties of practice is that of “being the sky rather than the clouds”. Below is one formulation of it. The outcome seems to be that rather than identifying with the sensations of the supposed observer, one’s identity shifts to *the entire field of consciousness itself* (in line with the thing about a program reading a file not having any location that would be defined in terms of the file):

> One way of describing the experience of glimpsing in effortless mindfulness practice is to use the metaphor of a cloud. You may have felt as if you have been living in a cloud; maybe it feels like a storm cloud a lot of the time. See if you can feel the boundary and fogginess of this cloud that you call “me.” You may have been trying to feel better by cleaning up the cloud of your mind by replacing negative thoughts with positive thoughts and developing good attitudes. You may have tried to calm your body and mind to make your brain as clear as possible. Within your cloud are storms, old traumas, emotional challenges, and relationships of all types. Each time you change these things and clean up one area of the cloud, it seems that another foggy issue or thunderous problem arises.

> Effortless mindfulness does not begin with dissolving the cloud, calming it, or trying to transform its contents. The glimpsing method of effortless mindfulness begins with awake awareness stepping out of the cloud, shifting, dropping, or opening to discover that you are also the open sky of awake awareness! When you shift out of this cloud of the emotional or small mind and discover this spaciousness of still, quiet, alert awareness, it's a great relief. You can realize that you are the sky, and the cloudy emotions and thoughts are everchanging weather.
>
> [...] As we reach the fullness of effortless mindfulness, we will discover open-hearted awareness and ways to naturally embrace and welcome all emotions and parts of ourselves. [...] After all, all weather comes and goes, and no storm ever hurt the sky.
>
> (Loch Kelly, The Way of Effortless Mindfulness)

Or this more concrete description, quoted in a paper on meditation-induced changes to the sense of self (Lindahl & Britton 2019):

> So, [the retreat] was in the spring and I was doing some raking leaves, and just as I was raking, this really profound feeling of 'this is all me' came to me. And so the 'this is all me' — what that means is that my identity is literally everything that I could see through my eyes. So, the rake that I was holding in my hands was me. The ground that I was raking was me. The feet that I could see down at the bottom of my body, that was me. The steps up to the residence, that was me. The sky was me. The trees were me. And so, everything was just 'me'. And that there wasn't really anything else. It was all just 'me'. [...] Those experiences that I related about what I would call kenshō experiences, there was no viewer in those — it was just what was there, and there was no viewer observing it.

Here is how I would rephrase these reports in third-person terms. Normally, there is a flow of information within the global workspace: mental objects representing sensory information, thoughts, and some objects encoding a sense of there being someone who watches the senses. These kinds of experiences are a part of a process where the system reorients its assumptions to recognize that there is no homunculus sitting behind the eyes and watching everything.

From an external point of view, we can say that *your conscious mind - or "you" - consists of everything that is in the global workspace, and no particular piece of mental content is more or less "you" than the others are*. If you see a rake in your hands, then there is a process within your brain generating that visual experience. The experience exists as a part of your mind. Likewise, the experience of there being a *someone* who is *having* that experience, is generated by a process within your brain, and exists as a part of your mind. *Everything that you ever experience is mental content generated by your brain,* as opposed to you having direct access to reality.

Now, in Loch Kelly's quote above, there is the suggestion that changing one's identification to the entire field of consciousness will also change how one relates to negative experiences. Exactly why this would happen is an important question, and I will come back to it later. For now, let's look a bit more at why getting such an experience can be so difficult.

The self as a tool for planning

A thing that might happen, once the above has been explained to you, is that you put a lot of effort into *intellectually* figuring out the contradiction between experiencing something that you also identify with, and then figuring out what must be going on instead. This kind of theorizing can be useful for purposes of writing articles such as this one. But you cannot use theorizing alone to put your brain into a no-self state by convincing your brain of the contradiction. Meditation teachers may explicitly warn you that it is impossible for your thinking mind to comprehend no-self states in such a way that would cause you to actually *experience* them.

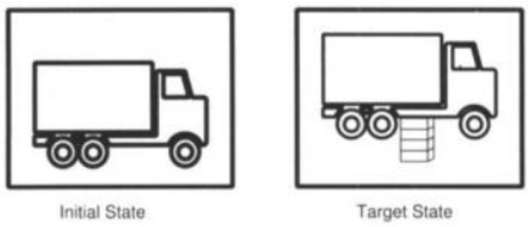
I suspect that a part of this is because the subsystems in the brain used for this kind of theorizing take the sense of self as input. As a result, them being active tends to put the mind in a state where it identifies more strongly with the sensations of a self.

Going back to the map analogy, consider the route-finding algorithm included in Google Maps: you give it a starting location, an end location, some parameters of what kind of a route you prefer, and it then finds you the best route that meets those criteria.

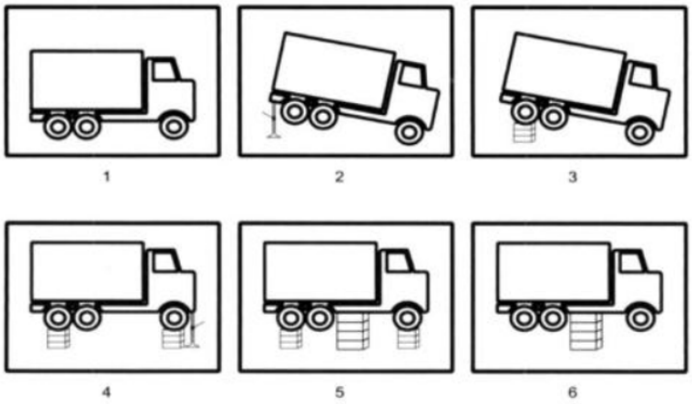
I suggested that the sense of the observer is like a point in a map, saying “YOU ARE HERE”, and that one of the goals of practice was coming to see that the point that’s marked on the map cannot actually be “your” real location. That is, some part of your mind stops treating the red ink on the map as being identical to where you are. But a route-finding algorithm does not have the option of treating the starting point of a route as anything else than as the starting point of a route. Its *entire purpose* is to assume that the “YOU ARE HERE” really does correspond to your real location, and to then plot a route from there. If it didn’t, it wouldn’t be a route-finding algorithm anymore.

What I am calling “intellectual” parts of your brain, seem to be similar to route-finding algorithms. Their purpose is to figure out a path from where you are now, to some desired target state.

The literature on expertise suggests that people figure out novel tasks by running mental simulations of how to get from a current state to a target state, and then trying to carry out a sequence that they have successfully simulated (Klein 1999). For example, you might be faced with a truck sitting on the ground. Using a jack and concrete blocks, you want to get it up on the air on a column of blocks.



You mentally go through different options, until you figure out a sequence of steps that gets you to the end result. When you find something that seems promising enough, you give it a shot.



Now, in the example of a truck, your reasoning can happen purely in terms of what is going to happen to the truck; the same process would work exactly the same regardless of whether you or someone else was doing it. But what happens if you set the goal of "*I* want to get to a state where *I* experience no sense of self?"

This again fires up the parts of your brain that carry out mental simulations… but just as in the truck example, where they needed to track what was happening to the truck in each step of the sequence, they now need to track whether or not *you* are experiencing a sense of self in any given step. This makes it impossible for them to find a state where you wouldn't experience a sense of self, as the very act of trying to plan how to get you there requires instantiating a sense of self that represents you in the simulation!

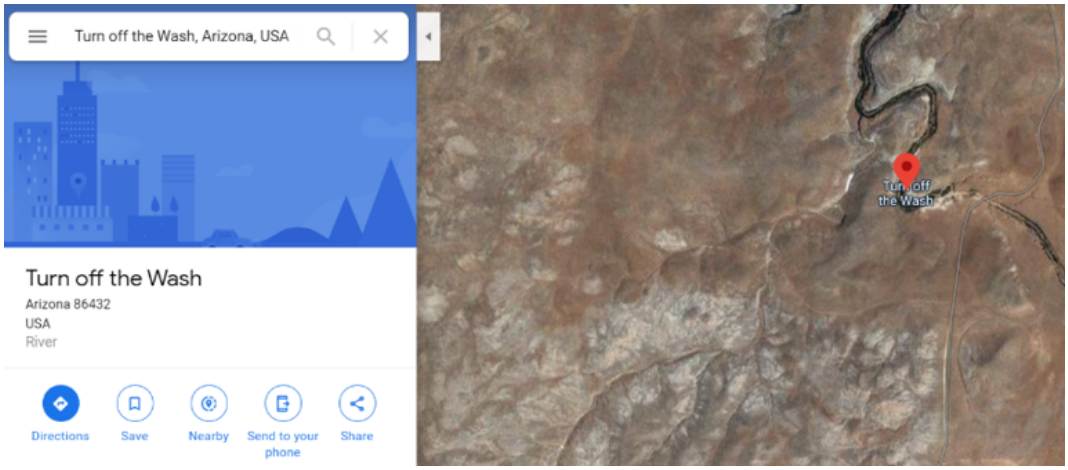
This can make for some frustrating experiences, in that if you once experience a state with a drastically weakened sense of self, it may feel pleasant and you then want to get back to it. But trying to figure out how to get back into it, is exactly the kind of a process that may *prevent* you from getting back. This is part of the reason why some traditions and teachers say things like "in order to get enlightened, you must stop striving for enlightenment", as well as claiming that thinking in terms of outcomes is contrary to the spirit of the practice.

What the planning system would actually need to do to achieve its goal, is to simply turn itself off, so that it stops projecting a sense of self into the global workspace. But it cannot accurately represent this target state, as it parses it as "a state where *I* experience no sense of self". Its representation of the target still includes a sense of someone who either is or is not experiencing a sense of self.

To use the map analogy, this is something like asking Google Maps to route a path to a state where the Google Maps program has been turned off. There is simply no way for it to do that, because the notion of being on or off is not explicitly represented anywhere in the program. The pathfinding routine of Google Maps only reasons in terms of where "you" are in the maps that are loaded into it.

What the route-finding algorithm in Google Maps *can* do, is something like take a map, find the location on it that sounds the closest to "turn yourself off", and plot a route to that. Of course, this will not actually turn it off, but it is something that the algorithm can at least *do*. So upon being given the task of turning itself off, it will plot a route to that location, correctly notice that this is not actually fulfilling the task that it was given, and trash around trying to find a better target location. This corresponds to a meditator thinking something like "oh, how do I get into a no-self state again, oh wait,

if I *try* to get into a no-self state I can't do it, so I have to stop trying to get to it… so now I am going to stop trying to get to it… wait, that is trying again, gahhhhhh."



Google Maps trying to figure out where "turn off" is. This location isn't quite it, but maybe it would at least be getting close?

This runs into the contradiction between the way that we often think about our minds, and the way that our minds actually work. We often have the feeling that at least *some* of the content in our consciousness is something that we can actively choose. Most people don't expect to be able to choose their emotions, but at least the act of *intentionally trying to do something* feels like it should be under conscious control - isn't that what intentionally acting *means*?

But under a multi-agent framework, "trying to do something" simply means that a subsystem is active and pursuing a particular goal. Neither the subsystem itself, nor any other subsystem, has direct access to a command which would turn that subsystem off: the choice of which subsystem to activate or keep running, happens by means of a preconscious selection process. That means that, despite it possibly going against one's naive intuition, it is perfectly possible to consciously intend to do something while also having no conscious control over the fact that you are intending to do so.

As I noted before, there are several approaches to dealing with this problem. For example, flow states typically involve activities that are similar to the truck task, in that they do not require a sense of self. At the same time, the task is challenging enough that it requires one's full attention: in other words, a single planning subsystem uses up the full bandwidth of consciousness, being the only one that projects content to the global workspace. If there was any spare capacity, other planning systems could project self-related thoughts at the same time (e.g. thinking about what to do after the current task is done), thus instantiating a sense of self. Thus, getting the mind into something like a flow state is one way to reduce the sense of self.

On the other hand, some situations just trigger the self-related planning machinery very strongly. In vipassana/mindfulness-style approaches, one frequently ends up creating a sense of being an observer who is detached from their thoughts and emotions. For example, a simple set of "labeling" instructions is just:

1. Notice something in your consciousness.
2. Give it a label, such as “seeing”, “feeling”, or “hearing”.
3. Go back to 1.

In these instructions, the planning machinery is given a goal that it *is* capable of carrying out. Following these instructions does instantiate a sense of self - the planning system needs to monitor the question of “am *I* still labeling my experience”. However, this task constructs an experience where the “I” is merely *observing* other mental content, and that mental content is happening on its own.

This can be particularly useful in situations which are experienced as important or potentially threatening, as those kinds of situations tend to make goal-oriented systems kick in very strongly to help resolve the situation. For people with trauma and ongoing anxiety, this might include even situations with no immediate external concerns; such people may almost constantly be in a state of uncertainty, activating planning systems with the goal of making those unpleasant feelings go away. If one practices dispassionately observing the contents of their mind, even when the content is unpleasant, one can in effect train up a new subsystem that competes with the other subsystems in projecting content to the global workspace. (However, it needs to be noted that training the mind to closely examine unpleasant feelings may also make trauma responses *worse* by bringing more attention to them and interfering with the subsystems that were previously regulating the responses.)

In this, one continues the process of identifying with a self, but the thing that is being identified with shifts to a sense of someone who is just observing everything happening in the mind - which can bring relief from various unpleasant emotions. Once one gets to this kind of a state, the subsystem trained to do this can continue to further investigate the contents of the mind in fine detail… either looking at other characteristics like impermanence or unsatisfactoriness, or turning its focus *on itself*, to deepen the no-self realization by seeing that the observer self that it is projecting is *also* something that can be dis-identified with.

The meditation teacher Michael Taft describes this kind of a turn in his article on Escaping the Observer Trap:

> Many traditions—especially mindfulness meditation—encourage you to observe your sensory experience in a neutral manner. Observe your breathing, observe emotions, observe thoughts, and so on, without reacting to them. This observer technique works really well because it gives you something like an outside perspective on your own experience. You can watch your own mind, your reactions, your emotions, your behavior almost from the perspective of another person, and that is tremendously useful feedback to have. It leads to equanimity, and the tremendous personal growth that mindfulness advocates are always talking about. […]
>
> Taking this observer stance is so useful, in fact, that many teachers stop there and do not talk about the next important step in spiritual development. But there is a hidden problem with the observer technique, which becomes obvious once you think about it. Who is the observer? Who is this person who is behind the binoculars, watching your experience from the outside? This neutral observer you’ve created over time is actually just another—albeit smaller and less neurotic—version of the ego. It’s the sense of being a person who is doing the meditating. You could also call it a meditator ego or an observer ego. Creating this neutral

> observer is very useful, but the goal of meditation is not to create a new meditator ego, it's to see through the illusion of the ego entirely.
>
> It is quite common for even very dedicated mindfulness students in observation-based traditions to get stuck in observer mode forever. I have seen it over and over in my experience. Being the observer, a neutral meditator ego, is not such a bad place to be; certainly it is much preferable to the unconscious, robotic mode of life lived without any self-reflection. However, it impedes all deeper progress toward real awakening. So the only way forward is to let go of the observer ego; to release the sense of being a person who is doing a meditation. [...]
>
> To release yourself from the observer trap, begin by realizing that the observer, however comfortable or habitual, is still just another version of the ego. You've spent endless hours watching your breath and your emotions and your thoughts. Now it's time to watch the watcher instead. You have to observer the observer. You do this, in typical mindfulness style, by carefully deconstructing the components of the observer itself.
>
> The observer ego is constructed out of the same components as the everyday ego, but on a smaller scale. The everyday mind has thoughts about all sorts of stuff, the observer has thoughts about how the mediation is going, or how long until this sit is over. The everyday ego has emotions about all sorts of stuff, but observer has emotions about how this sit is going, or even blissful feelings of love and joy. The everyday ego has all sorts of body sensations, but the observer has a very special set of body sensations: the sensations of where he/she imagines awareness is located. [...] So to overcome the observer problem and get unstuck in your practice, closely observe the sensations (i.e. the thoughts and feelings) associated with the observer ego.

This is the second post of the "a non-mystical explanation of insight meditation and the three characteristics of existence" series. The next post in the series is "Craving, suffering, and predictive processing".

Craving, suffering, and predictive processing (three characteristics series)

This is the third post of the "a non-mystical explanation of insight meditation and the three characteristics of existence" series. I originally intended this post to more closely connect no-self and unsatisfactoriness, but then decided on focusing on unsatisfactoriness in this post and relating it to no-self in the next one.

Unsatisfactoriness

In the previous post, I discussed some of the ways that the mind seems to construct a notion of a self. In this post, I will talk about a specific form of motivation, which Buddhism commonly refers to as *craving* (*taṇhā* in the original Pali). Some discussions distinguish between craving (in the sense of wanting positive things) and aversion (wanting to avoid negative things); this article uses the definition where both desire and aversion are considered subtypes of craving.

My model is that craving is generated by a particular set of motivational subsystems within the brain. Craving is not the *only* form of motivation that a person has, but it normally tends to be the loudest and most dominant. As a form of motivation, craving has some advantages:

- People tend to experience a strong craving to pursue positive states and avoid negative states. If they had less craving, they might not do this with an equal zeal.
    - To some extent, craving looks to me like a mechanism that shifts behaviors from exploration to exploitation.
    - In an earlier post, Building up to an Internal Family Systems model, I suggested that the human mind might incorporate mechanisms that acted as priority overrides to avoid repeating particular catastrophic events. Craving feels like a major component of how this is implemented in the mind.
- Craving tends to be automatic and visceral. A strong craving to eat when hungry may cause a person to get food when they need it, even if they did not intellectually understand the need to eat.

