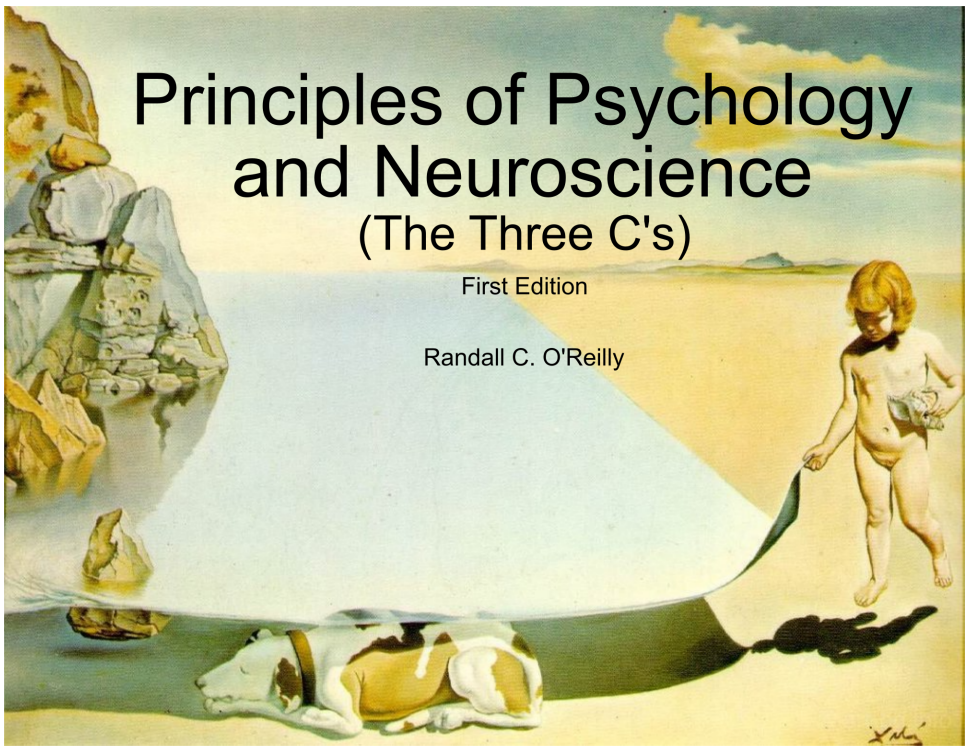
Principles of Psychology and Neuroscience, First Edition

Randall C. O'Reilly


Principles of Psychology
and Neuroscience
(The Three C's)
First Edition
Randall C. O'Reilly

Contents

Published by Open Textbook, freely available

https://github.com/PsychNeuro/ed1

To my family.

Preface

This is an in-progress experiment – feedback is more than welcome.

Chapter 0: Introduction

Introductory Psychology textbooks typically provide a rather fragmented, fact-laden view of the field, relying on colorful graphics, exciting news stories, and personal anecdotes to generate interest in the material. This book represents a radical departure from that approach. Instead, the goal here is to provide a simple, succinct, principled account of the human mental world and how it emerges from our brains, that is coherent across the scope of phenomena covered in typical Intro Psych texts.

The overall portrait painted of you looks something like this (caution: it is not overly flattering, and you might even think this song is not about you, but go ahead and be vain – it is): You are obsessed with controlling your environment to satisfy a range of core desires and to mitigate strong fears. You are unlikely to be swayed by other people's advice, but have no problem dishing it out. A challenge to your social standing or any other form of disrespect (the *diss*) is one of the worst offenses. You are willing to spin all manner of stories to maintain your sense of order in the world, *especially* when that sense is strongly challenged, often to the point of absurdity in the eyes of others.

You crave simple ways of understanding the world, to the point of massively oversimplifying the true complexities and ambiguities, preferring to think in terms of concrete anecdotes instead of broad abstractions, logical arguments, or, especially, statistics. You think you know how most stuff you use everyday works (bikes, cars, toilets..), but studies show that you are actually remarkably clueless – how exactly does that chain on a bike work? Perhaps most glaringly, you can't help but think in terms of stereotypes, and inevitably focus on information that is consistent with your existing views, while ignoring all those nagging hints that all may not be as simple as you might like.

You only care about things that are new and unexpected, and are constantly comparing and evaluating yourself and others with a keen eye for who is doing better or worse along any number of important dimensions (wealth, beauty, smarts, athletic ability, popularity – you name it!) You are hypersensitive to who might be cheating or gaming the system, but are perhaps not so aware of unfair advantages you might have. More generally, you tend to think of yourself as being "your own person" and strongly underestimate how strong of an influence other people actually have over you. If you're honest with yourself, you'll admit that you spend way too much time thinking about what what other people think of you – without recognizing that everyone else is doing the same thing, so that in fact the answer is a somewhat disappointing: "not much" (unless of course you do something embarrassing or strange or stupid, but even then, your memory of those events will typically far outlast those of others).

In other words, you are a *survivor*. You are a tough cookie. Your ancestors survived unbelievable hardships to get you here, to your relatively plush college-educated world. You are amazingly efficient. All those crazy details you don't know about the world are largely irrelevant anyway. Seriously, does it really matter that you don't know how the engine or transmission in your car works?

You can drive, and get to where you need to go – and that is what really matters. Your brain is exquisitely tuned into what really matters, and despite over 60 years of attempts to recreate the magic of your brain in a computer, nothing has come even close (despite all the recent media hype to the contrary).

And yet, despite all your toughness and amazing abilities, you are very likely to have at least some level of significant mental dysfunction. You are more likely than not to suffer from depression, anxiety disorders (and often both of those together), drug dependence, etc. Unfortunately, the promise of a magic pill to cure these afflictions has turned out to be yet another disappointment. In fact, regular old "talking to another human being about your problems" (i.e., therapy, which is actually somewhat more involved and structured than that) is likely to be more effective than medication for most people.

The Three C's

Surprisingly, we can make sense of all the above (and more!) using only three core principles:

Compression

Each neuron in the most important part of your brain (the *neocortex*) is wired for simplification, and the collective effect of the massive waves of electrical activity surging through your brain every millisecond is to compress, reduce, and simplify information. Each neuron receives input signals from roughly 10,000 or more other neurons, but guess how much it can then say about that flood of information coming in? Almost nothing. First of all, it only has *one* output signal, the *spike*, which is an all-or-nothing affair. Furthermore, a typical neocortical pyramidal neuron will fire at most around 100 spikes in a second. And a second is a relatively long time in the inner loops of the brain – there is evidence that 1/10th of a second represents a kind of fundamental time-frame for information processing, so those 100 spikes reduce down to just 10 spikes within that critical window. And most neurons are firing far less rapidly than that. It's like when you tell your friend all your deepest thoughts, and they just say "huh". Neurons are the strong, silent type most of the time. But still waters run deep: when neurons *do* get excited about something, it is likely to be *important*, and most of what they are doing is *shielding you from constant TMI* (too-much-information – but you knew that already, so, kind of a meta thing we got going there...)

The raw scene coming into your eyeballs is truly gory: all jumbles of light, motion and color. When you were a tiny baby, you were overwhelmed by this "blooming buzzing confusion", but now your neural networks have learned and developed to the point where you don't (can't!) even see that raw sensation anymore (unless of course you partake of various hallucinogenic substances, but even then, the level of disorder experienced is trifling compared to the pure chaos of the raw, unfiltered tidal wave of sensation coming in). We get small,

fascinating hints of the magic power of our perceptual systems through illusions, and the occasional "viral gold / blue / brown dress" controversy, where people strongly see or hear very different things from the very same stimulus. But overall, we really have absolutely no idea how much undercover cleanup work is going on inside our brains. If anyone was truly aware of the level of conspiracy operating in there, it would be scandalous. But, somehow, amazingly, we largely all end up converging on the same stable, boring illusions of simplicity. A table. A chair. Some french fries. People walking down the street. Cars driving by. Nothing strange going on here.

We would be utterly nonfunctional without this compression. For the same reason that those hallucinogenic drugs render people nonfunctional. If you want to do something useful with your time, you need to be able to make everything else in the world boring and irrelevant, so you can focus on *what matters*. If you're reading a book, or your tiny screen, it simply wouldn't work if every time you moved your eyes, the whole world was seen afresh, requiring you to reorient and rediscover what you were just reading and what you need to read next. Interestingly, this capacity for perceiving a stable, boring world seems to depend critically on a very active underlying process of *prediction* – your brain is stitching everything together in a seamless whole by filling in the gaps with what you *expect* or *predict* to see. You can easily see this, and relive some of your earliest experiences, by simply closing one eye, and then gently pushing on the bottom of the eyelid of your other, open eye. Suddenly, the world is a moving jumbly mess again! (Seriously, try it!)

Your brain's penchant for simplification (compression) does not stop with perception. Your highest levels of thought are similarly dominated by the same quest to render everything simple and predictable. Instead of recognizing the incredible high-dimensional diversity of our fellow beings, we inevitably reduce everyone to stereotypes. Even members of negatively stereotyped groups are caught in the evil maw of this process, exhibiting similar levels of stereotype-driven biases as everyone else. The ultimate expression of this compression process is the *anosagnosia of everyday life* (aka the Dunning-Kruger effect; NY Times Article: `https://opinionator.blogs.nytimes.com/2010/06/20/the-anosognosics-dilemma-1/`) – the lack of knowledge about our utter lack of knowledge. People can be remarkably unaware about what they don't know, and sometimes, this leads to funny situations. But, amazingly, most of the time, *it causes no obvious problems whatsoever*. We just keep getting on with our lives. And, as with perception, if we didn't, we'd never get anything done, because there is such a huge amount of stuff we routinely, safely ignore, that it would take many many lifetimes to process and understand it all.

Contrast

The next principle explains why we seem so fixated on comparing ourselves with others. Not just any others, but those certain people *who really get to you*. In that inexplicable, frustrating way. Why do I always have to be so jealous of those

people? Can't I convince myself that the "grass is always greener?" Nope. As with compression, your brain is wired at the lowest level for magnifying contrasts, in this case via a special class of neurons called *inhibitory interneurons*, coupled with other important properties of all neurons that we'll cover in Chapter 2. The net effect is that your brain only sees things *relatively* (yep, we can have our own, special, relativity law in Psychology too – actually it is pretty general). A classic example of this is when you come in from the bright sunny outdoors into a dimly-lit room. The difference in raw light energy coming into your eyeballs in these two situations is enormous, but, after a brief period of adaptation, you're seeing things in the dim room that differ by a few photons here or there, whereas outside those few photons would be a miniscule drop in the bucket. In other words, our neurons *normalize* away the raw strength of whatever signal is coming into them, and remain sensitive to the *relative differences* compared to that overall signal. Those inhibitory neurons play the critical role of mathematically *dividing away* the raw signal strength, leaving the principal pyramidal neurons "in the zone" for responding to relative differences.

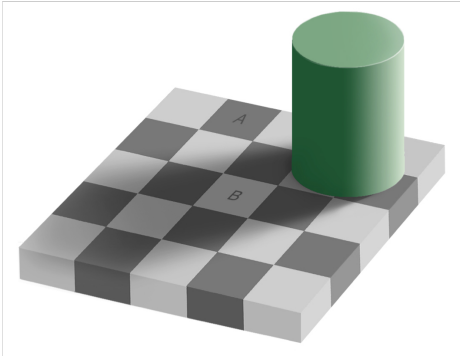


Figure 1: Fig 1-1: Illustration of the power of contrast in perception. Do you think the physical image-level color of square A is the same or different from that of B? Unbelievably, they are identical!

As with compression, perception provides some of the clearest windows into this phenomenon, for example Figure 1-1, showing the remarkable effects of contrast (and global scene understanding) on perception of color and brightness. Another remarkable example is the case of *perfect pitch* – why is it so unusual a skill for people to simply be able to recognize the absolute frequency of a sound? Mechanically, and mathematically, extracting such a frequency from a sound signal is trivial, and simple (a "Snark"-like guitar tuning device can be had for a few bucks). That this feat is so incredibly rare and difficult in humans just points to the pervasive power of the contrast relativity (most people can easily tell the relative pitch).

But contrast, like compression, is not restricted to the perceptual domain. It affects every level of thought, contributing to that insidious obsession with your relative standing among your peers. For example, studies routinely show

that the absolute amount of money that people make is largely unrelated to various measures of their happiness – instead, what matters is their perceived level of income *relative to their peers*.

Contrast operates over time as well, in several important ways. First, at the perceptual level, we are highly sensitive to the rate of change over time of stimuli. The classic example here is the slow approach to boil being unnoticed by a hapless frog until it is too late (this is not exactly true – you can't get all the way to boiling, but it is very likely true to at least some extent). Similarly, a cottage industry of amazing demonstrations of our inability to detect slow changes in visual scenes sprung up a few years ago: (YouTube Link to Dan Simons Video: `https://www.youtube.com/watch?v=1nL5ulsWMYc`). Once you become aware (upon repeated viewing or instruction) of the nature of these changes, it is truly astounding to realize how much you overlooked them the first time(s). If you rapidly flip between the start and end frames of these slow-moving videos, the changes pop immediately into view. Again, we see the *delta*, not the absolute value of things.

Nowhere is this more poignant, and pressing, than deep inside the *dopamine* system in the middle of your brain. As you already know (and would be annoyed to have me repeat, but I'm going to do it anyway, to prove the pertinent point), dopamine is widely believed to be the "pleasure drug" in your brain. It is associated with drugs of addiction, and actually most other major mental disorders in one way or another. However, this popular description of dopamine leaves out one of the most important points: dopamine is *not* about *raw* pleasure, but rather, about the *difference* between what you experience and what you *expected* to experience. Specifically, if you get exactly what you expected, your dopamine goes "meh". This soul-crushing response to your greatest accomplishments is exactly what critics do to performers, and indeed the dopamine system is best understood as being the *central critic* of your brain. Far from a center of epicurian delights, it is a hard-nosed bully that is never satisfied. And that dissatisfaction is what has driven us ever upward in all manner of exploits – many of them good, but many of them not so good.

Greed is really a byproduct of your dopamine system. Seriously, why in the world can't someone who already has *millions of dollars* just be happy with that, and give the rest away or do something else useful with it (and their all-too-brief lives). Because dopamine adapts quickly to that million-dollar feeling, and it keeps giving you back that critical "meh" response. You need more than that. You *deserve* more than that... It really is tough living with such an asshole critic in your head all the time. But then again, we really do owe every step of "progress", within our individual lives and as a society and a species, to that nasty little critic driving us ever upward and onward.

Finally, another obvious manifestation of the contrast principle is our collective obsession with the *news*. Especially with the advent of the 24hr news cycle and the constant updating of news information via electronic, online media, we are now living in a quickly-moving bubble of news that sweeps things up in its path and spits them out quicker than... yesterday's news. Or yesterhour's

news. If you don't check in quickly enough, you'll miss huge swaths of news that everyone would have definitely been aware of before. Everyone worries about this kind of thing, but really it is just what our brains are wired to do. Every conscious moment of our lives is driven by a thirst for knowing what has changed, what is different – anything that remains constant will quickly drift out of your mind, like that delicious aroma of dinner that I can no longer access, or, thankfully, that feeling of my butt sitting on this chair that I was thoroughly *so over with* until I just wrote that sentence..

Control

Last but certainly not least, is our obsession with control. Some of you may be thinking that you're not a control freak like those *other people*, but actually, every one of us is a crazy control freak at some level – it just differs in terms of what matters to us. Anyone want to have some stranger come pick you up and take you around to work with them all day? Or just invade your personal space? How would you feel if someone just started selling all your stuff on craigslist? Or how about those people who go door-to-door (or stop you on the street) and try to convince you to believe in some particular brand of religion? Or just your roomate who keeps nagging you about the dishes, or being to loud, etc. Yeah, there's definitely *something* for *everyone*, where it matters. And usually, if you have two or more people living together, you quickly become aware of all that stuff that you didn't realize really matters to you. A lot.

Starting again in the brain, virtually every neuron in the brain is serving the master of control at one level or another. At the most basic level, control is about *motor control*, and a great example of the dedication of the brain to this particular function comes from a lowly sea squirt that starts off life as a mobile tadpole, and flits around in the ocean for a bit, looking for a good place to settle down. As soon as it finds its special place on the reef, it promptly eats its own brain! Because, the whole point of the brain in the first place, evolutionarily speaking, is to process sensory inputs *in the service of producing useful motor outputs to improve survival and the overall quality of life*. There's a reason nobody thinks highly of layabouts and 30-year-old's living in their parent's basement: progress requires action, and our brains are wired for action. In the brains of most species, there are big chunks devoted to the compression and contrast processing of sensory inputs, and the rest is devoted to using that information to figure out what kinds of opportunities and threats are out there in the world, and how to best optimize chances of survival within the repertoire of available motor actions. Not much space left over for cultivating expertise in civil war battles, or fantasy role-playing games, or whatever other weird, seemingly non-functional things people spend their time doing.

The human brain takes this obsession with motor control to the next level, by building an internal fortress / castle of the *self*. We're not quite sure to what extent any other beast even has a similar kind of thing inside their own mental worlds. The self is a model, a construct, built up over years, that helps

us predict (again) how we are going to behave, and what we seem to really want (and not want). By having such a thing inside our own brains, we can use it to more accurately anticipate what kinds of motor actions are really going to get us what we want. This is especially important when dealing with other people, who are, compared to your average rock or tree, very complicated and unpredictable. I'm not saying you're a manipulative little jerk. I'm saying *everyone* is a manipulative little jerk, deep down. It is, again, just a logical extension of what brains are supposed to be doing. If they aren't good for maximizing pleasure and minimizing harm, then we might as well all just eat them for dinner!

This *self model* lying at the heart of our control system is like our secret nuclear power reactor inside our brains. It is the "nerve center" of our being. It does *not* take existential threats kindly. Anything that appears to threaten our internal sense of identity and control gets raised to the red alert level. This is why you can't just "mansplain" something to someone else, and expect them to instantly see the error of their ways, and instantly become a new, better self. We have a lot of investment in that *old* self, and it does not look kindly on being deposed from its despotic rule over its own internal kingdom.