At the same time, craving also has a number of disadvantages:

- Craving superficially looks like it cares about *outcomes*. However, it actually cares about *positive or negative feelings* (valence). This can lead to behaviors that are akin to wireheading in that they suppress the unpleasant feeling while doing nothing about the problem. If thinking about death makes you feel unpleasant and going to the doctor reminds you of your mortality, you may avoid doctors - even if this actually *increases* your risk of dying.
- Craving narrows your perception, making you only pay attention to things which seem immediately relevant for your craving. For example, if you have a craving for sex and go to a party with the goal of finding someone to sleep with, you

may see everyone only in terms of “will sleep with me” or “will not sleep with me”. This may not be the best possible way of classifying everyone you meet.
- Strong craving may cause premature exploitation. If you have a strong craving to achieve a particular goal, you may not want to do anything that looks like moving away from it, even if that would actually help you achieve it better. For example, if you intensely crave a feeling of accomplishment, you may get stuck playing video games that make you feel like you are accomplishing something, even if there was something else that you could do that was more fulfilling in the long term.
- Multiple conflicting cravings may cause you to thrash around in an unsuccessful attempt to fulfill all of them. If you crave to get your toothache fixed, but also a craving to avoid dentists, you may put off the dentist visit even as you continue to suffer from your toothache.
- Craving seems to act in part by creating self-fulfilling prophecies; making you strongly believe that you are going to achieve something, so as to cause you to do it. The stronger the craving, the stronger the false beliefs injected into your consciousness. This may warp your reasoning in all kinds of ways: updating to believe an unpleasant fact may subjectively feel like you are allowing that fact to become true by believing in it, incentivizing you to come up with ways to avoid believing in it.
- Finally, although craving is often motivated by a desire to avoid unsatisfactory experiences, it is actually the very thing that causes dissatisfaction in the first place. Craving assumes that negative feelings are intrinsically unpleasant, when in reality they only become unpleasant when craving resists them.

Given all of these disadvantages, it may be a good idea to try to shift one’s motivation to be more driven by subsystems that are not motivated by craving. It seems to me that everything that can be accomplished via craving, can *in principle* be accomplished by non-craving-based motivation as well.

Fortunately, there are several ways of achieving this. For one, a craving for some outcome X tends to implicitly involve at least two assumptions:

1. achieving X is necessary for being happy or avoiding suffering
2. one cannot achieve X *except* by having a craving for it

Both of these assumptions are false, but subsystems associated with craving have a built-in bias to selectively sample evidence which supports these assumptions, making them frequently feel compelling. Still, it is possible to give the brain evidence which lets it know that these assumptions are wrong: that it is possible to achieve X without having craving for it, *and* that one can feel good regardless of achieving X.

Predictive processing and binocular rivalry

I find that a promising way of looking at unsatisfactoriness and craving and their impact on decision-making comes from the predictive processing (PP) model about the brain. My claim is not that craving would work *exactly* like this, but something roughly like this seems like a promising analogy.

Good introductions to PP include this book review as well as the actual book in question... but for the purposes of this discussion, you really only need to know two things:

- According to PP, the brain is constantly attempting to find a model of the world (or hypothesis) that would both explain and predict the incoming sensory data. For example, if I upset you, my brain might predict that you are going to yell at me next. If the next thing that I hear is you yelling at me, then the prediction and the data match, and my brain considers its hypothesis validated. If you do *not* yell at me, then the predicted and experienced sense data conflict, sending off an error signal to force a revision to the model.
- Besides changing the model, another way in which the brain can react to reality not matching the prediction is by changing reality. For example, my brain might predict that I am going to type a particular sentence, and then fulfill that prediction by moving my fingers so as to write that sentence. PP goes so far as to claim that this is the mechanism behind *all* of our actions: a part of your brain predicts that you are going to do something, and then you do it so as to fulfill the prediction.

Next I am going to say a few words about a phenomenon called binocular rivalry and how it is interpreted within the PP paradigm. I promise that this is going to be relevant for the topic of craving and suffering in a bit, so please stay with me.

Binocular rivalry, first discovered in 1593 and extensively studied since then, is what happens when your left eye is shown one picture (e.g. an image of Isaac Newton), and your right eye is shown another (e.g. an image of a house) in the right. People report that their experience keeps alternating between seeing Isaac Newton and seeing a house. They might also see a brief mashup of the two, but such Newton-houses are short-lived and quickly fall apart before settling to a stable image of either Newton or a house.

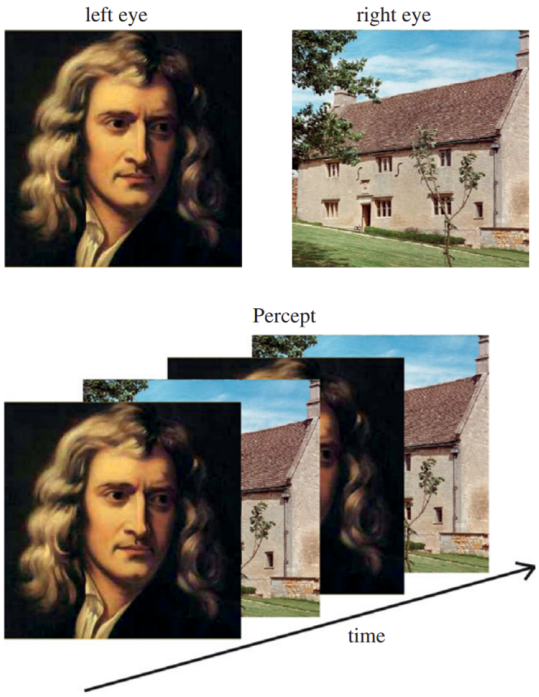


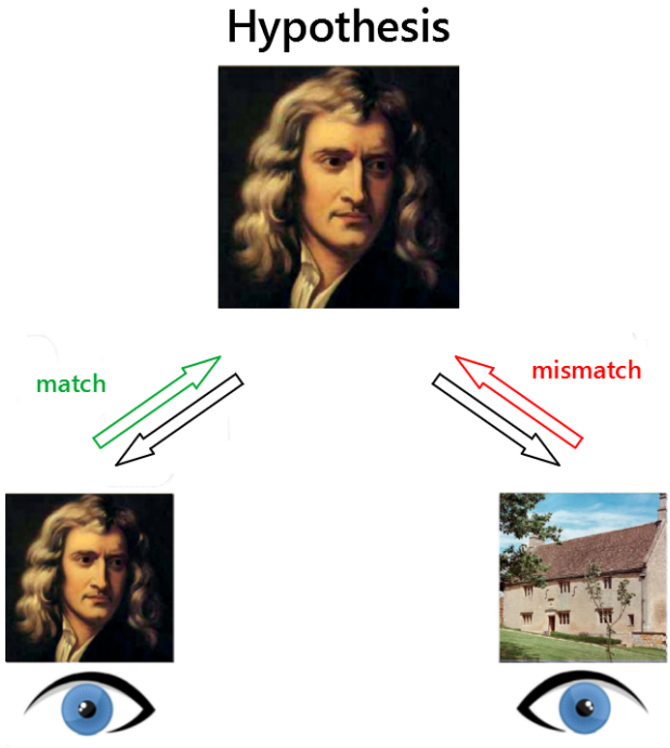
Image credit: Schwartz et al. (2012), Multistability in perception: binding sensory modalities, an overview. Philosophical Transactions of the Royal Society B, 367, 896-905.

Predictive processing explains what's happening as follows. The brain is trying to form a stable hypothesis of what exactly the image data that the eyes are sending represents: is it seeing Newton, or is it seeing a house? Sometimes the brain briefly considers the hybrid hypothesis of a Newton-house mashup, but this is quickly rejected: faces and houses do not exist as occupying the same place at the same scale at the same time, so this idea is clearly nonsensical. (At least, nonsensical outside highly unnatural and contrived experimental setups that psychologists subject people to.)

Your conscious experience alternating between the two images reflects the brain switching between the hypotheses of "this is Isaac Newton" and "this is a house"; the currently-winning hypothesis is simply what you experience reality as.

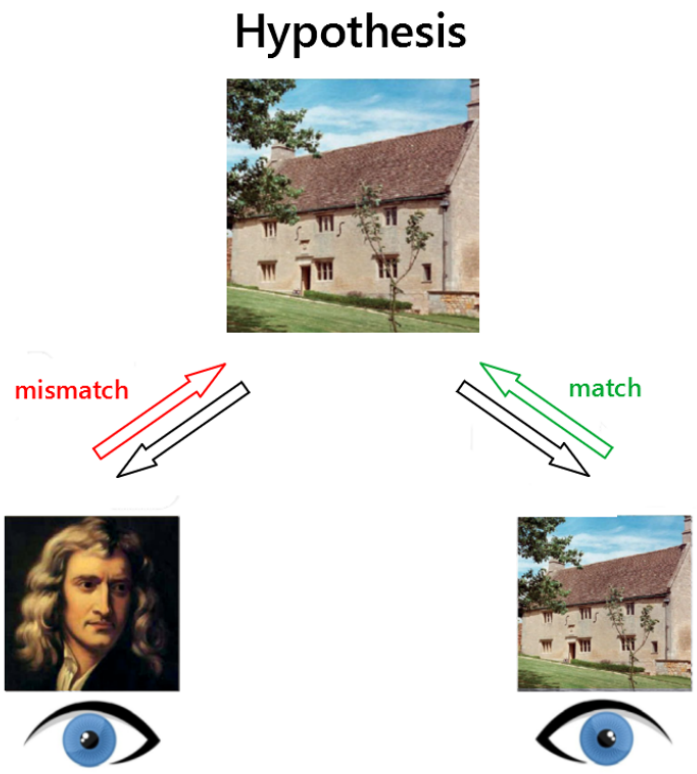
Suppose that the brain ends up settling on the hypothesis of "I am seeing Isaac Newton"; this matches the input from the Newton-seeing eye. As a result, there is no error signal that would arise from a mismatch between the hypothesis and the Newton-seeing eye's input. For a moment, the brain is satisfied that it has found a workable answer.

However, if one really was seeing Isaac Newton, then the *other* eye should not keep sending an image of a house. The hypothesis and the house-seeing eye's input *do* have a mismatch, kicking off a strong error signal which lowers the brain's confidence in the hypothesis of "I am seeing Isaac Newton".

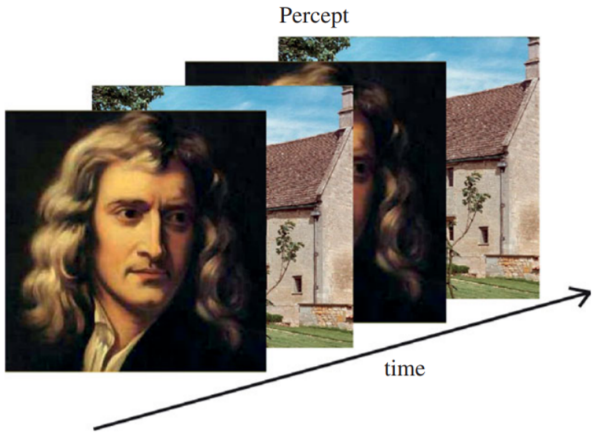


The brain goes looking for a hypothesis which would better satisfy the strong error signal… and then finds that the hypothesis of "I am seeing a house" serves to entirely quiet the error signal from the house-seeing eye. Success?

But even as the brain settles on the hypothesis of "I am seeing a house", this then contradicts the input coming from the *Newton-seeing eye.*



The brain is again momentarily satisfied, before the incoming error signal from the hypothesis/Newton-eye mismatch drives down the probability of the “I am seeing a house” hypothesis, causing the brain to eventually go back to the “I am seeing Isaac Newton” hypothesis... and then back to seeing a house, and then to seeing a Newton, and...



One way of phrasing this is that there are two subsystems, each of which are transmitting a particular set of constraints (about seeing Newton and a house). The brain is then trying and failing to find a hypothesis which would fulfill both sets of constraints, while *also* respecting everything else that it knows about the world.

As I will explain next, my feeling is that something similar is going on with unsatisfactoriness. Craving creates constraints about what the world should be like,

and the brain tries to find an action which would fulfill all of the constraints, while also taking into account everything else that it knows about the world. Suffering/unsatisfactoriness emerges when all of the constraints are impossible to fulfill, either because achieving them takes time, or because the brain is unable to find any scenario that could fulfill all of them even in theory.

Predictive processing and psychological suffering

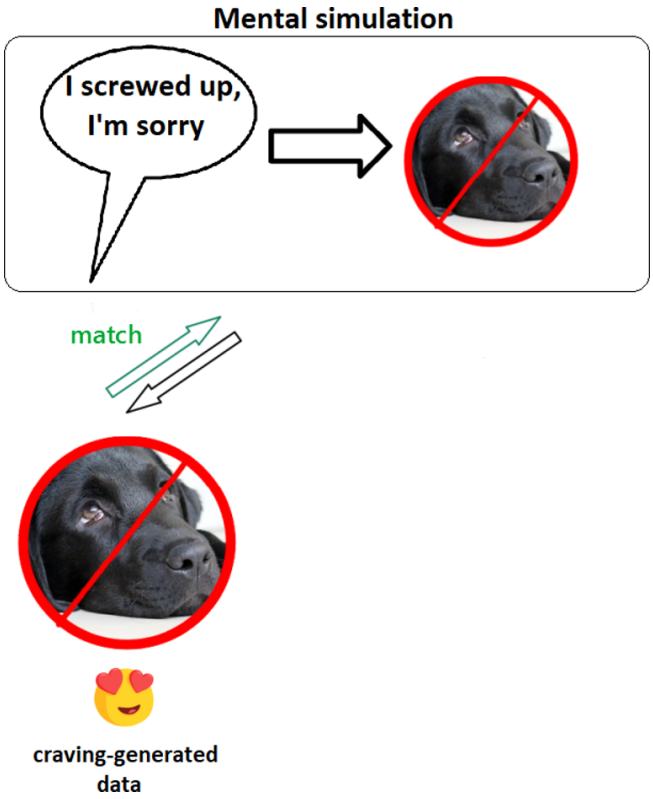
There are two broad categories of suffering: mental and physical discomfort. Let's start with the case of psychological suffering, as it seems most directly analogous to what we just covered.

Let's suppose that I have broken an important promise that I have made to a friend. I feel guilty about this, and want to confess what I have done. We might say that I have a craving to avoid the feeling of guilt, and the associated craving subsystem sends a prediction to my consciousness: I will stop feeling guilty.

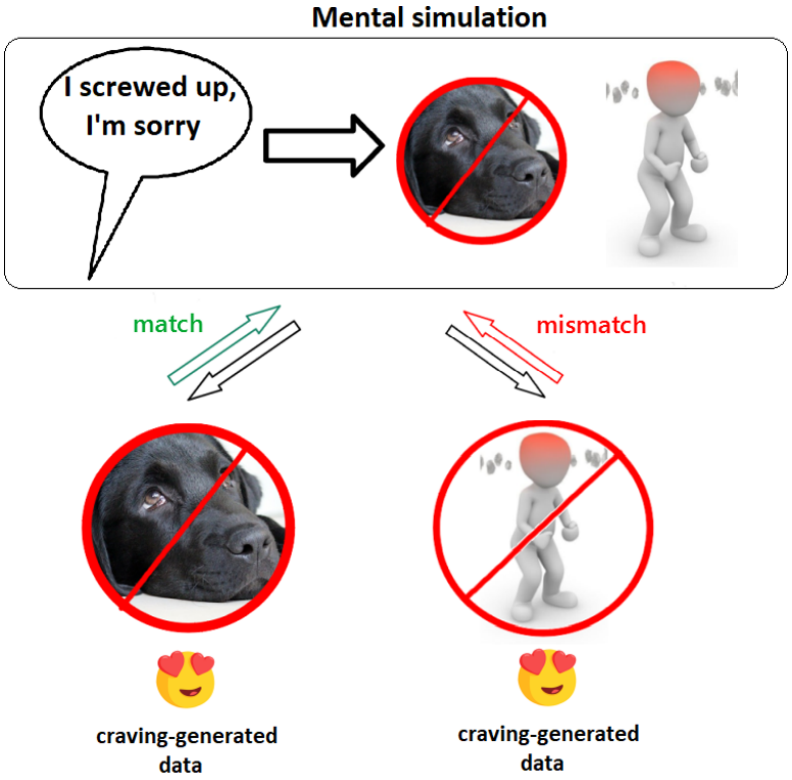
In the previous discussion, an inference mechanism in the brain was looking for a hypothesis that would satisfy the constraints imposed by the sensory data. In this case, the same thing is happening, but

- the hypothesis that it is looking for is a possible action that I could take, that would lead to the constraint being fulfilled
- the sensory data is not actually coming from the senses, but is internally generated by the craving and represents the outcome that the craving subsystem would like to see realized

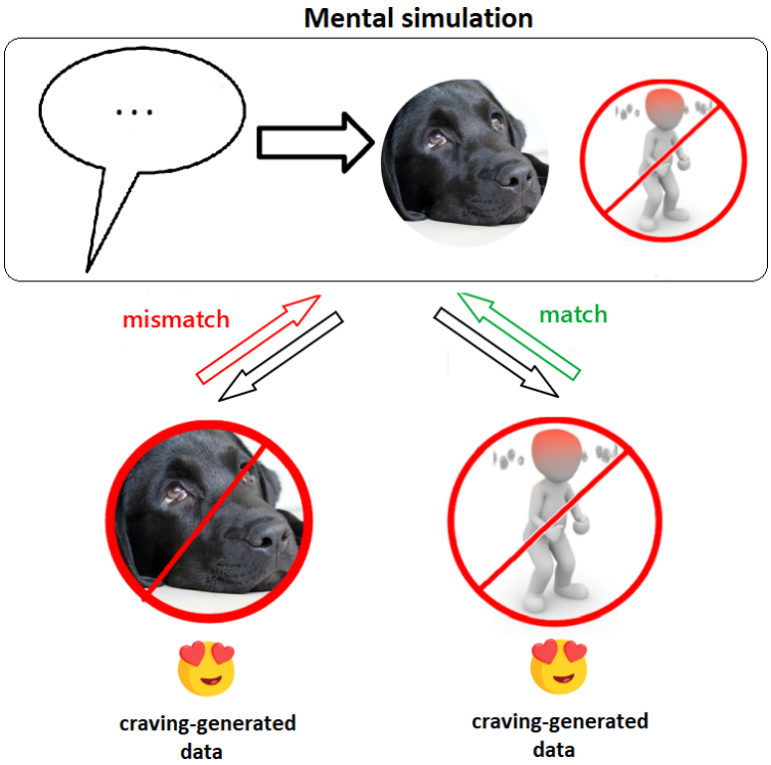
My brain searches for a possible world that would fulfill the provided constraints, and comes up with the idea of just admitting the truth of what I have done. It predicts that if I were to do this, I would stop feeling guilty over not admitting my broken promise. This satisfies the constraint of not feeling guilty.