Although the self is a despot at heart, it is also remarkably sensitive to external, social forces, creating one of the most fundamental and puzzling paradoxes of the human condition: We care deeply about what other people think of us, and are actually remarkably malleable in adapting our behavior under the influence of others. There are many demonstrations of the power of the social force, from the evil of Nazi Germany and controversial attempts to recreate those forces in the lab, to the seemingly more benign and amusing phenomenon of hypnosis. Biologically and ecologically, our very survival is utterly dependent on our ability to work together socially, and social motivations are undoubtedly wired directly into the depths of our brains, providing these "hijack" pathways past the watchful eye of the self-model.

And therein lies the likely explanation for this paradox: these social forces can only act when delivered in ways that the self either does not recognize as threatening, or even endorses. The minute you are aware someone is trying to convince you of something, is the minute that it fails. But when a social virus is neatly packaged in a nice sugar coating, often in terms of reinforcing a sense of belonging with an identified *in-group*, then it can easily slip past the guards. These kinds of in-group / out-group (tribalism) dynamics are the strongest of social forces and underlie all the greatest evils of humanity. And probably many of our greatest triumphs too.

Developmentally, the self emerges around age two, heralded by the onset of *tantrums*. Tantrums are the inevitable consequence of an emerging desire for control, coupled with an almost complete lack of *actual* control. This is really the defining battle of life, and it never really ends: the best you can hope for is some kind of truce as expressed in the Serenity Prayer of Reinhold Niebuhr: "God, grant me the serenity to accept the things I cannot change, Courage to change the things I can, And wisdom to know the difference."

The Breakdown of Control

Unfortunately, achieving *serenity now* is very difficult. And all those challenges to the self can end up leading to a bout of depression, often coupled with anxiety or other unpleasant mental states. Although widely characterized in terms of *anhedonia* or the inability to experience pleasure, current research supports the idea that the core disorder of depression is really about *control*, or the perceived lack thereof. When your self model is sufficiently challenged, it basically gives up on a lot of goals, and unfortunately, achieving those goals is a primary source of pleasure and satisfaction in life. So, yes, anehdonia is a consequence of depression, but the core of it is more about the inability to motivate yourself to get out of bed and do all those now-meaningless things that you used to find meaningful.

Consistent with this central role for control, one of the most promising components of modern therapy for treating depression is *behavioral activation*, which is essentially an attempt to reboot your core self-motivation control system. Indeed a major study found behavioral activation to be the most important element among a group of therapies, and as effective as medication (Dimidjian et al, 2006). And when you recognize the central role for control in depression, it is then less surprising that medications are relatively ineffective: for the vast majority of people, the problem is *not* about some kind of low-level imbalance in their brain chemistry: it is about their core mental power plant running out of steam. And it just takes hard mental work, aided by effective theraputic treatments, to reboot your own sense of mental self-control and efficacy.

For the smaller proportion of people who clearly do have a biologically-based mental disorder, it is still the case that the brain areas most centrally involved in self-control are the ones that are most likely to be affected. Schizophrenia and OCD for example involve the frontal cortex, basal ganglia, and dopamine systems of the brain, which are the main players in developing and sustaining our internal self control system. Thus, understanding how different parts of the brain function to support this critical self-model system is a major goal of current research in Psychology and Neuroscience, and this book is designed to get you started on a journey toward understanding this cutting-edge work.

Other Principles and Perspectives

There are many candidates for "the fourth C", and different names could have been chosen to refer to the above "three C's" (e.g., reduction, relativity, and... respect?), but being a slave to the simplifying force of *Compression*, it is useful to try to see as much as possible through the lens of these three principles. Furthermore, as briefly introduced above, these principles can be tied directly into the most fundamental properties of the nervous system, and thus provide a critical *bridging function* between Neuroscience and Psychology. Nevertheless, it is important to always remain aware of all the compressing taking place, and to acknowledge that this radical attempt at synthesis may strike many practicing scientists as overly simplistic or downright wrong-headed. However, my hope

is that the benefits outweigh the costs overall, without attempting to overly minimize those costs.

Where do we go from here?

This question can be asked at two levels: the short-term question of where this book is headed, and the longer-term question of where our species is headed!? Although it may seem like our current cultural and political environment reflects an extreme magnification of many of the negative aspects of human mental function as described above, another perspective is that these truly are perennial battles and challenges that we have struggled with since the dawn of human history, and that they are borne of fundamental properties of the human brain that also have many positive aspects. Like everything it seems, double-edged swords abound. And the core premise, and promise, of science is that by understanding something deeply, we are better positioned to make the best of it. This contrasts with the idea that by somehow reifying "bad" features of the human brain, we are therefore justifying the bad ends they produce. Clearly that is not the aim here, and my personal optimism leads me to believe that this endeavor will be a net benefit in the end (or at worst, simply irrelevant).

With those big picture questions out of the way, we can turn to the plot for the rest of this adventure story through the human brain. Unlike a good mystery story, we're going to ruin the whole thing right up front, in the hopes of achieving a better understanding and mental roadmap in the bargain.

Chapter 1 will provide a big-picture overview of the challenges and promise of achieving a scientific understanding of the brain and the mind (particularly the mind). The main challenge here is the *subjective* nature of the subject matter – however, we see that this subjectivity is actually primary and a fundamental challenge for all science, and a major challenge for people more generally!

Chapter 2 will cover the nuts and bolts of the brain, but always connected directly to the bigger picture via the three-C's principles and their applications. We'll see in detail how each neuron functions as such an amazing "information compactor", compressing those 1000's of signals into its single spiky output. We'll then take an amazing "connected" voyage through the pathways of the neocortex, seeing how the great chain of neurons locked in their long-lasting embraces create channels where information flows in different ways. We'll wrestle with the central question of whether brain areas are truly "specialized" for different functions or not, and whether there is any "there" there, as in, "where *is* that memory anyway?"

Chapter 3..

etc.

Chapter 1: Science and Subjectivity – The Fundamental Challenge of Psychology

Psychology is the science that attempts to understand the human mind. The human mind is the most fascinating and amazing "thing" in the known universe, and the idea that you can actually attempt to study it using the basic reductionistic approach of science may seem a bit of a stretch. And indeed it has been – but at this point in the development of the field, most practicing scientists are likely to feel rather confident that significant progress has been made, without fundamental, obvious limitations to how far we can go.

Despite all this progress and optimism, we will see in this chapter that there actually are fundamental boundaries to what science can penetrate, and these boundaries have shaped the field from its inception. Thus, understanding these limitations helps put the field of psychology and neuroscience into perspective in multiple ways, and in fact many of the limitations we discuss apply to science, and all human knowledge, more broadly.

The central issue we must confront head-on is the inescapable problem of *subjectivity*. By subjectivity we mean not just the fact that different people have different opinions or perspectives on things, though that is a big part of it. Instead, we need to step back a bit to look at the *really big picture* (i.e., Philosophy), starting with the fundamental problem of subjectivity as expressed by **Rene Descartes** (way back in 1637), in his famous statement: *Cogito Ergo Sum – I think therefore I am.*

There are two essential implications of this statement – we'll explore the first one in depth before turning to the second. The first implication is that *subjective experience is primary*. If you put yourself into the mindset of a very skeptical, doubting philosopher, you might just about be able to get yourself to question everything, *except* this one, primary fact: you are sitting there (wherever you are), *thinking*. If you really push it, you might appreciate that you can't really be sure that the world itself exists outside of your mind! This very challenging train of thought is well-captured in several modern movies, perhaps most notably in the *Matrix* series, where, in fact (in the movies at least), there turns out to be every reason to have such doubts. In philosophical circles, this line of thinking is known as *solipsism*, and lest you think that this is just an irrelevant and obscure way of thinking, one of the great innovators of our time, Elon Musk, is apparently convinced that we're all living in a giant simulation.

This is the kind of all-encompassing subjectivity that we want to more fully understand and appreciate. What does this line of thinking mean for the study of psychology, or science more generally?

This is where we can usefully bring in Descartes' second major implication from *Cogito Ergo Sum*: *dualism*. Dualism is the idea that there are two fundamentally different "substances" in the universe: the regular physical stuff of the everyday world, and this entirely separate, magical transcendent thing called *mind*, which lives apart from that other, regular stuff. The opposing view is

called *materialism*, where the mind is seen as just a product of the material world like everything else, and in particular a product of the physical processes taking place within the *brain*, as widely embraced in modern neuroscientific approaches to psychology.

You might be somewhat surprised to hear that many modern-day philosophers still embrace dualism, and one of the most outspoken advocates is David Chalmers, who argues that understanding the nature of subjective experience, or *qualia*, is the *hard problem* of consciousness and simply cannot be explained in objective, materialistic, scientific terms.

You might also be surprised to hear that, despite being one of those modern materialistic neuroscientists, I actually agree with Chalmers, and Descartes (in spirit at least, so to speak)! I think that there are two fundamentally different "somethings" in the universe, but, unlike Descartes and Chalmers, I don't think the dividing line is between *mind* and *matter*, but rather, between *subjective* and *objective* perspectives.

Following Descartes (again), we can take subjective experience as primary – it is the only thing I am fully certain of. But it is also primary in another, essential way: it is uniquely, completely, definitionaly, *mine*. It is literally impossible for *you* to experience *my* subjective experience, because, by definition, *my* subjective experience is exactly the sum-total of what it "feels like" to be me. If we somehow were to add *you* into my brain, my subjective experience would be irreparably altered. If you are somehow sharing in my subjective experience as it is happening, you would have to have direct access to every level of my brain, and not just "objective" access as you might get from a super-hi-tech future brain scanner, but *direct*, *internal*, *subjective* access, "from the inside out".

In other words, you would have to literally be inside my brain. And you can't be inside my brain because I'm already here. From the materialist perspective, we can identify my subjective experience as emerging directly from my brain – it is what it feels like to be my brain. If you truly appreciate this equivalence, then it should be readily apparent than there can be only one "mind" for every brain (we'll look into the fascinating phenomenon of multiple personality disorder later, but it doesn't change this fundamental conclusion – all those personalities are just as irrevocably trapped inside the one brain as you and I are, and in fact we all have something like multiple personalities too).

Another way of thinking about this is in terms of identical twins. Let's imagine we have the most identical of identical twins ever to exist. Their brains are *completely identical* in every way possible. Would those twins have the same subjective experience? No. They might have a great deal in common, but, fundamentally, they would not, and could not, directly experience exactly what the other is experiencing. Why not?

It all boils down to *perspective*. Each physical thing in the universe has its own unique perspective, if we take this term to mean a particular spatial location, and a particular trajectory through space and time in the past (and going onward into the future), that is fundamentally *unique* to that thing. This

is why the twins cannot share their subjective experiences: they are two separate, distinct things, and, inevitably, they "see the world" from two different vantage points. The only way they could share experiences is if they could somehow superimpose themselves into exactly the same point in space, and do so over a sufficiently long time period to synchronize their history of experience, which plays such a critical role in our subjective life, in addition to the immediate sensations coming in from the outside world.

Anyway, the key point of all this is that *if* you allow that subjective experience can never be shared among different brains, *then* it follows that there is a fundamental divide between this inner subjective world, and the "regular" outside *objective* world. I believe this divide captures the essence of what Chalmers is talking about in terms of the irreducible nature of the qualia of consciousness – the impossibility of trying to explain in objective terms "what it feels like" to experience things in our subjective, inner world. Furthermore, it does so without introducing anything particularly magical or fundamentally at odds with materialism: subjective experience is not separate from the physical world in terms of some kind of magical "substance" that it is constituted from – it is just separate in terms of this notion of *perspective* – the unique point of view (literally, where they are standing / sitting / looking) that each subjective being has all to themselves.

Subjectivity in Psychology: A Brief History

Stepping back from this big philosophical abyss, what does it all mean for the attempt to study psychology as a science? The primary, obvious problem is that psychology is the study of *what it is like to be a human being*, and if this is fundamentally a subjective thing that can never be directly shared with any other human being, how can we possibly hope to arrive at some kind of objective, scientific understanding? Well, the first step is to follow Chalmers and attempt to *partition the problems* – we can carefully attempt to set aside the *hard problems* associated with the nature of subjective experience, and focus instead on the so-called *easy problems* that are left over. *If* there is enough interesting stuff left over in this space of easy problems, then it probably makes pragmatic sense to just see how far we can get in trying to understand that stuff, and then, once we seem to have exhausted that space, perhaps we could circle back and start reconsidering some of those hard problems.

This overall approach provides a reasonable narrative for the history of psychology as a scientific discipline. The person most widely credited with founding the science of psychology, **Willhelm Wundt**, had the innovative idea in the late 1800's that, after millenia of armchair speculation, you could actually apply the techniques of empirical science to understanding the human mind / brain. Wundt made many groundbreaking contributions, but his legacy, at least at the level of introductory psychology texts, is as a founder of the *introspectionist* school of psychology, which also includes **William James**, who also made major lasting contributions to the field. When the next major paradigm shift took

place in the early 1900's, it emerged as as strong reaction and rejection of this introspectionist approach, which was characterized as being overly concerned with all those hard problems of subjective experience. Introspectionists would try to systematize and characterize the contents of subjective experience, and the hard-nosed *behaviorists* who came next regarded these investigations as insufficiently objective, rigorous, and replicable. Instead, they emphasized purely objective, externally-observable *behavior* as the only valid data in psychology (hence the term behaviorism). The main figures in this era (e.g., **John B. Watson, B. F. Skinner, and Ivan Pavlov**) focused on how external, objective factors such as reward and punishment affected subsequent behavior through *conditioning*.

Thus, these first two epochs of scientific psychology embody exactly this tension between the subjective and objective worlds. The next paradigm shift took place in the 1950's and 60's with the *Cognitive Revolution*, riding the wave of digital computers, which made it fashionable to start talking about internal mental operations in terms of the *information processing model* of the mind – i.e., the mind as a computational device. Scientists leading this new field, such as **Herbert Simon** and **Alan Newell**, started thinking about how the mind could perform complex mental operations such as scientific proofs, chess, and other challenging tasks (Newell and Simon 1972). People created running computer models of how these internal thought processes might work, which provided a compelling way to render that formerly "loosey-goosey" internal world in a much more rigorous, objectively-characterizable way.

However, as parallel work in the field of Neuroscience continued to advance, it gradually became clear that the brain really doesn't work anything like a standard digital computer. Instead, it is really a *massively parallel* computer with billions of computing elements (neurons) that combine the functions of computation and memory, which are otherwise separated in a standard digital computer. Psychologists **David Rumelhart** and **James McClelland** published a ground-breaking pair of books in the mid 1980's that popularized this new understanding of how information processing might work in the brain (Rumelhart and McClelland 1986; McClelland and Rumelhart 1986), and subsequent advances in the ability to take high-resolution pictures of the activity inside the human brain (*neuroimaging*) have led to the currently-dominant paradigm that integrates neuroscience and cognitive psychology (i.e., *cognitive neuroscience*) to come up with coherent understanding of how exactly the brain gives rise to the phenomena of the mind.

Fundamentals of Cognitive Neuroscience

This book is grounded squarely in this new paradigm of cognitive neuroscience, and attempts to provide a coherent set of core principles that connect directly from the basic processing carried out by individual neurons, all the way up to the highest levels of mental life. We are still largely avoiding significant consideration of the vast inner world of subjective life, but there is a robust field studying the *neural correlates of consciousness* (NCC) that we will discuss in depth in Chapter

X. Slowly but surely, we are building bridges between the objectively-identifiable properties of the human brain, and the subjective experiences that tend to co-occur with particular such brain states. Thus, we are developing a richer objective understanding about the kinds of neural mechanisms that give rise to our subjective mental life. But even with all of these advances, I don't think we could ever explain to a non-human-brain lifeform what it feels like subjectively to be a human brain. Thus, the subjective world remains our own private dominion, and probably literature, art, and movies provide the richest vehicles for sharing those experiences across the inevitable subjective gap between us all.

Subjectivity and Science: Working with the Method

The challenges imposed by the primacy of subjectivity have far-reaching implications beyond the field of psychology. First, given that some people can't even agree that there *is* an objective, external world outside the mind, how can we possibly even begin to start talking about *objective knowledge* and *facts*? This appreciation for the primary nature of subjective experience forces us to recognize that objective knowledge itself is entirely dependent on the subjective motivation of individuals to entertain a strong enough belief in this notion of objective reality, to put up with all the effort it takes to make any progress in understanding and advancing objective knowledge.

Those individuals are called "scientists", and they follow a particular method, the **scientific method**, which has the following basic steps:

1. Come up with a general question or problem, e.g., based on an informal **observation** about something of interest (e.g., Newton observes the apple falling on his head, which gets him thinking..)
2. Form a specific **hypothesis** about how that something might work, which makes testable **predictions** (e.g., there is an invisible force called *gravity* that causes all objects to experience the same accelleration, making the testable prediction that a feather and a hammer should fall at the same rate *in a perfect vaccuum* so as to eliminate the "confound" of friction).
3. **Collect data** that could actually test the predictions of the hypothesis, in comparison to other possible hypotheses (e.g., measure how fast things fall, ideally in a vaccuum if you happen to have one of those lying around). It is essential that the data be collected using a well-specified procedure that could be **replicated** by other scientists.
4. **Analyze the data** to determine whether any effects observed are strong enough to be clearly distinguishable from random chance and noise.
5. **Draw conclusions** – how compelling are the data, what holes are there in the data that would allow other hypotheses to explain the observed effects, etc?
6. Iterate! Plug the holes, think of other alternative explanations, test those, etc.

These steps can incrementally pull us out of our individual subjective

fortresses through the critical lever of **consistency**. If you articulate a clear sequence of steps to perform an experiment, and tell me exactly what you observe as results, and I do the same thing to the best of my ability, and get *consistent* results, then it seems like there might be something *real* and *objective* going on, or at least the world isn't completely random. As more and more people do the same thing, and continue to get consistent results, the odds that each one of us is just being individually tricked by some kind of subjective illusion would seem to go down.

As this scientific process continues, ever broader networks of interconnected hypotheses and associated empirical data accumulate, and if all of these remain somehow consistent with each other, it really starts to seem like there might be some kind of *laws* governing the behavior of the outside world. Furthermore, all this scientific knowledge makes its way into technology, which depends on those same laws, further bolstering the network of consistency. Fast forward to the modern world, and we now have the *standard model* of physics that provides a single consistent framework for understanding virtually all physical phenomena that have been subject to experiment, and drives incredible technology that would have been considered pure magic in times past.