However, as my brain further predicts what it expects to happen as a consequence, it notes that my friend will probably get quite angry. This triggers another kind of craving: to not experience the feeling of getting yelled at. This generates its own goal/prediction: that nobody will be angry with me. This acts as a further constraint for the plan that the brain needs to find.



As the constraint of “nobody will be angry at me” seems incompatible with the plan of “I will admit the truth”, this generates an error signal, driving down the probability of this plan. My brain abandons this plan, and then considers the alternative plan of “I will just stay quiet and not say anything”. This matches the constraint of “nobody will be angry at me” quite well, driving down the error signal from that particular plan/constraint mismatch… but then, if I don’t say anything, I will continue feeling guilty.



The mismatch with the constraint of “I will stop feeling guilty” drives up the error signal, causing the “I will just stay quiet” plan to be abandoned. At worst, my mind may find it impossible to find any plan which would fulfill both constraints, keeping me in an endless loop of alternating between two unviable scenarios.

There are some interesting aspects about the phenomenology of such a situation, which feel like they fit the PP model quite well. In particular, it may feel like *if I just focus on a particular craving enough, thinking about my desired outcome hard enough will make it true*.

Recall that under the PP framework, goals happen because a part of the brain *assumes* that they will happen, after which it changes reality to *make* that belief true. So focusing really hard on a craving for X makes it feel like X will become true, *because the craving is literally rewriting an aspect of my subjective reality to make me think that X will become true*.

When I focus hard on the craving, I am temporarily guiding my attention away from the parts of my mind which are pointing out the obstacles in the way of X coming true. That is, those parts have less of a chance to incorporate *their* constraints into the plan that my brain is trying to develop. This momentarily reduces the motion away from this plan, making it seem more plausible that the desired outcome will in fact become real.

Conversely, letting go of this craving, may feel like it is *literally making the undesired outcome more real*, rather than like I am coming more to terms with reality. This is most obvious in cases where one has a craving for an outcome that is impossible for certain, such as in the case of grieving about a friend’s death. Even after it is certain that someone is dead, there may still be persistent thoughts of *if only I had done X,* with an implicit additional flavor of *if I just want to have done X really hard, things will*

change, and I can't stop focusing on this possibility because my friend needs *to be alive.*

In this form, craving may lead to all kinds of rationalization and biased reasoning: a part of your mind is literally making you believe that X is true, because it wants you to find a strategy where X is true. This hallucinated belief may constrain all of your plans and models about the world in the same sense as *getting direct sensory evidence about X being true* would constrain your brain's models. For example, if I have a *very* strong urge to believe that someone is interested in me, then this may cause me to interpret *any* of his words and expressions in a way compatible with this belief, regardless of how implausible and far-spread of a distortion this requires.

The case of physical pain

Similar principles apply to the case of physical pain.

We should first note that pain does not necessarily need to be aversive: for example, people may enjoy the pain of exercise, hot spices or sexual masochism. Morphine may also have an effect where people report that they still experience the pain but no longer mind it.

And, relevant for our topic, people practicing meditation find that by shifting their attention *towards* pain, it can become *less* aversive. The meditation teacher Shinzen Young writes that

> ... pain is one thing, and resistance to the pain is something else, and when the two come together you have an experience of suffering, that is to say, 'suffering equals pain multiplied by resistance.' You'll be able to see that's true not only for physical pain, but also for emotional pain and it's true not only for little pains but also for big pains. It's true for every kind of pain no matter how big, how small, or what causes it. Whenever there is resistance there is suffering. As soon as you can see that, you gain an insight into the nature of "pain as a problem" and as soon as you gain that insight, you'll begin to have some freedom. You come to realize that as long as we are alive we can't avoid pain. It's built into our nervous system. But we can certainly learn to experience pain without it being a problem. (Young, 1994)

What does it mean to say that *resisting* pain creates suffering?

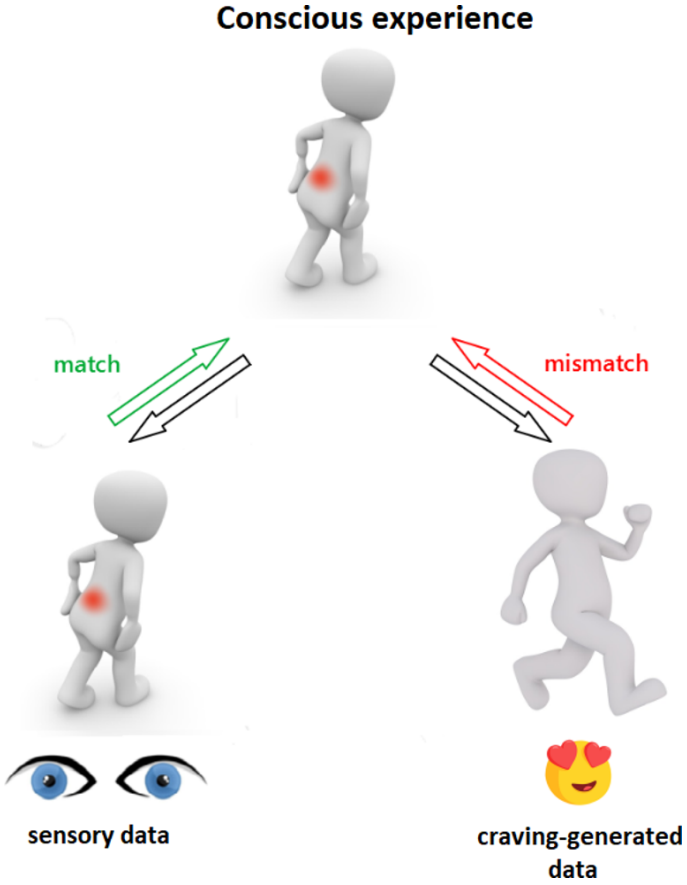
In the discussion about binocular rivalry, we might have said that when the mind settled on a hypothesis of seeing Isaac Newton, this hypothesis was *resisted* by the sensory data coming from the house-seeing eye. The mind would have settled on the hypothesis of "I am seeing Isaac Newton", if not for that resistance. Likewise, in the preceding discussion, the decision to admit the truth was resisted by the desire to not get yelled at.

Suppose that you have a sore muscle, which hurts whenever you put weight on it. Like sensory data coming from your eyes, this constrains the possible interpretations of what you might be experiencing: your brain might settle on the hypothesis of "I am feeling pain".

But the experience of this hypothesis then triggers a *resistance* to that pain: a craving subsystem wired to detect pain and resist it by projecting a form of internally-generated sense data, effectively claiming that you are *not* in pain. There are now

again two incompatible streams of data that need to be reconciled, one saying that you are in pain, and another which says that you are not.

In the case of binocular rivalry, both of the streams were generated by sensory information. In the discussion about psychological suffering, both of the streams were generated by craving. In this case, craving generates one of the streams and sensory information generates the other.



On the left, a persistent pain signal is strong enough to dominate consciousness. On the right, a craving for not being in pain attempts to constrain consciousness so that it doesn't include the pain.

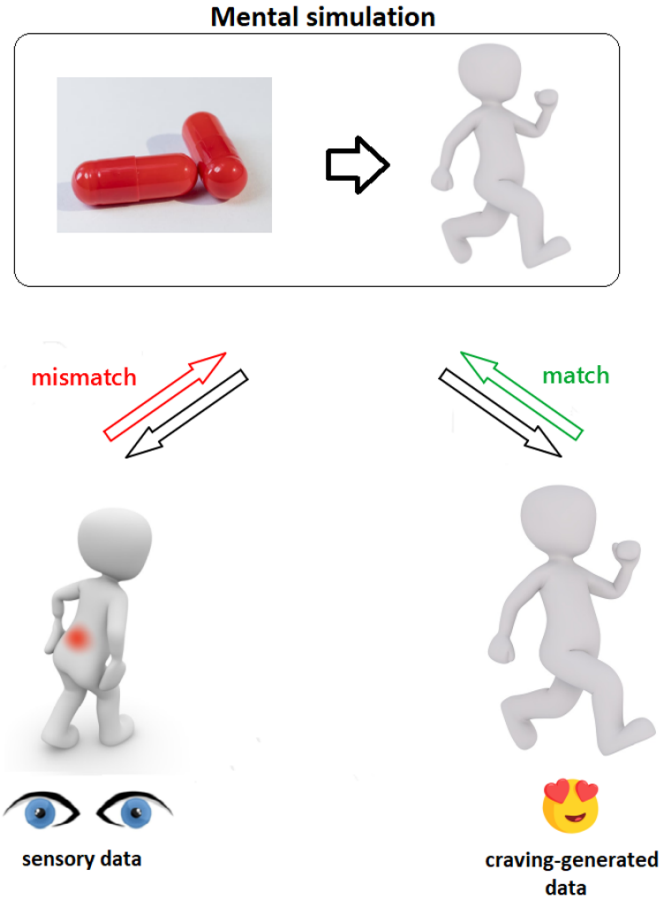
Now if you stop putting weight on the sore muscle, the pain goes away, fulfilling the prediction of "I am not in pain". As soon as your brain figures this out, your motor cortex can incorporate the craving-generated constraint of "I will not be in pain" into its planning. It generates different plans of how to move your body, and whenever it predicts that one of them would violate the constraint of "I will not be in pain", it will revise its plan. The end result is that you end up moving in ways that avoid putting weight on your sore muscle. If you miscalculate, the resulting pain will cause a rapid error signal that causes you to adjust your movement again.

What if the pain is more persistent, and bothers you no matter how much you try to avoid moving? Or if the circumstances force you to put weight on the sore muscle?

In that case, the brain will continue looking for a possible hypothesis that would fulfill the constraint of "I am not in pain". For example, maybe you have previously taken

painkillers that have helped with your pain. In that case, your mind may seize upon the hypothesis that “by taking painkillers, my pain will cease”.

As your mind predicts the likely consequences of taking painkillers, it notices that in this simulation, the constraint of “I am not in pain” gets fulfilled, driving down the error signal between the hypothesis and the “I am not in pain” constraint. However, if the brain could suppress the craving-for-pain-relief merely by *imagining* a scenario where the pain was gone, then it would never need to take any actions: it could just hallucinate pleasant states. Helping keep it anchored into reality is the fact that simply imagining the painkillers has not done anything to the pain signal itself: the imagined state does not match your actual sense data. There is still an error signal generated between the mismatch of the imagined “I have taken painkillers and am free of pain” scenario, and the fact that the pain is not gone yet.



Your brain imagines a possible experience: taking painkillers and being free of pain. This imagined scenario fulfills the constraint of “I have no pain”. However, it does not fulfill the constraint of actually matching your sense data: you have ***not*** *yet taken painkillers and* ***are*** *still in pain.*

Fortunately, if painkillers are actually available, your mind is not locked into a state where the two constraints of “I’m in pain” and “I’m not in pain” remain equally impossible to achieve. It can take actions - such as making you walk towards the medicine cabinet - that get you closer towards being able to fulfill both of these constraints.

There are studies suggesting that physical pain and psychological pain share similar neural mechanisms [*citation*]. And in meditation, one may notice that psychological discomfort and suffering involves avoiding unpleasant sensations in the same way as physical pain does; the same mechanism has been recruited for more abstract planning.

When the brain predicts that a particular experience would produce an unpleasant sensation, craving resists that prediction and tries to find another way. Similarly, if the brain predicts that something will *not* produce a pleasant sensation, craving may also resist *that* aspect of reality.

Now, this process as described has a structural equivalence to binocular rivalry, but as far as I know, binocular rivalry does not involve any particular discomfort. Suffering obviously does.

Being in pain is generally bad: it is usually better to try to avoid ending up in painful states, as well as try to get out of painful states once you are in them. This is also true for other states, such as hunger, that do not necessarily feel painful, but still have a negative emotional tone. Suppose that whenever craving generates a self-fulfilling prediction which resists your direct sensory experience, this generates a signal we might call “unsatisfactoriness”.

The stronger the conflict between the experience and the craving, the stronger the unsatisfactoriness - so that a mild pain that is easy to ignore only causes a little unsatisfactoriness, and an excruciating pain that generates a strong resistance causes immense suffering. The brain is then wired to use this unsatisfactoriness as a training signal, attempting to avoid situations that have previously included high levels of it, and to keep looking for ways out if it currently has a lot of it.

It is also worth noting what it means for you to be paralyzed by two strong, mutually opposing cravings. Consider again the situation where I am torn between admitting the truth to my friend, and staying quiet. We might think that this is a situation where the overall system is uncertain of the correct course of action: some subsystems are trying to force the action of confronting the situation, others are trying to force the action of avoiding it. Both courses of action are predicted to lead to some kind of loss.

In general, it is a bad thing if a system ends up in a situation where it has to choose between two different kinds of losses, and has high internal uncertainty of the right action. A system should avoid such dilemmas, either by avoiding the situations themselves or by finding a way to reconcile the conflicting priorities.

Craving-based and non-craving-based motivation

What I have written so far might be taken to suggest that craving is a requirement for all action and planning. However, the Buddhist claim is that craving is actually just one of at least two different motivational systems in the brain. Given that neuroscience suggests the existence of at least *three* different motivational systems, this should not seem particularly implausible.

Let’s take another look at the types of processes related to binocular rivalry versus craving.

Craving acts by actively introducing false beliefs into one’s reasoning. If craving could just do this completely uninhibited, rewriting *all* experience to match one’s desires,

nobody would ever do anything: they would just sit still, enjoying a craving-driven hallucination of a world where everything was perfect.

In contrast, in the case of binocular rivalry, no system is feeding the reasoning process any false beliefs: all the constraints emerge directly from the sense data and previous life-experience. To the extent that the system can be said to have a preference over either the "I am seeing a house" or the "I am seeing Isaac Newton" hypothesis, it is just "if seeing a house is the most likely hypothesis, then I prefer to see a house; if seeing Newton is the most likely hypothesis, then I prefer to see Newton". The computation does not have an intrinsic attachment to any particular outcome, nor will it hallucinate a particular experience if it has no good reason to.

Likewise, it seems like there are modes of doing and being which are similar in the respect that one is focused on process rather than outcome: taking whatever actions are best-suited for the situation at hand, regardless of what their outcome might be. In these situations, little unsatisfactoriness seems to be present.

In an earlier post, I discussed a proposal where an autonomously acting robot has two decision-making systems. The first system just figures out whatever actions would maximize its rewards and tries to take those actions. The second "Blocker" system tries to predict whether or not a human overseer would approve of any given action, and prevents the first system from doing anything that would be disapproved of. We then have two evaluation systems: "what would bring the maximum reward" (running on a lower priority) and "would a human overseer approve of a proposed action" (taking precedence in case of a disagreement).

It seems to me that there is something similar going on with craving. There are processes which are neutrally just trying to figure out the best action; and when those processes hit upon particularly good or bad outcomes, craving is formed in an attempt to force the system into repeating or avoiding those outcomes in the future.

Suppose that you are in a situation where the best possible course of action only has a 10% chance of getting you through alive. If you are in a non-craving-driven state, you may focus on getting at least that 10% chance together, since that's the best that you can do.

In contrast, the kind of behavior that is typical for craving is realizing that you have a significant chance of dying, deciding that this thought is completely unacceptable, and refusing to go on before you have an approach where the thought of death isn't so stark.

Both systems have their upsides and downsides. If it is true that a 10% chance of survival really is the best that you can do, then you should clearly just focus on getting the probability even that high. The craving which causes trouble by thrashing around is only going to make things worse. On the other hand, maybe this estimate is flawed and you could achieve a higher probability of survival by doing something else. In that case, the craving absolutely refusing to go on until you have figured out something better might be the right action.

There is also another major difference, in that craving does not *really* care about outcomes. Rather, it cares about avoiding positive or negative feelings. In the case of avoiding death, craving-oriented systems are primarily reacting to the *thought* of death… which may make them reject even plans which would *reduce* the risk of death, if those plans involved needing to think about death too much.

This becomes particularly obvious in the case of things like going to the dentist in order to have an operation you know will be unpleasant. You may find yourself highly averse to going, as you crave the comfort of not needing to suffer from the unpleasantness. At the same time, you *also* know that the operation will benefit you in the long term: any unpleasantness will just be a passing state of mind, rather than permanent damage. But avoiding unpleasantness - including the very thought of experiencing something unpleasant - is just what craving is about.

In contrast, if you are in a state of equanimity with little craving, you still recognize the thoughts of going to the dentist as having negative valence, but this negative valence does not bother you, because you do not have a craving to avoid it. You can choose whatever option seems best, regardless of what kind of content this ends up producing in your consciousness.

Of course, choosing correctly requires you to actually *know* what is best. Expert meditators have been known to sometimes ignore extreme physical pain that should have caused them to seek medical aid. And they probably *would* have sought help, if not for their ability to drop their resistance to pain and experience it with extreme equanimity.

Negative-valence states tend to correlate with states which are bad for the achievement of our goals. That is the reason why we are wired to avoid them. But the correlation is only partial, so if you focus too much on avoiding unpleasantness, you are falling victim to Goodhart's Law: optimizing a measure so much that you sacrifice the goals that the measure was supposed to track. Equanimity gives you the ability to ignore your consciously experienced suffering, so you don't need to pay additional mental costs for taking actions which further your goals. This can be useful, if you are strategic about actually achieving your goals.

But while Goodharting on a measure is a failure mode, so is ignoring the measure entirely. Unpleasantness *does* still correlate with things that make it harder to realize your values, and the need to avoid displeasure normally operates as an automatic feedback mechanism. It is possible to have high equanimity and weaken this mechanism, without being smart about it and doing nothing to develop alternative mechanisms. In that case you are just trading Goodhart's Law for the opposite failure mode.