Despite all this amazing progress made through the iterative application of the scientific method, you still have people like Elon Musk, one of the great *users* of physical laws, nevertheless concluding that it is all a giant simulation. And still plenty of people who believe that the Earth is flat, etc. And there is *nothing* you can do to convince these people otherwise. Such is the ultimate primacy of our subjective perspective on the world: the *only* porthole we have onto that supposed objective reality out there is through our very own, individual, subjective lenses. Because our subjective worlds are fundamentally uniquely our own, this also means that nobody can force anyone to believe anything that they aren't otherwise prepared to believe. Objective reality really is a second-class citizen, and is entirely dependent on the patronage of the ruling, sovereign subjectivty, just as scientists are still to this day dependent on the hard work and wealth of others to have the luxury of time and resources to create this huge network of consistent hypotheses and data.

Even within the scope of the scientific method, subjectivity abounds. Where, exactly are these hypotheses, or conclusions, supposed to come from? How many scientists looking at the exact same empirical data draw the same conclusions? You'd be surprised how subjective and inconsistent cutting-edge science really is. History is full of examples where a visionary pioneer was ridiculed by their colleagues, until enough evidence accumulated, and enough old people in power died, to allow the new ideas to flourish. The widely-accepted description of how science actually works, developed by Thomas Khun in 1962, emphasizes this sociological, psychological reality of science, with one major consequence being the strong suppression of ideas that are inconsistent with the current paradigm.

We can understand this phenomeon in terms of the three C's principles. Compression says that people crave simplicity, and the current paradigm embodies that: it is something that a large number of people know and agree

about. Having that overturned requires confronting a high level of uncertainty and complexity. Control is paramount here: that challenge to a widely-believed paradigm is experienced as a direct, personal challenge to your entire mental fortress – psychologically, it is really the same as challenging someone's belief in a particular religion. Furthermore, the uncertainty directly undermines the feeling of control as well. And control interacts with contrast – the "paradigm believers" constitute a social in-group, and anyone challenging the paradigm is immediately a strongly-contrasting out-group member, and all the deep tribal motivations are aroused in this case, causing the challenger to be treated like a real outcast and pariah.

In other words, science is just people being people. However, despite all our limitations and inevitable subjectivity, there is some indication that following some approximation of the scientific method really does seem to work, at least over the longer arc of history.

Before we get more into the nuts and bolts of actual experiments and statistical analysis techniques in psychology and neuroscience, there is one further perspective on the problem of subjectivity in science that bears mentioning. This comes from Robert Pirsig, who wrote the famous book, *Zen and the Art of Motorcycle Maintenance*, which is actually more about philosophy of science and personal autobiography, rather than Zen per se. Pirsig literally went insane (as in, institutionalized, electroconvulsive shock therapy, etc) in the course of struggling with the quesiton of where hypotheses come from – he realized that there was no rational explanation for how to come up with a good hypothesis, and it seems like there could easily be an infinite number of plausible hypotheses, so this throws a massive monkey wrench into the entire rational foundation of science.

Thus, subjectivity, creativity, and individual genius truly lie at the heart of science – most scientists are reasonably capable of evaluating hypotheses in terms of their consistency with data and with the larger network of other validated hypotheses, but relatively few scientists are responsible for coming up with the major hypotheses in the first place. Oh, and by the way, Pirsig suffered from Schizophrenia so that probably had more to do with his mental breakdown than the problem with hypotheses, but anyway it makes for a good story.

Research Methods in Psychology and Neuroscience

After all that philosophy, you might find a bit of concrete research methods a refreshing change! In this section, we'll discuss the specific types of data that psychologists and neuroscientists tend to collect, and what kinds of analyses are typically done with that data. This is the kind of thing that almost everyone agrees about, and we will cover it very succinctly because it all sounds perfectly logical, but actually applying it requires a good deal of practice and experience, which is beyond the scope of this book, and likely the course you're currently taking.

In psychology, there are three major ways in which data is collected, each with complementary trade-offs:

- **Descriptive Methods** – these tend to be the least *invasive* techniques, involving various ways of capturing what is actually happening in human behavior, such as observation, case studies, and surveys. A modern version employs cell phones with apps that ping people at random times during the day and ask them what they're doing, or thinking about, etc. The disadvantage of these techniques is in their relative inability to inform you about *why* people might be behaving the way they are – the other two techniques improve on that aspect of things, but, particularly with the experimental method, tend to require more artificial, less naturalistic kinds of experiments.
- **Correlational Studies** involve measuring multiple different **variables** (something that can be measured which varies across people, such as weight, IQ, vocabulary, diet, etc) and determining the extent to which these variables **correlate** or vary systematically in relationship to each other. For example, people's weight and height tend to be positively correlated, because as one goes up, the other does too. Critically, as with most real-world data, this is not a **perfect** correlation – there are many exceptions in either direction – but overall, on average, there is a relationship. The single most important limitation of correlational studies, is that the presence of a **correlation does not imply causation**. Typically, causation does imply correlation of some sort, but this relationship is not symmetric! Unfortunately, the human brain relies on correlation as a kind of "quick and dirty" shortcut for finding causal relationships in the world, and we find it remarkably difficult to recognize that the two are not equivalent. For example, most studies on the effects of diet on health are correlational, and yet the media and even scientific papers regularly interpret these as showing a causal link. "Drink more coffee because you'll live longer!" Well, what if in fact the observed correlation between coffee and longevity is due to the fact that more wealthy people drink more coffee, and it is really the wealth and all its associated benefits that is driving the longevity. Coffee is just "along for the ride". This is the **third variable problem** (in this case, the third variable is wealth), and it is the bane of correlational studies, because *there is always a third variable* (and a fourth, and a fith, etc). And it is typically very difficult to rule out the possiblity that everything is being caused by one of these unmeasured "third variables".
- **Experimental Studies** are the only way to truly establish a causal relationship, and even then it is still a major challenge to really accomplish this feat. The key trick is to use randomness and careful designs to attempt to systematically eliminate all possible "third variables". A huge source of third variables is each individual person participating in the study. Like all the bacteria on your skin, you are crawling with third variables. Your genes, your upbringing, your neighborhood, your schools, your friends, your... everything, is a teaming cesspool of third variables! The key trick

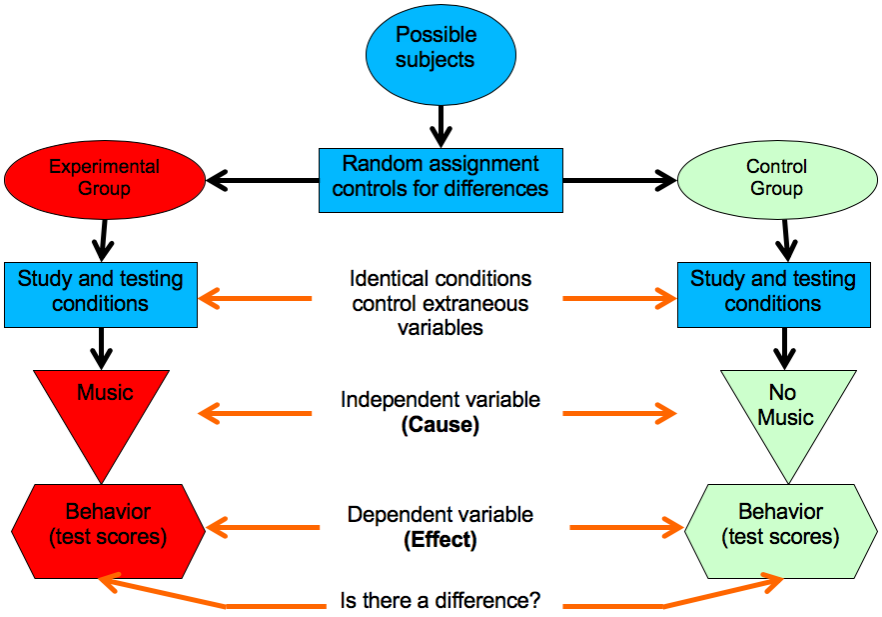


Figure 2: Fig 1-1: Logic of an experimental study, using random assignment to eliminate third variables from the study participants. It is also essential to minimze all other differences between the experimental and control conditions (i.e., *confounds*, or additional "third variables"), to more precisely identify the single *independent variable* (i.e., the *causal* variable) as truly being responsible for the differences measured in the *dependent variable*

in an experimental study is to use the cleansing power of randomness to wash away all those third variables, by **randomly assigning people to different conditions**. No third variable can withstand the incredible power of such random assignment – if we find a systematic difference between two completely random samples of the population, it cannot be due to their pre-existing conditions! However, random assignment is also the achilles heel of experimental studies, because it is often impossible to use random assignment for many questions of interest. Can you really look at the effects of parenting style on subsequent emotional development, by randomly assigning kids to parents!? Same goes with any long-term study on things like diet and lifestyle – you can sometimes sorta force people to eat some particular diet over a period of a few months or so, but that just isn't going to work for the decades it likely takes for most diet effects to really impact overall health outcomes. There are also other important ways of eliminating further possible third variables (typically called **confounds** in this context) from experiments, but random assignment is the most important (see Figure 2-1 for a diagram of the overall logic).