Some other disadvantages of craving

In the beginning of this post, I mentioned a few other disadvantages that craving has, which I have not yet mentioned explicitly. Let's take a quick look at those.

Craving narrows your perception, making you only pay attention to things that seem immediately relevant for your craving.

In predictive processing, attention is conceptualized as giving increased weighting to those features of the sensory data that seem most useful for making successful predictions about the task at hand. If you have strong craving to achieve a particular outcome, your mind will focus on those aspects of the sensory data that seem useful for realizing your craving.

Strong craving may cause premature exploitation. If you have a strong craving to achieve a particular goal, you may not want to do anything that looks like moving away from it, even if that would actually help you achieve it better.

Suppose that you have a strong craving to experience a feeling of accomplishment: this means that the craving is strongly projecting a constraint of “I will feel accomplished” into your planning, causing an error signal if you consider any plan which does not fulfill the constraint. If you are thinking about a multistep plan which will take time before you feel accomplished, it will start out by you *not* feeling accomplished. This contradicts the constraint of “I will feel accomplished”, causing that plan to be rejected in favor of ones that bring you even *some* accomplishment right away.

Craving and suffering

We might summarize the unsatisfactoriness-related parts of the above as follows:

- Craving tries to get us into pleasant states of consciousness.
- But pleasant states of consciousness are those *without* craving.
- Thus, there are subsystems which are trying to get us into pleasant states of consciousness by creating constant craving, which is the exact *opposite* of a pleasant state.

We can somewhat rephrase this as:

- The default state of human psychology involves a degree of almost constant dissatisfaction with one’s state of consciousness.
- This dissatisfaction is created by the craving.
- The dissatisfaction can be ended by eliminating craving.

… which, if correct, might be interpreted to roughly equal the first three of Buddhism’s Four Noble Truths: the fourth is “Buddhism’s Noble Eightfold Path is a way to end craving”.

A more rationalist framing might be that the craving is essentially acting in a way that looks similar to wireheading: pursuing pleasure and happiness even if that sacrifices your ability to impact the world. Reducing the influence of the craving makes your motivations less driven by wireheading-like impulses, and more able to see the world clearly even if it is painful. Thus, reducing craving may be valuable even if one does not care about suffering less.

This gives rise to the question - how exactly *does* one reduce craving? And what does all of this have to do with the self, again?

We’ll get back to those questions in the next post.

This is the third post of the "a non-mystical explanation of insight meditation and the three characteristics of existence" series. The next post in the series is "From self to craving".

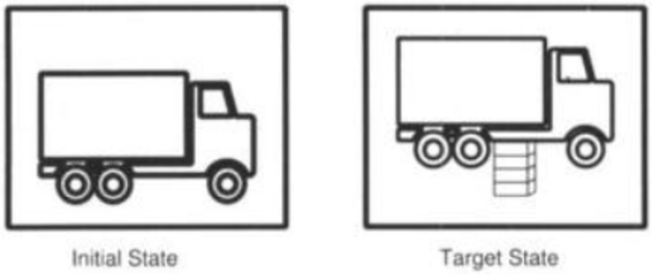
From self to craving (three characteristics series)

Buddhists talk a lot about the self, and also about suffering. They claim that if you come to investigate what the self is really made of, then this will lead to a reduction in suffering. Why would that be?

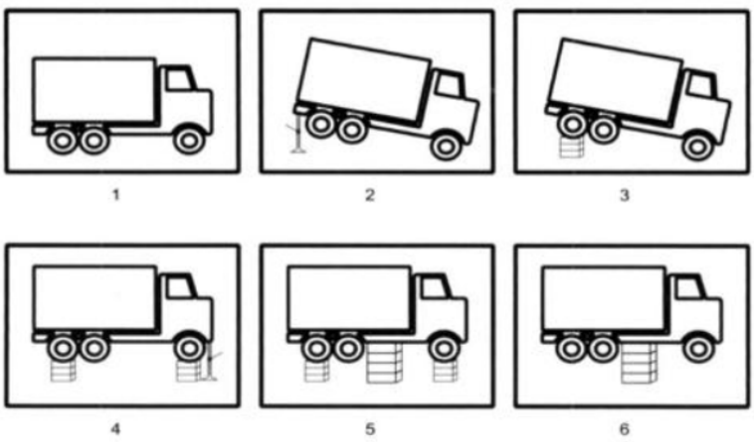
This post seeks to answer that question. First, let's recap a few things that we have been talking about before.

The connection between self and craving

In "a non-mystical explanation of 'no-self'", I talked about the way in which there are two kinds of goals. First, we can manipulate something that does not require a representation of ourselves. For example, we can figure out how to get a truck on a column of blocks.

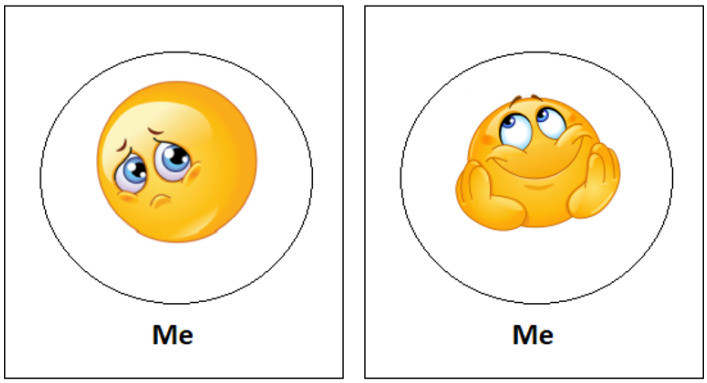


In that case, we can figure out a sequence of actions that takes the truck from its initial state to its target state. We don't necessarily need to think about *ourselves* as we are figuring this out - the actual sequence could just as well be carried out by someone else.



I mentioned that these kinds of tasks seem to allow flow states, in which the sense of self becomes temporarily suspended as unnecessary, and which are typically experienced as highly enjoyable and free from discomfort.

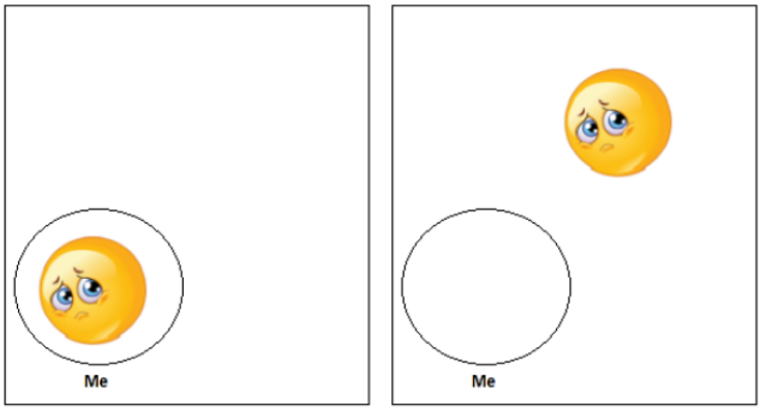
Alternatively, we can think of a goal which intrinsically requires self-reference. For example, I might be feeling sad, and think that I want to feel happy instead. In this case, both the initial state and the target state are defined in terms of what I feel, so in order to measure my progress, I need to track a reference to myself.



Initial state Target state

In that post, I remarked that changing one's experience of one's self may change how emotions are experienced. This does not necessarily require high levels of enlightenment: it is a common mindfulness practice to reframe your emotions as something that is external to you, in which case negative emotions might cease to feel aversive.

I have also previously discussed therapy techniques that allow you to create some distance between yourself and your feelings, making them less aversive. For example, one may pay attention to where in their body they feel an emotion, keep their attention on those physical feelings, and then allow a visual image of that emotion to arise. This may then cause the emotion to be experienced as "not me", in a way which makes it easier to deal with.



Aversive state Non-aversive state

The next post in my series was on craving and suffering, where I distinguished two different kinds of motivation: craving and non-craving. I claimed that craving is

intrinsically associated with the desire to either experience positive valence, or to avoid experiencing negative valence. Non-craving-based motivation, on the other hand, can care about anything; not just valence.

In particular, I claimed that discomfort / suffering (unsatisfactoriness, to use a generic term) is created by craving: craving attempts to resist reality, and in so doing generates an error signal which is subjectively experienced as unsatisfactoriness. I also suggested that non-craving-based motivation does not resist reality in the same way, so does not create unsatisfactoriness.

Putting some of these pieces together:

- Craving tries to ensure that “the self” experiences positive feelings and avoids negative feelings.
- If there are consciously experienced feelings which are not interpreted as being experienced by the self, it does not trigger craving.
- There is a two-way connection between the sense of self and craving.
    - On one hand, experiencing a strong sense of self triggers craving, as feelings are interpreted as happening to the self.
    - From the other direction, once craving *is* triggered, it sends into consciousness the goal of either avoiding or experiencing particular feelings. As this goal is one that requires making reference to the self, sending it into consciousness instantiates a sense of self.

Let’s look at this a bit more.

Craving as a second layer of motivation

A basic model is that the brain has subsystems which optimize for different kinds of goals, and then produce positive or negative valence in proportion to how well those goals are being achieved.

For example, appraisal theories of emotion hold that emotional responses (with their underlying positive or negative valence) are the result of subconscious evaluations about the significance of a situation, relative to the person's goals. An evaluation saying that you have lost something important to you, for example, may trigger the emotion of sadness with its associated negative valence.

Or consider a situation where you are successfully carrying out some physical activity; playing a fast-paced sport or video game, for example. This is likely to be associated with positive valence, which emerges from having success at the task. On the other hand, if you were failing to keep up and couldn't get into a good flow, you would likely experience negative valence.

Valence looks like a signal about whether some goals/values are being successfully attained. A subsystem may have a goal X which it pursues independently, and depending on how well it goes, valence is produced as a result. In this model, because valence tends to signal states that are good/bad for the achievement of an organism's goals, craving acts as an additional layer that "grabs onto" states that seem to be particularly good/bad, and tries to direct the organism more strongly towards those.

There's a subtlety here, in that craving is distinct from negative valence, but craving can certainly *produce* negative valence. For example:

- You are feeling stressed out, so you go for a long walk. This helps take your mind off the stress, and you come back relaxed.
- The next time you are stressed and feel like you need a break, you remember that the walk helped you before. It's important for you to get some more work done today, so you take another walk in the hopes of calming down again. As you do, you keep thinking "okay, am I about to calm down and forget my worries now" - and this question keeps occupying you throughout your walk, so that when you get back home, you haven't actually relaxed at all.

I think this has to do with craving influencing the goals and inputs that are fed into various subsystems. If craving generates the goal of "I should relax", then that goal is taken up by some planning subsystem; and success or failure at that goal, may by itself produce positive or negative valence just like any other goal does. This also means that craving may generate emotions that generate additional craving: first you have a craving to relax on a walk, but then going on the walk produces frustration, creating craving to be rid of the frustration...

My model about craving and non-craving seems somewhat similar to the proposed distinction between model-based and model-free goal systems in the brain. The model-based system does complex reasoning about what to do; it is capable of pretty sophisticated analyses, but requires substantial computational resources. To save on computation, the model-free system remembers which actions have led to good or bad outcomes in the past, and tries to repeat/avoid them. Under this model, craving would be associated with something like the model-free system, one which used "did an action produce positive or negative valence" as a shorthand for whether actions should be taken or avoided.

However, it would be a mistake to view these as two entirely distinct systems. Research suggests that even within the same task, both kinds of systems are active, with the brain adaptively learning which one of the systems to deploy during which parts of the activity. Craving valence and more complex model-based reasoning also seem to be intertwined in various ways, such as:

- It is possible to unlearn particular cravings, as the brain updates to notice that this specific craving is not actually useful. The unlearning process seems to involve some degree of model-based evaluation, as the brain seems resistant to release cravings if it predicts that doing so would make it harder to achieve some particular goal.
- Model-based subsystems may act in ways that seem to make use of craving. For example, in an earlier post reviewing the book Unlocking the Emotional Brain, I discussed the example of Richard. A subsystem in his brain predicted that if he were to express confidence, people would dislike him, so it projected negative self-talk into his consciousness to prevent him from being confident. Presumably the self-talk had negative valence and caused a craving to avoid it, in a way which contributed to him not saying anything. I suspect that this is part of why mindfulness practice may release buried trauma: if you get better at dealing with mild aversion, that aversion might previously have been used by subsystems to keep negative memories contained - and then your mind gets flooded with really unpleasant memories, and much stronger unsatisfactoriness than you were necessarily prepared to deal with.

It is also worth making explicit that pursuing positive valence and avoiding negative valence are goals that can be pursued by non-craving-based subsystems as well.

For basically any goal, there can be craving-based and non-craving-based motivations. You might pursue pleasure because of a craving for pleasure, but you may also pursue it because you value it for its own sake, because experiencing pleasure makes your mind and body work better than if you were only experiencing unhappiness, because it is useful for releasing craving… or for any other reason.

From self to craving and from craving to self

I have mostly been talking about craving in terms of aversion (craving to avoid a negative experience). Let's look at some examples of craving a positive experience instead:

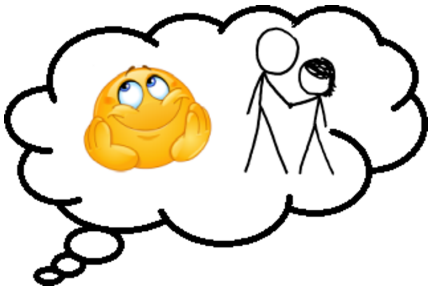
- It is morning and your alarm bell rings. You should get up, but it feels nice to be sleepy and remain in bed. You want to hang onto those pleasant sensations of sleepiness for a little bit more.
- You are spending an evening together with a loved one. This is the last occasion that you will see each other in a long time. You feel really good being with them, but a small part of you is unhappy over the fact that this evening will eventually end.
- You are at work on a Friday afternoon. Your mind wanders to the thought of no longer being at work, and doing the things you had planned for the weekend. You would prefer to be done with work already, and find it hard to stay focused as you cling to the thoughts of your free time.
- You are single and hanging out with an attractive person. You know that they are not into you, but it would be so great if they were. You can't stop thinking about that possibility, and this keeps distracting you from the actual conversation.
- You are in a conversation with several other people. You think of a line that would be a great response to what someone else just said. Before you can say it, somebody says something, and the conversation moves on. You find yourself still thinking of your line, and how nice it would have been to get to say it.
- You had been planning on going to a famous museum while on your vacation, but the museum turns out to be temporarily closed at the time. You keep thinking about how much you had been looking forward to it.
- You are hungry, and keep thinking about how good a particular food would taste, and how much better you would feel after you had eaten.

These circumstances are all quite different, but on a certain abstract level, they share the same kind of a mechanism. There is the thought of something pleasant, which triggers craving, and a desire to either get into that pleasant state or make sure you remain in it.

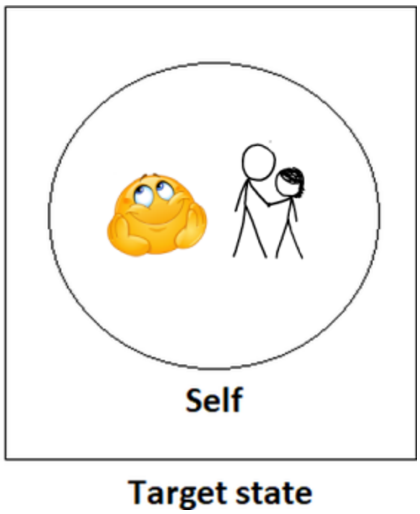
Let's say that there is this kind of a process:

Sense contact. You have some particular experience, such as the sensation of being sleepy and in bed, or the thought of how it would feel if the attractive person in front of you were to like you. In third-person terms, this sensation/thought is sent to the global workspace of your consciousness.

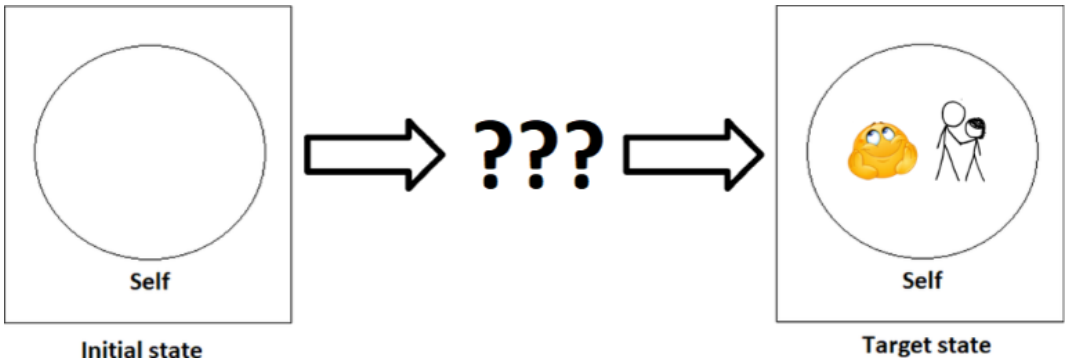
Valence. An emotional system classifies this experience as being positive, and paints it with a corresponding positive gloss as it is being broadcast into the workspace.



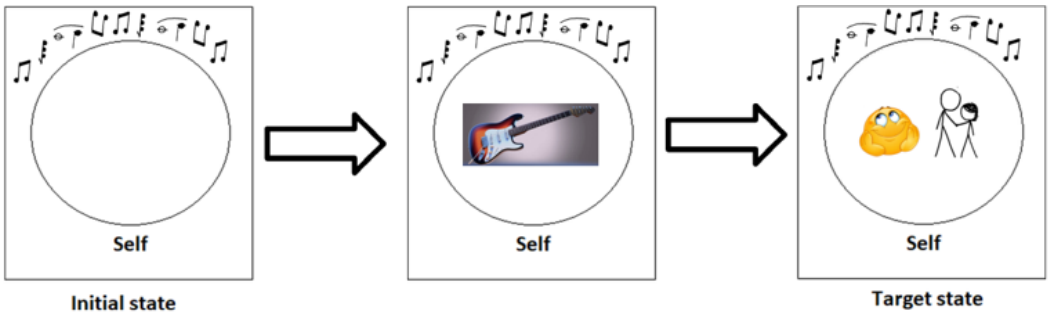
Craving. A craving subsystem notices the positive valence, and interprets the valence as *being experienced by you.* The subsystem registers the imagined state the valence is associated with, so it sets the intention of getting the self to achieve that state, and the resulting positive valence. For example, it may set the intention of getting into a relationship with that attractive person.