Thus, each of these different techniques is most appropriate for different kinds of questions, given the different tradeoffs. The key thing as a student and a citizen is to understand the limitations of any given study, so you can make an informed decision about what it really means. And don't expect the media to do this for you. Seriously, look at *any* correlational study on health / diet / etc and see how clearly the story, or the original article, discusses the limitations on any kind of causal implications from the study.

Neuroscience methods

Methods in neuroscience (and cognitive neuroscience) tend to be either correlational or experimental. The vast majority of **neuroimaging** studies are purely correlational, measuring the neural correlates of various different tasks or other manipulations performed while participants are in the brain scanner. By now, the neural correlates of just about every possible human activity (yes, including sex) have been measured in a scanner. But because of the correlational nature of these results, it is difficult to know whether the recorded brain activity is just *epiphenomenal* (i.e., just along for the ride), or whether it is really causal and somehow *responsible* for the behavior in question.

To attempt to address this causality question, scientists have used various forms of electrical and magnetic stimulation, which can disrupt or enhance neural firing in a relatively localized region of the brain. For example, **transcranial magnetic stimulation (TMS)** applied over the primary motor cortex can cause your muscles to flinch. However, just as with other experimental studies, the resulting brain states after TMS are not very "naturalistic", and it becomes difficult to interpret whether any changes in observed behavior are due to the disruption of the "normal" functioning of that brain area, or whether they just reflect the weird stuff that happens when you tweak that brain area in a

completely unnatural way.

In animal neuroscience, much more precise causal inferences can be made by employing much more "invasive" techniques, such as directly cutting out different parts of the brain, or using modern **optogenetic** techniques to instantly and reversibly activate or deactivate a given population of neurons. These optogenetic techniques allow very specific populations of neurons to be targeted, and have produced a powerful new wave of causal empirical data, showing that very precise manipulations to very specific neural populations can sometimes have impressive overall effects. However, often even these results are over-interpreted and one must look very carefully for confounds in the resulting activity of other neural populations. Virtually every neuron in the brain is within a few synapses of every other neuron (i.e., the "6 degrees of separation" (from Kevin Bacon) phenomenon), so it remains very difficult to isolate what each specific subset of neurons is uniquely contributing. Indeed, as we'll see in the next chapter, the very premise of isolating specific functions may be entirely misguided.

Finally, animal neuroscience also affords much higher-resolution neuroimaging techniques which can resolve the activity of individual neurons, while also recording many such neurons at the same time. Such techniques provide the most powerful descriptive methods for characterizing what neurons actually do, and historically have been some of the most important data for fueling our theorizing and understanding of how the brain works.

Thus, truly each different type of technique plays a critical role in the overall arsenal of science.

Statistics

Finally, it is useful to be aware of the most widely used statistical techniques in psychology and neuroscience. Here is a brief overview:

- **Descriptive Statistics** – like descriptive methods, descriptive statistics are used to describe data, and differ from **inferential** statistics which are used to *infer* causality or correlation, as described below. The primary descriptive statistics are probably familiar to you: *mean*, *median*, *mode*, *range* and *standard deviation*. For a *normal* (bell-shaped, *gaussian*) distribution, the mean, median, and mode are all the same, and they tell you where the *middle* of the distribution is (i.e., the "average" person, etc). It is only when the distribution is *skewed* that they differ, with the mode and median being less "pulled" by the long-tailed side of the distribution. You may have heard of income being reported in terms of medians – this is because income is a skewed distribution, with progressively fewer people making a *lot* more money than the mass of the "middle class" and below. The median and the mode more accurately capture this "middle class" salary because they don't get pulled upwards as much by all the rich people.
- **Correlation Coefficient and Scatterplots** – these are the primary tools for correlational studies. The correlation coefficient is a number,

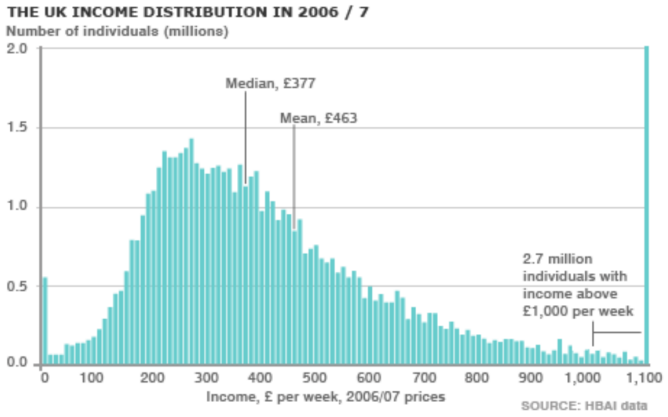


Figure 3: Fig 1-2: Mean, Median, and Mode tell different stories when the distribution is skewed (in this case, it is *right*-skewed – the skewer is the long tail to the right). The mean is pulled up by the tail much more than the median or mode, which do a better job of capturing the "middle class" income.

typically labeled r, which goes between -1 and 1, where -1 represents a perfect negative correlation, 0 is the complete absence of a correlation, and 1 is a perfect positive correlation. Imporantly, both a strong negative and a strong positive correlation are equally important statistically, and indeed you can almost always just flip one of your variables around and turn one into the other (e.g., height vs. weight is positive, but "shortness" vs. weight is negative). A scatterplot simply plots the value along each variable (one on the X or horizontal axis, and the other on the Y or vertical axis), with each dot representing a different person (or whatever else is being measured). Thus, you can usually directly see the strength of the correlation in the shape of the "cloud" of such points (todo: include standard figs). One critical "pro tip" for looking at such scatter plots is finding "outlier" points that might be carrying a huge amount of weight. Just as a person sitting further out on a see-saw has more impact than one sitting further in, data points that are far away from the center of the cloud carry much stronger weight, and if they happen to lie along one of the positive or negative diagonals, they can produce a strong apparent correlation, even when all the rest of the points in the middle are clearly just milling about and going nowhere in relation to each other.

- **t-test, F-test (ANOVA) and the GLM** The "Student's" t-test is the most basic of the *inferential* statistics used in experimental studies. It is *not* so-named because it is only for use by students, but rather it was the pen-name of the guy who invented it (William Gosset), to improve the quality of beer brewed by Guiness brewery in Ireland, no less! Too bad it isn't called the "Stout" t-test. Anyway, it basically tells you if the difference

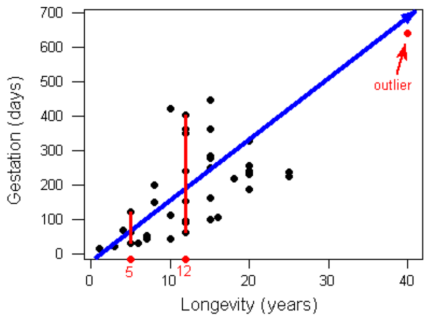


Figure 4: Fig 1-3: Scatterplot showing the positive correlation between length of gestation in the womb and overall lifespan, for different species of animals. The Elephant in the figure is the outlier, carrying undue amount of weight on the overall correlation coefficient. In this case, it is actually consistent with the rest of the data, but sometimes it is not, and yet the correlation still looks positive according to the r value. Thus, it is *essential* to *always* plot your raw data and ensure that the summary statistics are reflective of real aggregate effects!

between your experimental group and your control group is big enough to *not* be due to random chance. Thus, in applying this test, we "reject the null hypothesis" that our data is just random noise, but, critically, we're not actually *proving* that our favored hypothesis is correct. We're just saying it is relatively unlikely to be pure noise. There are more "advanced" versions of this test, specifically the F-test used in the ANOVA (analysis of variance) procedure, and the full *generalized linear model (GLM)*, which can tell you about the importance of multiple different factors and their potential interactions. You may have heard about the *replicability crisis* in various fields of science, including psychology, where many results that were thought to be "true" have "failed to replicate" – meaning that the original paper(s) reported a *significant* t-test result, and the subsequent ones did not (they instead found results consistent with pure noise). This is actually to be expected about 5% of the time, given the standard for publication is set at this 5% level. However, when you take into account how science is *actually* done, there are major systematic biases that enter into the process, which are not taken into account by these statistical tests, such that the actual effective probability of publishing garbage is closer to 50%! There are now important changes afoot to combat the worst of these biases, and help ensure that this garbage probability goes back down to closer to 5%. But 5% itself is still a rather large number – in physics the standard is one in 3.5 million! And, amazingly, results that end up going into the "garbage" pile appear significant at levels below this standard, so randomness can sometimes be a challenging foe.

Chapter 2: Neuroscience

From a materialist, neuroscientific perspective, *everything* that happens in your mind is due to underlying physical processes taking place in your brain. As we discussed in the last chapter, this does *not* mean that we can *reduce* your mind to the brain, but it does mean that there is a really huge mystery here: how is it even remotely possible for a physical system to produce the amazing subjective delights (and terrors, and everything in between) that we all experience?

We start with a time-honored scientific approach: reduce the problem to the simplest possible system that exhibits the relevant behavior, and see if that makes it easier to understand. Consider the two gears as shown in Figure 2.1. As elaborated in the figure caption, there is something kind of "magical" that emerges out of the interaction between the two gears, which cannot be reduced directly to either gear separately. These **emergent** properties depend critically on the relationship and interaction between the two different parts – their relative sizes, rotational speeds, etc. If the larger gear interacted with a different, even larger gear, the overall system of interacting gears would exhibit very different emergent properties. Thus, you really can't isolate these emergent properties to either gear in isolation. Furthermore, the actual material that the gears are made of is largely irrelevant, as long as it is reasonably solid. Thus, there truly is some kind of seemingly mysterious new "substance" being created out of this interaction, which can "transcend" its material basis. And yet, it nevertheless depends entirely and directly on having an actual material basis – the *picture* of those gears doesn't work at all like two actual gears!

This simple two-interacting-gears scenario captures the strange relationship between mind and brain, where the mind depends entirely on the brain, and yet it fully transcends it. As we'll see in a moment, the brain has billions of tiny, interacting parts (*neurons*), which, like the gears, interact in ways that produce emergent properties transcending their material substance. Moreover, there are so many neurons in the brain, and each one interacts with so many *other* neurons (receiving roughly 10,000 inputs and sending a similar number of outputs), that there is a vastly greater degree of emergent interactions taking place in the brain compared to our simple gear example. Thus, although it is essentially impossible for us to wrap our own minds around it, it should be possible to at least imagine in a vague way how something as fantastic and complex as the mind could indeed emerge out of all those billions and billions of interactions taking place every nanosecond, right inside your very own brain.

To try out another metaphor, you can also think about the brain as a massive LEGO set, with parts that *learn* to interconnect with each other in myriad ways. As you might have experienced in your youth, the number of different ways even a small pile of LEGO's can be combined to make different things quickly exhausts the imagination. This *combinatorial explosion* of possibilities is an essential feature of the brain – our neurons can be interconnected in so unimaginably many different ways, that the possibilities are effectively infinite. Due to the explosive nature of combinations, even small numbers of elements can be combined in

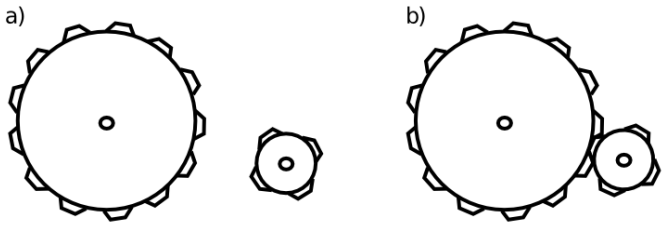


Figure 5: Fig 2-1: The principle of *emergence*, simply illustrated. The gears on the left do not interact, and nothing interesting happens. However, on the right, the interaction between the gears produces interesting, useful phenomena that *cannot* be reduced to the individual gears *separately*. For example, the little gear will spin faster, but the larger one will have higher torque at its axel – these properties would be entirely different if either gear interacted with different sized gear. Furthermore, the material that the gear is made from really doesn't matter very much – the same basic behavior would be produced by plastic, metal, wood, etc. Thus, even in this simple case, there is something just slightly magical and irreducible going on – when two gears get together, something emerges that is more than the sum of the parts, and exists in a way independent of the parts, even while being entirely dependent on actually *having* those parts to make it happen. This seems like a good analogy for the relationship between the mind and the brain.

more different ways than there are atoms in the universe. To see this for yourself, type in `69!` (factorial) on your calculator (or just google it), and you'll get a number that is 1.7... with 98 zeros! This factorial function lies at the heart of combinatorial explosion, and gives a rough sense of the number of different combinations of 69 parts. You can't even begin to conceive of (or even calculate) the value of 100,000,000,000! (i.e., the factorial of the 100 billion neurons in your brain).

Simple Neurons Make Complex Work

The magic of LEGO is that all the different parts interconnect using a single, simple principle, so that you really can make all those different combinations work. The same is true of the brain: each neuron operates according to surprisingly simple, easily-understood principles, and the power emerges through all the interactions / combinations of these simple parts. For full disclosure, not everyone agrees with this perspective, and many scientists have spent a long time exploring more complex kinds of languages that neurons might speak (Reike et al. 1996), but that is a topic for a more advanced textbook. Through the use of computer models of neurons, we can at least say with confidence that the simple ideas presented here can account for a surprisingly large amount of what we know about brain function.

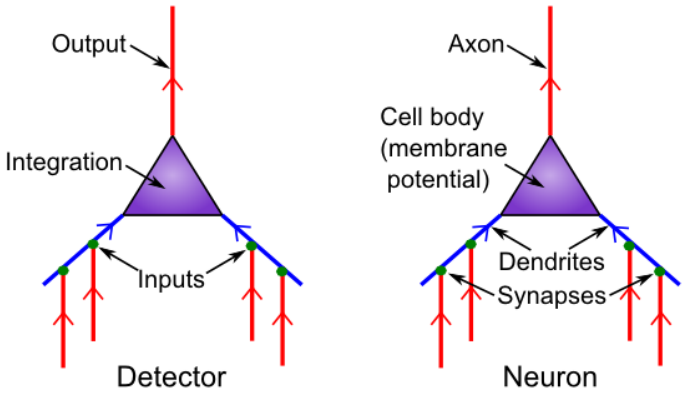


Figure 6: Fig 2-2: The neuron as a detector. Inputs come in via *synapses* connecting the *axons* of other neurons to the *dendrites* of a given neuron. This neuron *integrates* these inputs, resulting in an overall *electrical potential* (called the *membrane potential*, because it is the electrical difference between the inside and outside of the neuron's cell membrane), in the cell body. At the start of the axon (the axon *hillock*), a critical *go / nogo* "decision" is made – if the membrane potential is sufficiently elevated, then the neuron triggers an *action potential* (aka a "spike"), which races up the axon and delivers its signal to the many thousands of other neurons that are "listening" to this signal, via their own synaptic connections. Thus, the essence of neural function is *communciation* – neurons are highly social little things, and our brain is really a huge social network of chattering naybobs.

The easiest way to understand what neurons are doing is in terms of **detection** (Figure 2-2). A neuron acts much like a smoke detector, constantly sampling its local environment (i.e., its inputs from other neurons), and looking for some set of incoming signals that indicate that something *important* might be going on (e.g., a fire in the case of the smoke detector). When it detects whatever it is looking for, it sends a signal out to other neurons, alerting them to the news, so they can incorporate this as one of many other pieces of information that they are sampling in their own detection process. And so on, and so on... Examples of the kinds of things different neurons have been shown to detect include: faces, specific people's faces (e.g., a famous case of a neuron tuned to Halle Barry, and another for Bill Clinton; (Quiroga et al. 2005)), eyes, letters, numbers, houses, different levels of visual depth, specific sounds, etc. Basically, anything that you can be aware of when looking out at the world is the result of neurons detecting those things from among all the possible configurations of visual features, including the words you're reading now, or your laptop, or your phone, or that pizza slice... everything!

Each of the major biological parts of the neuron take on a clear functional role within this overall detector model:

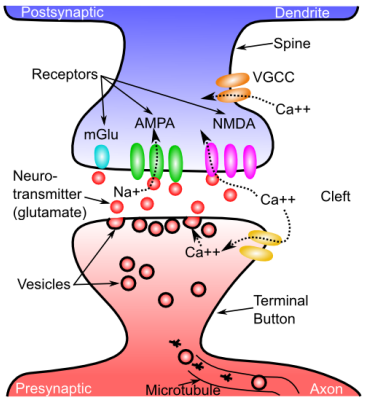


Figure 7: Fig 2-3: Details of the communication process across the synapse between a sending axon and a receiving dendrite. Neurotransmitter is released from the *terminal button* (pronounced French-style by those in the know), and binds to corresponding *receptors* on the dendrites, causing them to un-twist and thus open up small *channels* that allow electrically-charged *ions* to flow into the receiving neuron. Once neurotransmitter is released, it is taken back into the axon (*reuptake*) and is also broken down by enzymes, so that it tracks the rate of spiking by the sending neuron, and doesn't just hang out indefinitely. In addition to the primary excitatory *AMPA* channels that bind *glutamate* and allow *Na+* to enter, glutamate also binds to NMDA and mGlu receptors that are involved in learning, and other synapses use other neurotransmitters such as *GABA* which are inhibitory and allow *Cl-* ions to enter.

- **Synapses**: are the tiny gaps between neurons, where the output signals from one neuron cross over and become the input signals to the next (Figure 2-3). Most synapses are *chemical*, involving the release of a *neurotransmitter* from the *presynaptic* axons, which then bind to *receptors* on the *postsynaptic* dendrites. These receptors twist open as a result of neurotransmitter binding, and allow *ions* (i.e., electrical charge) to flow into the dendrites, through the resulting open channels. The most common neurotransmitter in the neocortex is *glutamate*, and it opens up *AMPA* receptors, that allow sodium (Na+) ions to flow into the dendrites, thus creating a *positive* electrical potential in the receiving neuron. After release, neurotransmitters are taken back into the terminal button (*reuptake*) and broken down by enzymes – this ensures that the amount of neurotransmitter binding accurately reflects the spiking rate of the sending neuron. All this complex-sounding machinery and terminology is actually very simple: Neurons like to excite other neurons by sending them exciting signals! The basic machinery is chemical and electrical, but the bottom line is just: how strongly do the input signals to a given neuron excite it? This is determined by the detailed function of each of the roughly 10,000 synapses coming into a given neuron. This point bears emphasis, as we will return to it repeatedly: **the pattern of its synaptic connections determines what a given neuron detects, and thus, ultimately, what the brain knows**.