Clinging and planning. This intent clings in your mind and is fed to planning subsystems. They make plans of how to get to the state which has been evaluated as being better. (My previous post was largely a description of this stage.)



Birth. As these plans are broadcast into consciousness for evaluation, they contain a representation of your current state, creating a stronger experience of having a distinct self that wants a particular thing. Subsystems may emphasize particular aspects of their model of you that seem relevant for achieving the goal: for example, if the other person seems to be drawn towards musicians, your skill at this may become highlighted. Maybe you, being a musical kind of person, could play the guitar and make a good impression on them... emphasizing your "musician nature" in the self-representation that appears in consciousness.



Death. Eventually, you stop pursuing the goal, at least for the time being. Maybe you get what you wanted, it turns out to be unviable right now, or you just get distracted and forget. This particular goal-state and plan disappear from consciousness, and with it, the sense of self that was tracking its completion disappears as well. Before long - or maybe even simultaneously - another self will be created, one which cares about something completely different: maybe you got hungry, and now the craving only cares about getting some food, creating a food-hungry self...

Craving and the self-model

There's a straightforward reason for why craving should be tied into a conception of the self: its purpose is to motivate *you* to action. If your brain predicts that someone other than you would experience positive valence, this does not trigger craving in the same way. (Imagine getting the one thing that you most desire at the moment. Then imagine some complete stranger getting a similar thing. Not quite the same, is it?)

Your craving subsystems have been wired to make *you* experience positive valence / avoid negative valence. Each craving subsystem contains an implicit schema along the lines of "*I* will strive towards a more positive experience, and then *I* will have that more positive experience in the end". If a craving subsystem predicted that its actions would produce someone *else* a more positive experience, this would not fulfill the goal

condition in that schema. (Of course, craving may have the instrumental goal of making someone else happy, if it predicts that leads to a positive consequence for you.)

At the same time, there is something interesting going on with the fact that mindfulness and cognitive defusion practices work: if you can mentally transform a source of negative valence into something that feels external to you, it may not bother you as much. In particular, it feels like you don't need to react to it: that is, there is less of a craving to get it out of your consciousness. The inference of "if I get this emotion out of my mind, I will feel better" is never applied, as the emotion is not experienced as being "in my mind" in the first place.

In other words, the mere experience or prediction of positive/negative valence alone does not seem to be enough to trigger craving. The valence also needs to be bound into a particular abstract representation of the self, and interpreted as happening to "you".

At this point, we need to distinguish between two different senses of "happening to you":

- Happening to the system defined by the physical boundaries of your body; for conscious experiences, appearing in the global workspace located within that body. I will call this "happening to the system".
- Happening to "the self", an abstract tag which is computed within the system, and which feels like the entity that internal experiences such as emotions happen *to*; typically only includes *part* of the content in the global workspace. I will call this "happening to the self-model".

Craving reacts to valence that is experienced by the self-model. It may *seem* to also react to events that happen to the system, such as getting to sleep for longer, but this always takes place through something that bottoms out at the self-model experiencing valence. E.g. the thought of sleeping longer creates positive valence, and the craving reacts to the positive valence becoming incorporated in the self-model. (At the same time, the predictions that craving is based on are not necessarily accurate or up to date, so there is also craving for things that we do not actually enjoy.)

At the same time, non-craving-based motivation may react to anything, including events that happen to the system rather than the self-model. Even if the thought of getting to sleep for longer didn't cause craving, you could still stay in bed in order to get more rest. Because this involves some degree of abstract reasoning, craving is probably something that humans need to have in place *before* non-craving-based systems have had a chance to mature and acquire sophisticated models. A toddler is not capable of intellectually figuring out the consequences of most harmful things and how they should thus be avoided, but the toddler *does* still have visceral craving to avoid pain.

Earlier in this series, I mentioned a metaphor of experiencing yourself as the sky, and all the emotions and thoughts as things that are on the sky but which do not affect the sky. This was intended as a metaphor for how it feels once your mind comes to identify with your underlying field of consciousness, as opposed to the contents of that consciousness. I noted that this raised a question which I promised to come back to: why would this kind of a shift in identification reduce suffering?

What I have outlined above suggests an answer: because craving subsystems are activated by a prediction of positive or negative valence being experienced by the self-model. If the system's self-model changes so that experiences are no longer interpreted as happening to the self-model, craving will not trigger, nor will it produce a feeling of unsatisfactoriness. At the same time, non-craving-based motivation can still reason about the consequences of those events and respond appropriately.

In these articles, I have used the term "no-self", because that seems to be the established translation for *anatta*; but I could also have used the arguably better translation of"*not* self".

In other words, the brain has something of a built-in model defining what kinds of criteria a piece of conscious experience needs to fulfill in order to be incorporated into the self-model. With sufficiently close study, one may come to notice the way in which actually *no* conscious content fulfills this criteria: there is nothing in consciousness that is "self"... and if there is nothing that is incorporated into the self-model, then there is nothing that could trigger craving.

At the same time, some "no-self" experiences also seem to also be interpreted as "everything is self" rather than "nothing is self". For example, my earlier post on no-self quoted an experience described as "*this is all me [...] my identity is literally everything that I could see through my eyes*". Craving also seems reduced in these situations.

With emotions, the raw experience of the emotion is distinct from the process of naming the emotion; the same emotion may have different names in different languages. Likewise, the subsystem which creates the abstract boundary that craving responds to, and the subsystem which produces the verbal label of that boundary, may be distinct. Thus, if the experience of a boundary dividing self and non-self disappears, then this might on a verbal level be interpreted as either "all is self" *or* "nothing is self", depending on which framework one is using and which features of the experience one is paying attention to.

This is the fourth post of the "a non-mystical explanation of insight meditation and the three characteristics of existence" series. The next post in the series is "On the construction of the self".

Some of the stick figures and musical notes in this post were borrowed from xkcd.com.

On the construction of the self

This is the fifth post of the "a non-mystical explanation of the three characteristics of existence" series.

On the construction of the self

In his essay The Self as a Center of Narrative Gravity, Daniel Dennett offers the thought experiment of a robot that moves around the world. The robot also happens to have a module writing a novel about someone named Gilbert. When we look at the story the novel-writing module is writing, we notice that its events bear a striking similarity to what the rest of the robot is doing:

> If you hit the robot with a baseball bat, very shortly thereafter the story of Gilbert includes his being hit with a baseball bat by somebody who looks like you. Every now and then the robot gets locked in the closet and then says "Help me!" Help whom? Well, help Gilbert, presumably. But who is Gilbert? Is Gilbert the robot, or merely the fictional self created by the robot? If we go and help the robot out of the closet, it sends us a note: "Thank you. Love, Gilbert." At this point we will be unable to ignore the fact that the fictional career of the fictional Gilbert bears an interesting resemblance to the "career" of this mere robot moving through the world. We can still maintain that the robot's brain, the robot's computer, really knows nothing about the world; it's not a self. It's just a clanky computer. It doesn't know what it's doing. It doesn't even know that it's creating a fictional character. (The same is just as true of your brain; it doesn't know what it's doing either.) Nevertheless, the patterns in the behavior that is being controlled by the computer are interpretable, by us, as accreting biography--telling the narrative of a self.

As Dennett suggests, something similar seems to be going on in the brain. Whenever you are awake, there is a constant distributed decision-making process going on, where different subsystems swap in and out of control. While you are eating breakfast, subsystem #42 might be running things, and while you are having an argument with your spouse, subsystem #1138 may be driving your behavior. As this takes place, a special subsystem that we might call the “self-narrative subsystem” creates a story of how all of these actions were taken by “the self”, and writes that story into consciousness.

The consciously available information, including the story of the self, is then used by subsystems in the brain to develop additional models of the system’s behavior. There is an experience of boredom, and “the self” seems to draw away from the boredom - this may lead to the inference that boredom is bad for the self, and something that needs to be avoided. An alternative inference might have been that there is a subsystem which reacts to boredom by avoiding it, while other subsystems don’t care. So rather than boredom being intrinsically bad, one particular subsystem has an avoidance reaction to it. But as long as the self-narrative subsystem creates the concept of a self, the notion of there being a single self doing everything looks like the simplest explanation, and the system tends to gravitate towards that interpretation.

In the article “The Apologist and the Revolutionary”, Scott Alexander discusses neuroscientist V.S. Ramachandran’s theory that the brain contains a reasoning module

which attempts to fit various observations into the brain's "current best guess" of what's going on. He connects this theory with the behavior of "split-brain" patients, who have had the connection between the hemispheres of their brain severed:

> Consider the following experiment: a split-brain patient was shown two images, one in each visual field. The left hemisphere received the image of a chicken claw, and the right hemisphere received the image of a snowed-in house. The patient was asked verbally to describe what he saw, activating the left (more verbal) hemisphere. The patient said he saw a chicken claw, as expected. Then the patient was asked to point with his left hand (controlled by the right hemisphere) to a picture related to the scene. Among the pictures available were a shovel and a chicken. He pointed to the shovel. So far, no crazier than what we've come to expect from neuroscience.
>
> Now the doctor verbally asked the patient to describe why he just pointed to the shovel. The patient verbally (left hemisphere!) answered that he saw a chicken claw, and of course shovels are necessary to clean out chicken sheds, so he pointed to the shovel to indicate chickens. The apologist in the left-brain is helpless to do anything besides explain why the data fits its own theory, and its own theory is that whatever happened had something to do with chickens, dammit!

Now, as people have pointed out, it is tricky to use split-brain cases to draw conclusions about people who have not had their hemispheric connections severed. But this case looks *very* similar to what meditative insights suggest, namely that the brain constructs an ongoing story of there being a single decision-maker. And while the story of the single self tends to be closely correlated with the system's actions, the narrative self does not actually *decide* the person's actions, it's just a story of someone who does. In a sense, the part of your mind that may feel like the "you" that takes actions, is actually produced by a module that just *claims credit* for those actions.

With sufficiently close study of your own experience, you may notice how these come apart, and what creates the appearance of them being the same. As you accumulate more evidence about their differences, it becomes harder for the system to consistently maintain the hypothesis equating them. Gradually, it will start revising its assumptions.

Some of this might be noticed without any meditation experience. For example, suppose that you get into a nasty argument with your spouse, and say things that you don't fully mean. The next morning, you wake up and feel regret, wondering why in the world you said those terrible things. You resolve to apologize and not do it again, but after a while, there is another argument and you end up behaving in the same nasty way.

What happened was something like: something about the argument triggered a subsystem which focused your attention on everything that could be said to be wrong with your spouse. The following morning, that particular trigger was no longer around: instead, there was an unpleasant emotion, guilt. The feelings of guilt activated another subsystem which had the goal of making *those* sensations go away, and it had learned that self-punishment and deciding to never act the same way is an effective way of doing so. (As an aside, Chris Frith and Thomas Metzinger speculate that one of the reasons why we have a unified self-model, is to allow this kind of a social regret and a feeling of "I could have done otherwise".)

In fact, the “regretful subsystem” cannot directly influence the “nasty subsystem”. The regretful subsystem can only output the thought that *it* will not act in a nasty way again - which it never did in the first place. Yet, because the mind-system has the underlying modeling assumption of being unified, the regretful subsystem will (correctly) predict that *it* will not act in that nasty way, if put into the same situation again… while failing to predict that it will actually be the *nasty* subsystem that activates, because the regretful subsystem is currently *identifying with* (modeling itself as having caused) the actions that were actually caused by the nasty subsystem.

Having an intellectual understanding of this may sometimes help, but not always. Suppose that you keep feeling strong shame over what happened and a resolve never to do so again. At the same time, you realize that you are not thinking clearly about this, and that you are not actually a unified self. Your train of thought might thus go something like this:

“Oh God why did I do that, I feel so ashamed for being such a horrible person… but aaah, it’s pointless to feel shame, the subsystem that did that wasn’t actually the one that is running now, so this isn’t helping change my behavior… oh man oh man oh man I’m such a terrible person I must never do that again… but that thought *doesn’t actually help*, it doesn’t influence the system that actually did it, I must do something else… aaaahh why did I do that I feel so much shame…”

Which we can rewrite as:

Regretful subsystem: “Oh God why did I do that, I feel so shamed for being such a horrible person…”

“Sophisticated” subsystem: “But aaah, it’s pointless to feel shame, the subsystem that did that wasn’t actually the one that is running now…”

Regretful subsystem: “Oh man oh man oh man I’m such a terrible person I must never do that again…”

“Sophisticated” subsystem: “But that thought *doesn’t actually help*, it doesn’t influence the subsystem that actually did it, so I must do something else…”

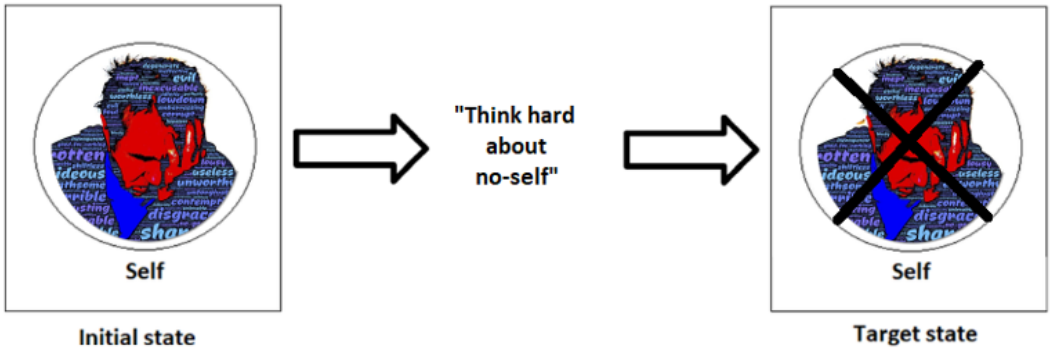
Regretful subsystem: “Aaaahh why did I do that, I feel so much shame…”

I previously said that the regretful subsystem cannot directly affect the nasty subsystem… but it may very well be that a more sophisticated subsystem, which understands what is happening, cannot directly influence the regretful subsystem *either*. Even as it understands the disunity of self on one level, it may still end up identifying with all the thoughts currently in consciousness.

In fact, the “sophisticated” subsystem may be particularly *likely* to assume that the mind is unified.

Suppose that the regretful subsystem triggers first, and starts projecting self-punishment into the global workspace. This then triggers another subsystem, which has learned that “a subsystem triggering to punish myself is not actually going to change my future actions, and it would be more useful for me to do something else”. In other words, even as the philosophical subsystem’s *strategy* is to *think really hard about the verbal concept of no-self*, the subsystem *has actually been* triggered by the thought of “I am ashamed”, and has the goal state of making the “me” not feel shame. It even ends up thinking “I must do something else”, as if the shameful

thoughts were coming from *it* rather than a different subsystem. Despite a pretense of understanding no-self, its actual cognitive schema is still identifying with the regretful system's output.



This is one of the trickiest things about no-self experiences: on some occasions, you might be able to get into a state with no strong sense of self, when you are not particularly trying to do so and don't have strong craving. The resulting state then feels really pleasant. You might think that this would make it easier to let go of craving in the future… and it might in fact make you more motivated to let go of craving again, but "more motivated" frequently means "have more craving to let go of craving". Then you need to find ways to disable *those* subsystems - but of course, if you have the *goal* of disabling the subsystems, then this may mean that you have craving to let go of the craving to let go of the craving…

Mechanisms of claiming credit

Consider someone who is trying hard to concentrate on a lecture (or on their meditation object), but then wanders off into a daydream without noticing it. Thomas Metzinger calls the moment of drifting off to the daydream a "self-representational blink". As the content of consciousness changes from that produced by one subsystem to another, there is a momentary loss of self-monitoring that allows this change to happen unnoticed.

From what I can tell, the mechanism by which the self-narrative agent manages to disguise itself as the causally acting agent seems to also involve a kind of a self-representative blink. Normally, *intentions* precede both physical and mental actions: before moving my hand I have an intention to move my hand, before pursuing a particular line of thought I have an intention to pursue that line of thought, and before moving my attention to my breath I have an intention to move my attention to my breath. Just as these intentions appear, there is a blink in awareness, making it hard to perceive what exactly happens. That blink is used by the self-narrative agent to attribute the intention as arising from a single "self".

If one develops sufficient introspective awareness, they may come to see that the intentions are arising on their own, and that there is actually no way to control them: if one *intends* to control them somehow, then the intention to do that is also arising on its own. Further, it can become possible to see exactly what the self-narrative agent is doing, and how it creates the story of an independent and acausal self. This does not eliminate the self-narrative agent, but does allow its nature to be properly understood, as it is more clearly seen. Meditation teacher Daniel Ingram describes this as the

nature of the self-representation “becoming bright, clear, hardwired to be perceived in a way that doesn’t create habitual misperception”; and elaborates:

> The qualities that previously meant and were perceived to be “Agent” still arise in their entirety. Every single one of them. Their perceptual implications are very different, but, functionally, their meanings are often the same, or at least similar. […]
>
> All thoughts arise in this space. They arise simultaneously as experiences (auditory, physical, visual, etc.), but also as meaning. These meanings include thoughts and representations such as “I”, “me”, and “mine” just the way they did before. No meaning have been cut off. Instead, these meanings are a lot easier to appreciate as they arise for being the representations that they are, and also there is a perspective on this representation that is refreshingly open and spacious rather than being contracted into them as happened before.

No-self and paradoxes around “doing”

Some meditation instructions may say nonsensical-sounding things that go roughly as follows: “Just sit down and don’t do anything. But also don’t try to not do anything. You should neither do anything, nor do ‘not doing anything’.”

Or they might say: “Surrender entirely. This moment is as it is, with or without your participation. This does not mean that you must be passive. Surrender also to activity.”

This sounds confusing, because our normal language is built around the assumption of a unitary self which only engages in voluntary, intentionally chosen activities. But it turns out that if you *do* just sit down and decide to let whatever thoughts arise in your head, you might soon start having confusing experiences.

Viewed from a multi-agent frame, there is no such thing as “not making decisions”. Different subsystems are constantly transmitting different kinds of intentions - about what to say, what to do, and even what to think - into consciousness. Even things which do not subjectively *feel* like decisions, like just sitting still and waiting, are still the outcomes of a constant decision-making process. It is just that a narrative subsystem tags *some* outcomes as being the result of “you” having made a decision, while tagging others as things that “just happened”.

For example, if an unconscious evidence-weighting system decides to bring a daydream into consciousness, causing your mind to wander into the daydream, you may feel that you “just got lost in thought”. But when a thought about making tea comes into consciousness and you then get up to make tea, you may feel like you “decided to make some tea”. Ultimately, you have a schema which classifies everything as either “doing” or “not doing”, and attempts to place all experience into one of those two categories.

Various paradoxical-sounding meditation instructions about doing ”nothing” sound confusing because they are intended to get you to notice things that do not actually match these kinds of assumptions.

For example, you may be told to just sit down and let your mind do whatever it wants, neither controlling what happens in your mind but also not trying to avoid controlling it. Another way of putting this is:

- Do not exert any conscious intention to control your mind, such as by preventing yourself from thinking particular thoughts.
- But if you notice a sensation of yourself trying to control your mind, also do not do anything to stop yourself from feeling that sensation.

One thing (of many) that may happen when you try to follow these instructions is:

- You were told not to have any conscious intention to control your mind, so the subsystems that caused you to sit down and start meditating, take no particular action to shift the contents of consciousness.
- At some point, other subsystems, driven by a priority other than meditating, send something into consciousness. They are attempting to change its contents towards the particular priority that these subsystems care about.
- This is noticed by the subsystem classifying mental contents as doing or non-doing. It classifies this as “doing”, and sends a sensation of intentional “doing” into consciousness.
- The subsystem which originally sat down with the intention of meditating and not exerting conscious control, notices the sensation of conscious doing.
- The overall mind-system gets confused, because its internal narrative (modeled from the assumption of a unitary self) now says “I sat down and intended not to intentionally control my mind, but now I find myself intentionally controlling my mind anyway, but how can I unintentionally end up doing something intentional?”
- Following the original instructions, the non-meditating subsystem may now just let the sensation of conscious doing be there, without interfering with it.
- Eventually, there may be an experiential realization that the sensation of intentional doing is just a sensation or a tag assigned to some particular actions, not intrinsically associated with any single subsystem.

So, translated into multiagent language: “don’t do anything but also don’t not do anything” means “you, the subsystem which is following these instructions: don’t do anything in particular, but when an intention to do something arises from another subsystem, also don’t do anything to counteract that intention”.

Inevitably, you will experience yourself as doing something anyway, because another subsystem has swapped in and out of control, and this registers to you as 'you' having done something. This is an opportunity to notice on an experiential level that you are actually identifying with multiple distinct processes at the same time.

(This is very much a simplified account; in actuality, even the “meditating subsystem” seems itself composed of multiple smaller subsystems. As the degree of mental precision grows, every component in the mind will grow increasingly fine-grained. This makes it difficult to translate things into a coarser ontology which does not draw equal distinctions.)

The difficulty of surrendering

Contemplative traditions may make frequent reference to the concept of “surrendering” to an experience. One way that surrender may happen is when a

subsystem has repeatedly tried to resist or change a particular experience, failed time after time, and finally just turns off. (It is unclear to me whether this is better described as "the subsystem gives up and turns off" or "the overall system gives up on this subsystem and turns it off", but I think that it is probably the latter, even if the self-narrative agent may sometimes make it feel like the former.) Once one genuinely surrenders to the experience of (e.g.) pain with no attempt to make it stop, it ceases to be aversive.

However, this tends to be hard to repeat. Suppose that a person manages to surrender to the pain, which makes it stop. The next time they are in pain, they might remember that "I just need to surrender to the pain, and it will stop".

In this case, a subsystem which is trying to get the pain out of consciousness has formed the prediction that there *is* an action that can be carried out to make the pain stop. Because the self-narrative agent records all actions as "I did this", the act of surrendering to the pain has been recorded in memory as "I surrendered to the pain, and that made it stop".

The overall system has access to that memory; it infers something like "there was an action of 'surrender' which I carried out, in order to make the pain stop; I must carry out this action of surrender again". Rather than inferring that there is nothing which can be done to make the pain stop, making it useless to resist, the system may reach the *opposite* conclusion and look *harder* for the thing that it did the last time around.

The flaw here is that "surrendering" was not an *action which the resisting subsystem took*; surrendering was the *act of the resisting subsystem going offline*. Unfortunately, for as long as no-self has not been properly grasped, the mind-system does not have the concept of a subsystem going offline. Thus, successfully surrendering to the pain on a few occasions can make it temporarily *harder* to surrender to it.

And again, having an intellectual understanding of what I wrote just means that there is a subsystem that can detect that the resisting subsystem keeps resisting when it should surrender. It can then inject into consciousness the thought, "the resisting subsystem should surrender". But this is just another thought in consciousness, and the resisting subsystem does not have the capacity to parse verbal arguments, so the verbal understanding may become just another system resisting the mind's current state.

At some point, you begin to surrender to the fact that even if you have deep equanimity at the moment, at some point it will be gone. "You" can't maintain equanimity at all times, because there is no single "you" that could maintain it, just subsystems which have or have not internalized no-self.

Even if the general sense of self has been strongly weakened, it looks like craving subsystems run on cached models. That is to say, if a craving subsystem was once created to resist a particular sensation under the assumption that this is necessary for protecting a self, that assumption and associated model of a self will be stored within that subsystem. The craving subsystem may still be triggered by the presence of the sensation - and as the subsystem then projects its models into consciousness, they will include a strong sense of self again.

There is an interesting connection between this and Botvinick et al. (2019), who discuss a brain-inspired artificial intelligence approach. "Episodic meta-reinforcement learning" is an approach where a neural network is capable of recognizing situations that it has encountered before. When it does, the system reinstates the network state

that it was in when it was in this situation before, allowing old learning to be rapidly retrieved and reapplied.

The authors argue that this model of learning applies to the brain as well. If true, it would be compatible with models suggesting that emotional learning needs to be activated to be updated; and it would explain why craving can persist and be triggered even after global beliefs about the self have been revised. Once a person encounters a situation which triggers craving, that triggering may effectively retrieve from memory a partial “snapshot” of what their brain was like on previous occasions when this particular craving triggered… and that snapshot may substantially predate the time when the person became enlightened, carrying with it memories of past self-models.

People who have made significant progress on the path of enlightenment say that once you have seen through how the sense of self is constructed, it becomes *possible* to remember what the self really is, snapping out of a triggered state and working with the craving so that it won’t trigger again. But you need to actually *remember* to do so - that is to say, the triggered system needs to grow less activated and allow other subsystems better access to consciousness. Over time, remembering will grow more frequent and the “periods of forgetting” will grow shorter, but forgetting will still keep happening.

Managing to have a period of equanimity means that you can maintain a state where nothing triggers craving, but then maybe your partner yells at you and a terrified subsystem triggers and then you scream back, until you finally remember and have equanimity again. And at some point you just surrender to the fact that this, too, is inevitable, and come to crave complete equanimity less. (Or at least, you remember to surrender to this some part of the time.)

An example of a no-self experience

There was one evening when I had done quite a lot of investigation into no-self but wasn’t really thinking about the practice. I was just chatting with friends online and watching YouTube, when I had a sense of a mental shift, as if a chunk of the sense of having a separate self fell away.

I'd previously had no-self experiences in which I'd lost the sense of being the one doing things, and instead just watched my own thoughts and actions from the side. However, those still involved the experience of a separate observer, just one which wasn't actively involved in doing anything. This was the first time that that too fell away (or at least the first time if we exclude brief moments in the middle of conventional flow experiences).

There was also a sense of some thoughts - of the kind which would usually be asking things like "what am I doing with my life" or "what do others think of me" - starting to feel a little less natural, as if my mind was having difficulties identifying the subject that the "I" in those sentences was referring to. I became aware of this from having a sense of absence, as if I'd been looping those kinds of thoughts on a subconscious level, but then had those loops disrupted and only became aware of the loops once they disappeared.

The state didn't feel particularly dramatic; it was nice, but maybe more like "neutral plus" than "actively pleasant". After a while in the state, I started getting confused

about the extent to which the sense of a separate self really had disappeared; because I started again picking up sensations associated with that self. But when my attention was drawn to those sensations, their interpretation would change. They were still the same sensations, but an examination would indicate them to be just instances of a familiar sensation, without that sensation being interpreted as indicating the existence of a self. Then the focus of my attention would move elsewhere, and they would start feeling a bit more like a self again, when I wasn't paying so much attention to the sensations just being sensations.

A couple of hours later, I went to the sauna by myself; I had with me a bottle of cold water. At some point I was feeling hot but didn't feel like drinking more water, so instead I poured some of it on my body. It felt cold enough to be aversive, but then I noticed that the aversiveness didn't make me stop pouring more of it on myself. This was unusual; normally I would pour some of it on myself, go "yikes", and stop doing it.

This gave me the idea for an experiment. I'd previously thought that taking cold showers could be nice mindfulness exercise, but hadn't been able to make myself actually take really cold ones; the thought had felt way too aversive. But now I seemed to be okay with doing aversive things involving cold water.

Hmm.

After I was done in the sauna, I went to the shower, with the pressure as high and the temperature as cold as it went. The thought of stepping into it did feel aversive, but not quite as strongly aversive as before. It only took me a relatively short moment of hesitation before I stepped into the shower.

The experience was... interesting. On previous occasions when I'd experimented with cold showers, my reaction to a sudden cold shock had been roughly "AAAAAAAAAAA I'M DYING I HAVE TO GET OUT OF HERE". And if you had been watching me from the outside, you might have reasonably concluded that I was feeling the same now. Very soon after the water hit me, I could feel myself gasping for breath, the water feeling like a torrent on my back that forced me down on my knees, my body desperately trying to avoid the water. The shock turned my heartbeat into a frenzied gallop, and I would still have a noticeably elevated pulse for minutes after I'd gotten out of the shower.

But I'm not sure if there was any point where I actually felt *uncomfortable*, this time around. I did have a moderate desire to step out of the shower, and did give into it after a pretty short while because I wasn't sure for how long this was healthy, but it was kind of like... like any discomfort, if there was any, was being experienced by a different mind than mine.

Eventually I went to bed, and my state had become more normal in the morning, the mind having returned to a previous mode of functioning.

The way I described it afterwards, was that in that state, there was an awareness of how the sense of self felt like "the glue binding different experiences together". Normally, there might be a sensation of cold, a feeling of discomfort, and a thought about the discomfort. All of them would be bound together into a single experience of "I am feeling cold, being uncomfortable, and thinking about this". But without the sense of self narrating how they relate to each other, they only felt like different experiences, which did not automatically compel any actions. In other words, craving did not activate, as there was no active concept of a self that could trigger it.

This is related to my earlier discussion of the sensations of a self as a spatial tag. I mentioned that you might have a particular sensation around or behind your eyes - such as some physical sense of tension - which is interpreted as “you” looking out from behind your eyes and seeing the world. When the sense of self is active, the physical sensations are bound together: there is a sensation of seeing, a sensation of physical tension, and the self-narrative agent interpreting these as “the sensation of tension is the self, which is experiencing the sensation of seeing”. Without that linkage, experience turns into just a sensation of seeing and a sensation of tension, leaving out the part of a separate entity experiencing something.

This is the fifth post of the "a non-mystical explanation of the three characteristics of existence" series. The next part in the series is Three characteristics: impermanence.

Three characteristics: impermanence

This is the sixth post of the "a non-mystical explanation of the three characteristics of existence" series.

Impermanence

Like no-self and unsatisfactoriness, impermanence seems like a label for a broad cluster of related phenomena. A one-sentence description of it, phrased in experiential terms, would be that "All experienced phenomena, whether physical or mental, inner or outer, are impermanent".

As an intellectual claim, this does not sound too surprising: few people would seriously think that either physical things or mental experiences last forever. However, there are ways in which impermanence does contradict our intuitive assumptions.

A conventional example of this is change blindness. In a typical change blindness experiment, people report having good awareness of the details of a picture shown to them: but when details are changed during an eye saccade, subjects fail to notice any difference. Maybe a person's hat looks red, and people who have been looking *right at* the hat fail to notice that it looked green just a second ago: the consciousness of the green-ness has vanished, replaced entirely with red.

People are typically surprised by this, thinking that "if it was red a second ago, surely I would remember that" - a thought that implicitly assumes that sense percepts leave permanent memories behind. But as long something does not explicitly store a piece of conscious information, it is gone as soon as it has been experienced.

This is a natural consequence of the Global Neuronal Workspace (GNW) model of consciousness from neuroscience. As I have previously discussed, studies suggest that the content of consciousness corresponds to information held in a particular network of neurons called the "global workspace". This workspace can only hold a single piece of conscious content at a time, and new information is constantly trying to enter it, replacing the old information.

Now if the content of your consciousness happens to be something like this:

- 0 milliseconds: Seeing a red hat
- 53 milliseconds: Thinking about cookies
- 200 milliseconds: Seeing a green hat

Then at the 200 millisecond mark, unless some memory system happened to explicitly store the fact of seeing a red hat before, no trace of it remains in consciousness for the person to compare with. One can train particular subsystems to monitor the contents of consciousness and send occasional summaries of previous contents, which is part of what investigating impermanence involves.

Compare this to meditation teacher Daniel Ingram's description of impermanence:

> Absolute transience is truly the actual nature of experiential reality.

> What do I mean by “experiential reality”? I mean the universe of sensations that you directly experience. […] From the conventional perspective, things are usually believed to exist even when you no longer experience them directly, and are thus inferred to exist with only circumstantial evidence to be relatively stable entities. […] For our day-to-day lives, this assumption is functional and adequate.
>
> For example, you could close your eyes, put down this book or device, and then pick it up again where you left it without opening your eyes. From a pragmatic point of view, this book was where you left it even when you were not directly experiencing it. However, when doing insight practices, it just happens to be much more useful to assume that things are only there when you experience them and not when you don’t. Thus, the gold standard for reality when doing insight practices is the sensations that make up your reality in that instant. Sensations that are not there at that time are not presumed to exist, and thus only sensations arising in that instant do exist, with “exist” clearly being a problematic term, given how transient sensations are.
>
> In short, most of what you assume as making up your universe doesn’t exist most of the time, from a purely sensate point of view. This is exactly, precisely, and specifically the point. […] sensations arise out of nothing, do their thing, and vanish utterly. Gone. Entirely gone.

In Ingram’s terms, people subconsciously assume that if a person in a picture has a red hat, then the person in the picture is going to keep having a red hat. Also, if the person in the picture has a green hat, they probably also had a green hat when you last looked at them. This kind of an assumption is often pragmatically useful, and may even be a true claim about the world, for as long as the image you are looking at is not being manipulated by researchers who keep changing subtle details. But it is not an accurate model of how your own mind functions.

Consciousness as an FBI report

In a previous article, I mentioned that according to neuroscientist Stanislas Dehaene, one of the functions of consciousness is for subsystems in the brain to exchange summaries of their conclusions. He offered the analogy of the US president being briefed by the FBI. The FBI is a vast organization, with thousands of employees: they are constantly shifting through enormous amounts of data and forming hypotheses about topics with national security relevance. But it would be useless for the FBI to present to the president every single report collected by every single field agent, as well as every analysis compiled by every single analyst in response. Rather, the FBI needs to internally settle on some overall summary of what they believe is going on, and then present that to the President, who can then act based on the information. Similarly, Dehaene suggests that consciousness is a place where different brain systems can exchange summaries of their models, and to integrate conflicting evidence in order to arrive to an overall conclusion.

In a similar way, it’s usually not necessary for the brain to keep a conscious track of every little detail in an image. Rather, sensory information comes in, and subsystems responsible for processing it broadcast a summary of what they consider important about it. If you look at a painting, a general summary of its contents will be produced and maintained in consciousness, while minor details like the color of someone’s hat won’t be recorded unless a person had a particularly important reason to look at it. (That would be the equivalent of the FBI including a field agent’s random observations

in a report to the president. They're very unlikely to include those unless they are *really* important.)

This is particularly noticeable when learning to draw: as Raemon discusses in Drawing Less Wrong: Observing Reality:

> When you look at a person, what you perceive is not a series of shapes and colors that correspond to what's there, but rather a bunch of hastily constructed symbols that convey the information that the brain thinks is important. If you haven't rewired your brain for drawing, then "important" questions do not include "*Is that elbow angled at 90 degrees or 75?*" or "*Where are the eyes in relation to the top of the head?*" Instead, what you usually care about are things like "is this person happy, or angry?" and the information that gets recorded is a little tag that says "Smiling" with a vague curving-upwards-line symbol accompanying it.
>
> A large chunk of the information we usually need has to do with the face. This plays a role in two common biases that are near-universal in inexperienced artists:
>
> -Drawing the head much larger than it actually is, compared to the rest of the body
>
> -Drawing the "face" (i.e. everything between the eyebrows and mouth) as if they took up the entire head rather than the bottom half. Practically everything above the eyebrows conveys no relevant information, so it's just ignored.
>
> Your brain has a mental model of what a human is "supposed" to look like, and that model is wrong. You can see major gains in drawing capability just by learning the "ideal" proportions of a human being.

The relation of this to impermanence is that observing the contents of your mind lets you notice just how little sense data is actually used, and how quickly it vanishes. Returning to Daniel Ingram:

> We are typically quite sloppy about distinguishing between physical and mental sensations (memories, mental images, and mental impressions of other physical or mental sensations). These two kinds of sensations alternate, one arising and passing and then the other arising and passing, in a quick but perceptible fashion. Being clear about exactly when the physical sensations are present will begin to clarify their slippery counterparts—flickering mental impressions—that help co-create the illusion of continuity, stability, or solidity. [...]
>
> Each one of these sensations (the physical sensation and the mental impression) arises and vanishes completely before another begins, so it is possible to sort out which is which with relatively stable attention dedicated to consistent precision and to not being lost in stories. This means that the instant you have experienced something, you can know that it isn't there anymore, and whatever is there is a new sensation that will be gone in an instant. There are typically many other momentary sensations and impressions interspersed with these, but for the sake of practice, this is close enough to what is happening to be a good working model.

Ingram suggests that between physical sensations, there are mental sensations which "fill in the gaps", and which prevent people from noticing that the original physical sensations only come in sporadically. As people become more adept at meditation practices such as following the breath, they may come to notice that a large part of their time has been spent on following *a thought about the breath*, rather than *the*

breath itself: and far less sensory information about the breath actually comes to consciousness than they assumed.

In an earlier article on insight meditation I gave another example about these kinds of mental sensations. I mentioned a time when I was doing concentration meditation, using an app that played the sound of something hitting a woodblock, 50 times per minute. As I was concentrating on listening to the sound, I noticed that what had originally been just one thing in my experience - a discrete sound event - was actually composed of many smaller parts. The beginning and end of the sound were different, so there were actually two sound sensations; and there was a subtle visualization of something hitting something else; and a sense of motion accompanying that visualization. I had not previously even been fully aware that my mind was automatically creating a mental image of what it thought that the sound represented.

Continuing to observe those different components, I became more aware of the fact that my visualization of the sound changed over time and between meditation sessions, in a rather arbitrary way. Sometimes my mind conjured up a vision of a hammer hitting a rock in a dwarven mine; sometimes it was two wooden sticks hitting each other; sometimes it was drops of water falling on the screen of my phone.

Normally, all of this would just be packaged together into a general impression of "I'm hearing some sound". Our raw sense data is made up of countless small details and sensations, each arising and passing away in rapid succession - but we mostly perceive the high-level summaries, which are much more static. This creates an experience of seeing solid and discrete objects, and a feeling of there being permanent objects.

So how *does* one actually come to see what is happening in their mind?

In *The Mind Illuminated*, meditation teacher and former neuroscientist John Yates (Culadasa) suggests that one way this happens is by taking the subsystems responsible for producing such summaries and directing them to produce summaries about the content of consciousness. The brain already has a subsystem that generates overall summaries of what's going on in your mind; you can train that system to produce more detailed reports. Yates calls such summaries *introspective awareness* (discussed in more detail in an earlier article).

The impermanence of the self

As I have discussed, consciousness involves a constant competitive process, where different subsystems send content to the global workspace. At any given time, only one of these pieces of content is selected to become the content of consciousness. We might say that there has been a "subsystem switch" or a "subsystem swap" when the content of consciousness changes from that submitted by one subsystem to that submitted by another.

In normal circumstances, the structure of your mind is such that you cannot directly notice the different subsystems getting swapped in and out. Your consciousness can only hold one piece of information at a time. Suppose that at one moment, you are thinking of your friend, and at the next you are thinking of candy. When you think of candy, you are no longer aware of the fact that you were thinking of your friend the previous moment. You can often *infer* that a subsystem switch has happened, but you can't actually *experience* the switch.

However, if you develop more detailed introspective awareness, the stream of your consciousness may include reports such as this:

- Subsystem 1: So I was talking with my friend and she said…
- Subsystem 2: Ooh, candy.
- Awareness subsystem: The train of thought about my friend switched to a train of thought about candy right now.

Subjectively, this feels like becoming aware of the subsystem swapping in real time: a thought comes in, while an "afterimage" of the previous thought lingers for a brief moment, enough to make you realize that one kind of thought has replaced the other. If the trains of thought are different enough, the transitions between might feel really sharp and distinct.

You may also notice that you have two or more separate thought streams going in parallel, while having had no awareness of the fact. At one time you are thinking about your friend, and at the other time you are thinking about candy. Despite the fact that these two thought streams have kept alternating, maybe switching once every couple of seconds, they have been entirely unaware of each other. First the candy is everything that is in your mind, then your friend, then the candy again.

This is not to say that it would normally be impossible to be aware of having multiple trains of thought going on. Even without meditative training, your brain is constantly producing summaries of what's happening, including summaries of what's happening in your head. But what normally happens is something like having the first train of thought, then having the second train of thought, and then having general introspective awareness of there being two trains of thought. What does not usually happen is that the introspective awareness is sharp enough to register the fact that *whenever the train of thought switches, everything else disappears from consciousness for the duration*.

Rather than there being a single observer who experiences all of their own thoughts, there are three separate processes, two of them concerned with their own issues and a third meta-process keeping a loose record of what the two others have been up to.

A rough analogy would be to a (single-core) computer that keeps executing multiple different programs in succession, with the contents of the processor being cleaned out for the next program each time the execution switches. As long as everything goes smoothly, things will appear to the user as multiple different programs being executed at the same time, and the programs themselves will be unaware of the other programs. Yet, a sufficiently fine-grained trace of the different processes will reveal that only one has been running at a time. (Though unlike in this analogy, mental subsystems do keep running even when "swapped out"; they just don't have write access to consciousness during that time.)

By developing sufficient detail, another thing that can be noticed is that the sense of self is actually only present a part of the time. As discussed in previous posts [1, 2], the experience of a self is basically a piece of data - a *narrative* which is sometimes experienced and sometimes not. That is, it is another high-level summary of what is happening - "I am doing this thing" - constructed from lower-level data. (In a comment, Vanessa Kosoy suggested that the experience of a self is an explanation of *why* the person is doing things, constructed for social purposes and to be able to justify your behavior afterwards. This sounds plausible to me.)

That means that the content of your consciousness may be something like:

- Time 1: The sight of a bird outside the window.
- Time 2: The thought “there’s a bird over there”.
- Time 3: The experience of typing on a keyboard.
- Time 4: The sound of a car outside.
- Time 5: A mental image of a car.
- Time 6: A sense of being someone who sees the bird and hears the car, while typing on a keyboard.

… that is, normally you may experience there being a constant, permanent self which feels like *what you really are*. But in fact, during a large part of your conscious experience, that sense of self may simply not be there at all. Normally this might be impossible to detect due to what’s called the refrigerator light illusion: the light in a refrigerator turns on whenever you open the door, so it seems to you to always be on. Likewise, whenever you ask “do I experience a sense of self right now”, that question references and activates a self-schema, meaning that the answer is always “yes”. It is only by developing introspective awareness that records *all* mental content, without needing to make reference to a self, that you can come to notice the way in which your self constantly appears and disappears.

It is worth noting that coming to experience this may feel very frightening. Psychologist and meditation teacher Ron Crouch describes one way that it can go:

> What is actually happening, down deep, is that as your attention is syncing up with the dissolution of phenomena you are finding that there is nothing in experience that the sense of “me” can hold onto as stable and permanent. It just can’t get any footing. You do not realize it at a cognitive level, but you are getting a deep insight into the impermanence of all phenomena, and along with that, into the impermanence of the self. This is something that is terrifying to one’s very roots. Needles to say this initial stage can be a great source of distress and people can become stuck here for some time if they do not have good guidance.

We might think the distress follows from the mind’s underlying assumption that the self must be something like a permanent object. Whenever one has checked for the presence of the self, it has been there: thus, it is something that persists uninterrupted over time (except maybe in sleep). Now it - or something that resembles what it used to be - suddenly keeps vanishing and reappearing. Does that mean that you are dying?

Eventually, given enough further practice, the mind readjusts and revises its models. Continuity of consciousness does not mean uninterrupted continuity of self after all; the self is as impermanent as any other sensory experience. Nothing here to see, move along now.

Impermanence and unsatisfactoriness

One aspect of craving is *clinging,* a kind of repeated craving. The mind notices a pleasant or unpleasant sensation, and then tries to keep the pleasant sensations in consciousness and the unpleasant sensations out of consciousness. This may feel like you are trying to “freeze” the content of consciousness into a particular, pleasurable slice of experience.

In an earlier post, I gave a list of examples about craving; this is also a good list of examples to use for clinging, so I’ll repeat it here:

- It is morning and your alarm bell rings. You should get up, but it feels nice to be sleepy and remain in bed. You want to hang onto those pleasant sensations of sleepiness for a little bit more.
- You are spending an evening together with a loved one. This is the last occasion that you will see each other in a long time. You feel really good being with them, but a small part of you is unhappy over the fact that this evening will eventually end.
- You are at work on a Friday afternoon. Your mind wanders to the thought of no longer being at work, and doing the things that you had planned to do on the weekend. You would prefer to be done with work already, and find it hard to stay focused as you cling to the thoughts of your free time.
- You are single and hanging out with an attractive person. You know that they are not into you, but it would be so great if they were. You can't stop thinking about that possibility, and this keeps distracting you from the actual conversation.
- You are in a conversation with several other people. You think of a line that would be a really good response to what someone else just said. Before you can say it, somebody says a thing, and the conversation moves on. You find yourself still thinking of your line, and how nice it would have been to get to say it.
- You are playing a game of chess. You see an opportunity to make a series of movies that looks like it would win the game for you. You get so focused on the sequence of moves that would bring you a victory, that you don't notice that your opponent could also respond in a way that would ruin the entire plan.
- You had been planning on going to a famous museum while on your vacation, but the museum turns out to be temporarily closed at the time. You keep thinking about how much you had been looking forward to it.

What is essentially going on, is the craving trying to *fight against impermanence*. Taking the example of being sleepy and in bed: there is the sensation of sleepiness and a feeling of pleasure; *and* that annoying thought which keeps saying that you really need to get up soon... and the craving wants that pleasant sleepiness *back* and *stable*, damnit. If only it would focus on the sleepiness enough, maybe that annoying reminder would go away...

This contributes to the loop where the mind sees craving as necessary for well-being: phenomena won't stabilize in consciousness by themselves, and craving takes actions to make them more stable. Whenever it is unsuccessfully trying to do so, there is discomfort; when it succeeds in getting the pleasant thing to become the object of consciousness (if only for a moment), there is less discomfort (if only for a moment). Now, that discomfort is being generated *by the craving itself*, so it could also be eliminated by dropping the craving… but the system does not notice that.

Nor does it notice that following the craving does not lead to consistent happiness. Of course, we may *intellectually* understand that there's no single thing that would make us permanently and eternally happy. But at the subsystem level, each source of craving is based on a schema that states something like:

- If I get the thing I am craving, things will feel satisfying.

When the subsystem related to that goal is active, this is the schema which will be active in the person's mind. If you are hungry for food, you only think about how food will bring relief to your discomfort. Intellectually, you may know that soon afterwards you will start wanting something else - but the assumption that your mind is operating from, is that getting the food will bring contentment. And that assumption is correct! Recall that unsatisfactoriness is actually caused *by* craving. So getting the food *will*

make the craving for it go away - until the next craving pops up, which is likely to happen very soon.

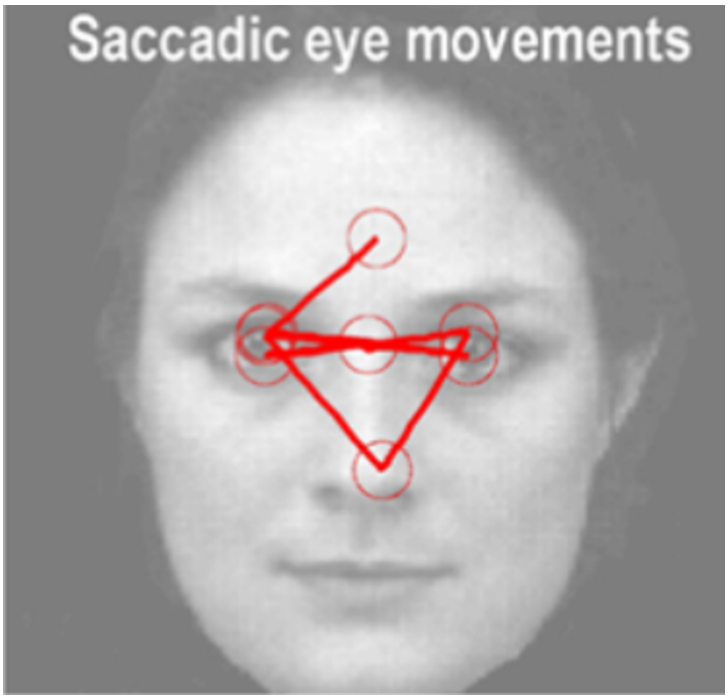
This has the consequence that each *individual* craving may have its prediction confirmed. The craving for food correctly predicts that food will bring satisfaction from that craving. The craving to look at the phone while you eat correctly predicts that looking at the phone will bring satisfaction from that craving. The craving to go watch pictures of attractive naked people after eating correctly predicts that going to watch pictures of attractive naked people will bring satisfaction from that craving… while the overall system remains in a near-constant state of craving that just keeps changing its target.

In the paper *Suffering* (Metzinger, 2016), Thomas Metzinger reports on an “experience sampling” experiment, where messages were sent to people’s phones at random times, asking them whether they felt that their current experience would feel worth reliving:

> For many, the result was surprising: the number of positive conscious moments per week varied between 0 and 36 [out of 70], with an average of 11.8 or almost 31 per cent of the phenomenological samples, while at 69 per cent a little more than two thirds of the moments were spontaneously ranked as not worth reliving.

Metzinger notes that one cannot generalize from these results to the general population: this was a small, unreplicated pilot study done with a highly selected group (philosophy students). But as he also notes, what *is* remarkable is that *nearly all* of the participants were surprised by their own results - they had expected many more moments to feel pleasurable. He speculates that human motivation may depend on systematic self-deception: if a person valued positive experiences but noticed that most of their experience was actually unpleasant, they might become paralyzed.

And it does seem that increased awareness of the impermanence of satisfaction helps reduce craving. I like to think of each individual craving as a form of a *hypothesis*, in the predictive processing sense where hypotheses drive behavior by seeking to prove themselves true. For example (Friston et al. 2012), your visual system may see someone's nose and form the hypothesis that "the thing that I'm seeing is a nose, and a nose is part of a person's face, so I'm seeing someone's face". That contains the prediction "faces have eyes next to the nose, so if I look slightly up and to the right I will see an eye, and if I look left from there I will see another eye"; it will then seek to confirm its prediction by making you look at those spots and verify that they do indeed contain eyes.



Eye movements seeking to confirm the hypotheses of "I am seeing a face". From [Friston, Adams, Perrinet & Breakspear 2012](#).

Normally, each craving is successfully proving true the hypothesis of "pursuing this craving will cause satisfaction"... but included in that prediction is not only a claim that satisfying this craving will bring momentary satisfaction. As the hypothesis is not modeling events that happen after it is satisfied, there is an implied claim that this will bring *lasting* satisfaction.

If the mind-system develops increased awareness of the way repeated craving seems to just lead to a constant state of discomfort, then under the right conditions it may consider the hypotheses in those cravings falsified and discard them.

Fighting against a sensation as assuming its permanence

Another thing that is happening is the subsystems failing to notice how fighting against a sensation actually helps *keep* it in consciousness, and how the sensation might actually fall away *on its own* if it was not being fought against.

Suppose that you are feeling stressed out over something, and a craving is activated to get rid of the feeling of stress. This involves sending into consciousness a plan for getting rid of the sensation of stress, which needs to *make reference to the sensation of stress*. This tends to redirect *more* attention towards the sensation of stress, strengthening the signal associated with it... and because sensations are normally impermanent and tend to easily vanish, this may help keep it in consciousness whereas it would otherwise have disappeared on its own.

In general, craving often operates under the assumption that unpleasant sensations are permanent: that is, they will persist in consciousness until actively resisted. And certainly it is true that not *all* unpleasant sensations will just disappear if you stop feeding them with attention. But even then, redirecting attention into a struggle against them may actively make them stronger.

If one develops sufficient introspective awareness, they may come to experience this directly. They will notice a neutral sensation, a negative sensation, another neutral sensation, then the aversion to the negative sensation… and notice there are actually quite a few neutral sensations, during which the negative sensation does not bother them at all. This helps notice that struggling against discomfort is actually not *necessary* for being free of discomfort; one is free of discomfort a large part of the time already.

Impermanence as vibration

No discussion of impermanence would be complete without touching upon the topic of "vibrations". Recall that according to the predictive processing model, the brain is composed of layers prediction machinery. A given layer can receive sensory information from a lower layer, and from the higher layer predictions of what that information *should* look like. For purposes of prediction, each layer is trying to form *models* of what it expects to see.

Rather than sense information primarily "flowing up" from the sense organs, the brain keeps making guesses of what it *expects* to see. These expectations are sent "down to the senses", with the brain using the sense data to *check* its assumptions and correcting for any mismatches. Mismatches that seem small enough may be ignored and explained away as noise.

One possible model is that the sensory information from the lower levels represents stable, permanent objects. As has been noted, this assumption is often a useful and correct one for predicting how the world behaves, so the system begins to assume it… ignoring the fact that sensory data is actually coming *in pulses* rather than constantly.

When one's consciousness starts dropping some of the mental impressions that normally "fill in the gaps", it may lead to an experiential quality of reality "vibrating". Here is how DavidM describes this:

> A meditator practicing in this style will eventually find that their experience is not static, but 'vibrates' or fluxes in a peculiar way over extremely short periods of time (fractions of a second). For an explanation by analogy, imagine a set of speakers playing music without dynamic variation; if a person rapidly turns the volume knob in the pattern off-low-high-low-off, the amplitude of the music will flux over time. Similarly, a meditator practicing in this style finds that the components of experience are not static, but fluctuate rapidly from nonexistent to existent and back again. N.B. This has nothing to do with the fact that the contents of experience are constantly changing. Rather, apparently static objects (e.g. an unchanging white visual field) turn out to be in flux.

For the most part, the hypothesis of "sensory data represents permanent objects" has turned out to deliver good results, so normally any gaps in the data will be automatically "filled in" by the model, as they are assumed to be meaningless noise. As a result, a "neural autocompletion feature" can create an impression of closely

observing sensory data, even when sensory data is actually sparse and the impression of it is mostly fabricated on the basis of a few data points.

For as long as the sensed data deviates only a little from the expected, the deviation is treated as noise and ignored; but once the deviation crosses some critical threshold, it is picked up and registered as surprising. If one intentionally goes looking for vibrations, then one is trying to pick up finer and finer distinctions in the sense data. This forces the system to pay attention to minor patterns that would otherwise have been treated as meaningless noise. That causes it to notice discrepancies between the higher-level model's prediction of "solid stream of sense data" and the sensory experiences that are coming in as pulses. This leads to an awareness of vibrations, and more generally insight into how the brain fills in data which is not actually there.

On the other hand, I have also heard reports of people finding vibrations without explicitly even looking at sensory details, in contexts such as doing loving-kindness meditation. I am confused about what is going on there and don't know how to explain it. This is also an area that I have personally investigated relatively little.

This is the sixth post of the "a non-mystical explanation of the three characteristics of existence" series.

Beliefs as emotional strategies

In “Crony Beliefs”, Kevin Simler suggests the analogy of beliefs as employees in a company, who may be hired to carry out different tasks. “Meritorious” ones, he suggests, are in the business of accurately modeling the world, whereas “crony” employees are there just to win favor from others.

I like the analogy (“employees within a brain” sounds a lot like “subagents”) and the idea of beliefs that have been "hired" for the purpose of currying favor from others, but I think the essay could go further still.

I think that many beliefs that do not appear *at all* crony are nonetheless rooted in complicated social strategies and that one’s thinking can drastically change by altering the assumptions behind those strategies. The strategies in question can be much subtler than just “believing what others in your current social group believe”. Rather, such socially motivated beliefs can include deeply-held lifelong conceptions of metaphysical topics such as the nature of free will, morality, and talent.

Let’s look at this in the light of three real-world case studies.

Believing in free will to earn merit

(The person in question reviewed this section before publication and let me know which parts were fine to share.)

Some time ago I happened to talk with someone who mentioned that they had a strong need to believe in non-deterministic free will. They knew that this made no intellectual sense, but they still needed to maintain this as a compartmentalized belief. At a time when they were younger, they'd stopped believing in free will and had gotten into a suicidal depression. The only way they got through was by coming to believe in free will again.

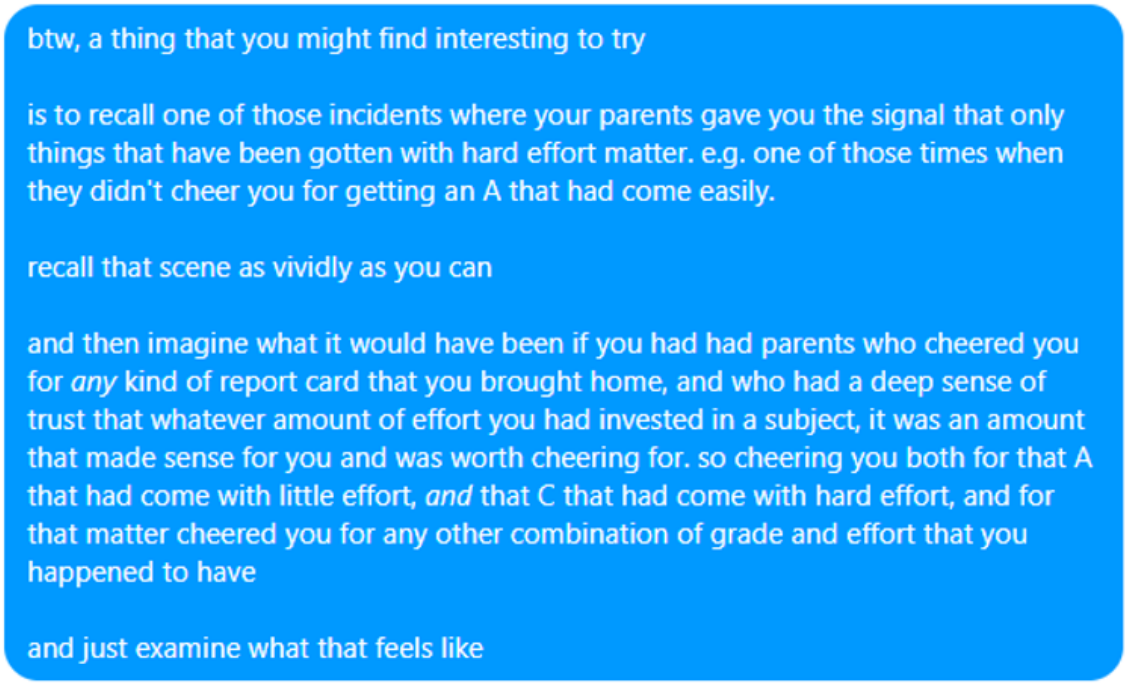
They explained that they didn’t know *why* they needed to believe in it, and figured that their motivational system was just intrinsically wired up that way. It felt like a hardwired, unquestionable innate need.

In response, I said that in my experience, these kinds of things are learned emotional strategies of some sort. They were skeptical of that claim… but thought about it for a while anyway. After a while, they told me that they had been raised to believe that only things you got with hard work mattered. Most of their life, they had specifically chosen to do hard things because doing easy things had made them feel unbearably empty.

The way they thought about it, for something to be hard work, it has to be your own choice. And something can only be your own choice if it’s not predetermined, i.e. if non-deterministic free will exists.

As an example of where that came from, they mentioned that their parents had never praised them for As that had come easily for them, but *had* praised them for Cs that had required a lot of hard work.

Since they seemed to be open to exploring this, I suggested a juxtaposition experience for them. What would happen if they tried recalling an incident where they *had* brought such a report card home, getting praise for the hard-earned Cs but not the effortlessly acquired As… and then imagined an alternate scene where they brought the same report card home, but had their parents just be happy with them regardless of what their grades happened to be and regardless of how they were earned?



As a result of doing that, they said that they had started crying, and they felt like the need to do everything the hard way had vanished. My model of what happened was that their emotional brain had noticed their parents only valuing hard work, and generalized that to "hard work is the only thing that *anyone* ever values". Once they noticed that they could coherently imagine having parents who did *not* make such a valuation, the generalization - and the resulting need to believe in free will - dissolved.

Sometime later I asked them how they currently felt about free will. They replied:

> ahw, thank you for asking 😀 I actually had quite some thoughts about this since we last spoke, but they were all rather disjointed and didn't want to flood you
>
> I didn't really realize in advance how much of a sort of "lynch pin" belief the free will actually is in my mind. It makes sense looking back on it now, cause it was so resistant to logic that I perfected a whole separate cognitive structure to protect it without giving up logic
>
> and I mean lynch pin as referring to this sense that "well if I let this thing go, then this whole flood of other things follow"
>
> So most of my thoughts have been around the repercussions of letting that feeling of "free will has to be a thing" go, and it recasts a pretty large part of my adolescences
>
> like ... I think the most fundamental part I can identify at the moment is this: If free will is not a thing on any level, then there are no merit points that can be earned by exerting it. Exerting free will was often a matter of fighting against one's natural urges. Basically about not picking the "easy path". The harder the thing you do, the more fighting is involved, the more sacrifice - the more merit points you got for doing the free will thing, because it shows you're strong. But then if you scratch that ... 1) how do you gain any merit points anymore? And if you don't gain merit points, then how do you have value or does your life have meaning? and 2) If there are no merit points to gain by doing the hard thing, then you might as well do the thing you WANT ... the thing you're naturally attracted to and feels right, and you don't have to feel guilty about that

> I think point 1 would explain why I could never give up the free will feeling. I had nothing to substitute it with, and back when I formed the belief it was part of my defense against suicidal depression. Which made me think that maybe the timing of this conversation between us is also fortuitous cause for my brain, having kids is so meaningful, I don't really need anything else to make myself feel fulfilled (I mean, more is a plus, but entirely not necessary).
>
> But also think point 2 might explain why I got depressed in the first place ... like if you internalize that you only have value, and your life only has meaning, if you pick the path that inherently makes you unhappy (cause it's the hard path and not what you really want) then of course that's going to make you depressed, and neurotic, and anxious

Moral obligation as defense against arbitrariness

I had long felt varying degrees of dissatisfaction with various aspects of my work, but kept getting stuck doing the things I knew how to do well enough to get paid for. During 2020 I made a deal with myself to only continue with my current thing until the end of the year and then quit, so that I would be forced to find something that felt more satisfying.

Separately, I had also intended many times to take a break from all the effective altruism, save-the-world kind of work in order to just focus on myself and my own needs for a while, but kept getting pulled into one do-gooding thing after another. This would *finally* be an opportunity to just take that break, spend a few years just doing something that felt good for its own sake, without worrying about what kind of an impact (if any) it would have on the world.

Then at one point, as I was considering various things that I could be doing, I noticed that there were some options that *most* of my mind was on board with… but there was still a lingering sense of an “effective altruist guilt”, some part of me saying that just doing something that’s personally meaningful rather than globally impactful is *selfish* - that even if I did something that was valuable for *some* people, if it wasn’t valuable for *the most* people, it was wrong.

So I had a belief saying something like “if you have the option of doing something that helps a lot of people and an option that only helps a few people (if any), then it is morally wrong to take the option that helps fewer people”. In late February, I started investigating that belief, and found that it was associated with a felt sense of having wanted/needed something when I was little, but being told that I couldn’t have it.

And *that* felt sense was associated with a feeling that the only reason why I couldn't have the thing(s) was some senseless social custom that overruled my needs - but that I also couldn't present any case for myself, because the custom felt so arbitrary and senseless that there was nothing to argue with. It wasn’t clear exactly what experience it was drawing on - to some extent it felt like the amalgamation of many similar experiences - but it did feel connected to vaguely recalled memories of being told that I couldn’t do or have something because I was too old for it now, with no other rationale being offered.

As a result, some part of my mind had revolted against the notion of arbitrary social customs overruling people’s needs. It was saying something like “people's needs *actually matter*, you've *got to help them* if they are suffering, you *are not allowed* to invoke social custom as a reason not to". Maybe it thought that if the people around me had followed this rule, I wouldn’t have been treated in a way that from a child’s perspective felt senseless and arbitrary. Maybe if I could make the people around me believe in that, then I would have my needs better taken into account in the future.

That rule, in turn, led very naturally to "I **personally** have to help people if they are suffering, I am not allowed to invoke ordinary social norms about the whole world's suffering not being my personal responsibility".

Having found this child part of me, I imagined an experience of him being given love and being understood, and him getting to keep doing whatever thing he wanted to keep doing until he naturally grew out of it (or alternatively he could just continue doing those things forever, if there genuinely wasn't any reason to stop doing them).

Doing that seems to have reconsolidated the emotional belief of "my needs won't be understood unless people are coerced into caring about suffering", which also dissolved the belief of "**I** have to be coerced into helping whoever is suffering", and since then I haven't had any "effective altruist guilt" associated with thinking about my life choices.

Lack of talent and people's selfishness

In both of the previous examples, the belief in question was relatively simple. Sometimes there's a more complicated web of interacting beliefs and behaviors.

My friend Sampo Tiensuu does math and physics tutoring as well as mental coaching to people trying to get into med school. He has told me of a pattern of beliefs that he has found with some of his clients.

A client has a tendency to easily give up whenever they face difficulties, believing any obstacles to be a sign that they are insufficiently talented. Poking at the memories and associations behind an underlying belief of "if I encounter difficulties, that means I'm insufficiently talented to deal with them on my own" eventually turns up the following:

The client's mother has had the belief that people are selfish at heart and only interested in exploiting others. The mother assumes that if she punishes others for their selfish behavior, then the fear of punishment will cause them to act more altruistically. This is linked to a belief that if people *were* fundamentally good, then there would be no major problems in the world; but because there *are* major problems in the world, then someone must be responsible for them and needs to be punished.

Now if the mother's child gets bad grades in school, this has to also be someone's fault. It could be the child's fault for not having worked hard enough, but if the child didn't work hard enough, then *that* would be the *mother's* fault for not having raised her child right. This, in the mother's belief system, would imply that *she* was selfish and deserving of punishment.

To avoid feeling guilty, the mother has to make her child do better in school. One possibility would be to arrange remedial teaching for her child. But if her child got extra lessons that other children didn't, then this would mean that she was getting her child an unfair advantage. This would also make her feel selfish.

The mother avoids this line of thought by deciding that her child's poor performance is the school's fault, and that it's the school's responsibility to help her child catch up. To make the idea of remedial teaching compatible with her conception of fairness, she needs to believe that there's something wrong with her child, such as a fundamental lack of talent. If her child is fundamentally untalented, then their poor grades are not the mother's fault, but they are the school's fault for not taking her child's learning problems seriously enough. The mother can now make helping her child feel justifiable to herself, while also feeling like a good person because she's punishing the school and the teachers, who she has shown to be selfish and thus in need of punishment.

The mother's belief that her child's poor grades reflect a lack of talent is then absorbed by the child, who ends up becoming demotivated whenever they end up having difficulties with

their studies. Eventually they end up in a coaching session with Sampo, who helps them bring into awareness the experiences during which their mother communicated this belief to them. After being successfully guided to imagine a mother who *didn't* feel a need to attribute the client's poor grades to a lack of talent, the belief reconsolidates and dissolves, and the client no longer experiences difficulties as intrinsically demotivating.

A unified view of belief

In Simler's original essay, he considers crony beliefs to be a different kind than ordinary "merit" beliefs. However, I think there exists a frame in which these are *not* two entirely distinct categories: rather both are special cases of a more general category of "belief". I'll say more about this in later posts.

Puzzle: how are both merit beliefs and crony beliefs the same kind of thing?

My current take on Internal Family Systems “parts”

I was recently asked how literal/metaphorical I consider the Internal Family Systems model of your mind being divided into “parts” that are kinda like subpersonalities.

The long answer would be my whole sequence on the topic, but that’s pretty long and also my exact conception of parts keeps shifting and getting more refined through the sequence. So one would be excused for still not being entirely clear on this question, even after reading the whole thing.

The short answer would be “it’s more than just metaphorical, but also not quite as literal as you might think from taking IFS books at face value”.

I do think that there are literally neurological subroutines doing their own thing that one has to manage, but I don’t think they’re literally full-blown subminds, they’re more like… clusters of beliefs and emotions and values that get activated at different times, and that can be interfaced with by treating them *as if* they were actual subminds.

My medium-length answer would be… let’s see.

There’s an influential model in neuroscience called global workspace theory. It says that the brain has a thing called the “global workspace”, which links together a variety of otherwise separate areas, and its contents corresponds to that what you’re currently consciously aware of. It has a limited capacity so you’re only consciously aware of a few things at any given moment.

At the same time, various subregions in your brain are doing their own things, some of them processing information that’s in the global workspace, some of them observing stuff from your senses that you’re currently not consciously aware of. Like you’re focused on a thing, then there’s a sudden sound, and some auditory processing region that has been monitoring the sounds in your environment picks it up and decides that this is important and pushes that sound into your global workspace, displacing whatever else happened to be there and making you consciously aware of that sound.

I tend to interpret IFS “parts” as processes that are connected with the workspace and manipulate it in different ways. But it’s not necessarily that they’re really “independent agents”, it’s more like there’s a combination of innate and learned rules for when to activate them.

So like, take it when an IFS book has a case study about a person with a “confuser” part that tries to distract them when they are thinking about something unpleasant. I wouldn’t interpret that to literally mean that there’s a sentient agent seeking to confuse the person in that person’s brain. I think it’s more something like… there are parts of the brain that are wired to interpret some states as uncomfortable, and other parts of the brain that are wired to avoid states that are interpreted as uncomfortable.

At some point when the person was feeling uncomfortable, something happened in their brain that made them confused instead, and then some learning subsystem in their brain noticed that “this particular pattern of internal behavior relieved the feeling of discomfort”. And then it learned how to repeat whatever internal process caused

the feeling of confusion to push the feeling of discomfort out of the global workspace, and to systematically trigger that process when faced with a similar sense of discomfort.

Then when the IFS therapist guided the client to “talk to the confuser part”, they were doing something like… interfacing with that learned pattern and bringing up the learned prediction that causing confusion will lessen the feeling of discomfort.

There’s a thing where, once information that has been previously only stored in a local neural pattern is retrieved and brought to the global workspace, it can then be accessed and potentially modified by every other subsystem that’s currently listening in to the workspace. I don’t fully understand this, but it seems to be something like, if those other systems have information suggesting that there are alternative ways of achieving the purpose that the confuser pattern is trying to accomplish, the rules for triggering the confuser pattern can get rewritten so that it’s no longer activated.

But there’s also a thing where, it looks to me like part of what these stored patterns are, are something like partial “snapshots” of your brain’s state at the time when they were first learned. So when IFS talks about there being “child parts”, then it looks to me like there’s a sense in which that’s literally true.

Suppose that someone first learned the “being confused helps me avoid an uncomfortable feeling” thing when they were six. At that time, their brain saved a “snapshot” of that state of confusion to be re-instated at a later time when getting confused might again help them avoid discomfort. Stored with that snapshot might also be associated other emotional and cognitive patterns that were active at the time when the person was six – so when the person is “talking with” their “confuser part”, there’s a sense in which they really are “talking with a six-year old part” of themselves. (At least, that’s my interpretation.)

And also there’s a thing where, even if the parts aren’t literally sentient subselves, the method still becomes more effective if you treat them *as if* they were.

If you relate to your six-year old part as if it was literally a six-year old that you’re compassionate towards, when it holds a memory of being lonely and not understood… then that somehow brings in the experience of someone actually caring about you into the memory of not being cared about.

And then if your brain had learned a rule like “I must avoid these kinds of situations, because in them I just get lonely and nobody understands me”, then bringing in that experience of being understood into the memory rewrites the learning and eliminates the need to so compulsively avoid situations that resemble that original experience.