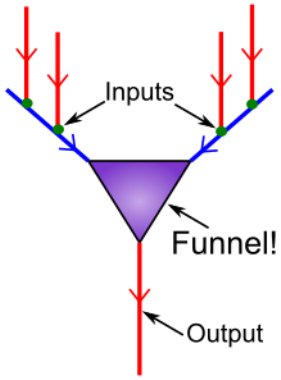


Figure 8: Fig 2-4: The neuron as a *funnel*, *compressing* its 10,000-odd inputs down into a *single* output signal, conveyed through its axonal output. This is the genesis of the *Compression* principle of brain function.

- **Dendrites**: provide a broad tree-like (dendrite literally means tree-like) "arbor" for all the synaptic inputs into a neuron, and they funnel the resulting electrical charges up into the cell body. This funnel-like property, illustrated in Figure 2-4, is the origin of the **compression** principle of brain function, one of our three C's. As we noted in the introduction, each neuron is compressing its 10,000 different inputs into a single output signal,

producing a roughly 10,000-to-1 compression factor. As an aside, one of the big debates in neuroscience is the extent to which these dendrites perform various kinds of more complex "processing" of their synaptic inputs, or simply convey the overall signal. There is evidence on both sides, and, as usual, the truth is likely somewhere in between.

- **Cell Body**: The neuron is a cell, and, despite its long tendrils, it has a cell body like other more compact kinds of cells, where the nucleus and other cellular machinery hangs out. It is here that all the dendritic signals converge, to produce the final compressed electrical potential that somehow summarizes everything coming into the cell at that moment. This electrical potential is called the *membrane potential* because the electrical signal is measured as a difference in electrical potential across the cell's membrane (that fatty lipid bilayer that you might recall reading about in high school science). If this membrane potential is sufficiently excited, then special channels at the start of the axon (the axon *hillock*) will get *extra* excited and essentially flip a switch, causing the initiation of the *action potential* or *spike*. The details of this process were worked out by Hodgkin and Huxley in the 1950's (Hodgkin and Huxley 1952), and have stood the test of time, forming the basis of modern detailed mathematical models of neuron firing.
- **Axon**: The spike propagates down the axon, effectively broadcasting this one signal out to the roughly 10,000 other neurons that it sends input to, continuing the great chain of communciation among neurons. Axons can have varying amounts of *myelin*, provided by helpful *glia* cells called *oligodendrocytes* (no you won't be tested on those!), which serve to insulate the electrical "wire" that is the axon. Myelinated axons convey information more quickly, and *multiple sclerosis* is one of various disorders that involves the degeneration of this myelin, resulting in slowed signal conduction. There are many other forms of glia cells, but all of them are generally thought to play various supporting roles in the overall function of the brain, whereas the neurons are the "stars" of the show. These supporting roles are essential for keeping the brain functioning, and may affect various processes such as learning, but we'll nevertheless generally ignore them in this introductory treatment.

To summarize, each neuron is receiving a huge amount of input through its roughly 10,000 synapses, and it then compresses this all down into a single discrete spiking signal that it then broadcasts back out to the roughly 10,000 other neurons listening to its little story. Only when a neuron detects something "interesting" does it get excited enough to send this spiky signal out, and this *thresholding* is really the defining characteristic of a detector, making it respond *selectively*. Thresholding is just as important in neurons as it is in people: it can quickly get tiresome listening to someone with a *low threshold* who is always blabbering on about the most uninteresting things. The advent of Facebook and other forms of social media has greatly magnified this problem.

We'll explore more about the kinds of interesting conversations neurons might be having in a bit, but first we'll examine how this electrical magic operates

within the neuron in more detail. These details are typically not presented at this introductory level, but a really simple analogy helps make it accessible, and this machinery ends up producing the **contrast** effects that are so central to our overall framework, so we're motivated to take this brief detour.

The Tug-of-War in Your Brain

There are two major classes of synaptic inputs converging on each neuron: the excitatory ones described above (via the neurotransmitter glutamate opening AMPA receptors), and separate *inhibitory* synaptic inputs that are driven by a neurotransmitter called *GABA* which activates... *GABA* receptors, which allow negatively-charged *Cl-* (chloride) ions to enter the cell. Thus, your brain runs primarily on table salt: *NaCl*. These inhibitory inputs come from an entirely separate set of specialized neurons known as *inhibitory interneurons*, which are somewhere between the principal, excitatory neurons and the glia in overall status within the pecking-order of the brain. These interneurons only act relatively locally, like glia, and they also play a largely *regulatory* role, regulating the overall level of electrical excitation surging through the brain. In contrast, the main excitatory *pyramidal* neurons (which constitute roughly 85% of neurons in the neocortex) can broadcast their exciting messages over long distances to far-flung regions of the brain, and are regarded as the primary *information processing* neurons (i.e., they are primarily responsible for all the chatter and compression going on).

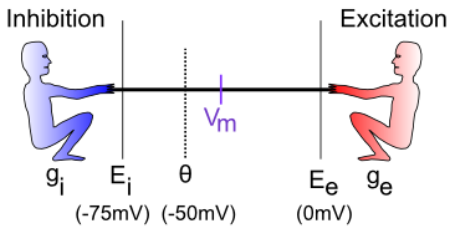


Figure 9: Fig 2-5: The Tug-of-War between excitation and inhibition, producing a beneficial balancing act, and the source of **contrast** coding in the brain. Inhibition pulls the membrane potential (written as Vm, where V=voltage and m=membrane), down toward the *resting potential* of roughly -75 mV via the influx of negative Cl- ions. Excitation pulls up toward roughly 0 to +55 mV (depending on type of neuron, etc) via the influx of positive Na+ ions. Thus, the components of ordinary table salt (*NaCl*) are driving this perpetual battle inside every one of your neurons. Theta represents the *threshold* electrical potential, above which the neuron will fire a spike. The ability to do so depends only on the *relative* balance between excitation and inhibition, not the absolute levels.

Inside each neuron, excitation and inhibition are forever locked in a pitched battle, which can be pictured as a tug-of-war, with each side pulling with varying

strength, but each side always pulling in the same direction (Figure 2-5). The "pitch" on which this battle is taking place is the amount of electrical charge in the cell, i.e., the membrane potential. The excitatory end is always pulling this potential upwards, while the inhibitory side is pulling it back down, and the actual potential represents the balance between these two forces.

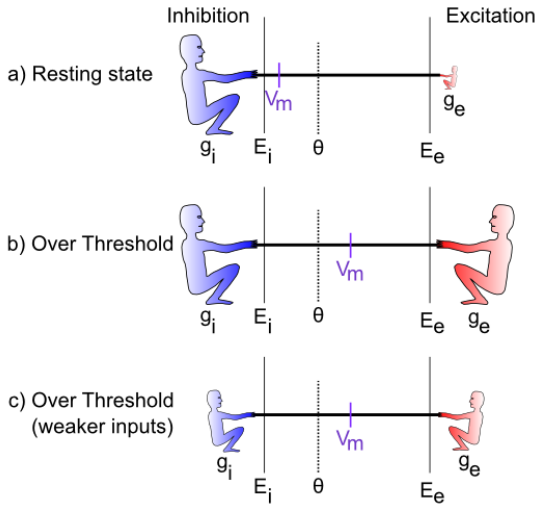


Figure 10: Fig 2-6: Illustration of the contrast or relative nature of the tug-of-war: only the relative strengths of excitation and inhibition matter, not their absolute values. Thus, neurons respond to the contrast between their inputs and overall levels of inhibition, which typically represent the the rough average of activity coming into a given brain area.

The essential point here is that **contrast**, or **relativity** emerges as the result of this tug-of-war battle. Specifically, it doesn't matter how strong the two different sides on the tug-of-war are in absolute terms – all that matters is the *relative* strength of the two sides (Figure 2-6). Excitation could be relatively weak, but if inhibition is also week, then the net balance between the two will be the same as if each was proportionally stronger.

Typically, the amount of inhibition is roughly proportional to the "average" amount of activity in the brain in any given area, so in effect, each neuron is effectively comparing how excited it is against this overall "average" level. Only those neurons that are getting *above average* level of excitation will actually get excited enough to fire spikes.

In real-world terms, this "average" inhibition is very much like the amount of money that your peers are making (or the amount of fun they appear to be having on their various social media accounts) – it forms the baseline or standard against which you measure yourself. Likewise, neurons are constantly comparing themselves against *their* peers, and all of the spiking going on in your brain is

therefore always and inexorably *relative* to these peer-standards. For example, when you step outside into the bright sunlight, all the visual neurons suddenly get a huge wave of excitation relative to the dim indoor light from before. But you avoid suffering an epileptic seizure from all that excitement because those inhibitory interneurons are also getting this wave of excitation, causing them to send a proportional amount of damping inhibition on the party, keeping everyone in balance and on a more level keel. Yes, inhibition is the wet rag of the brain, but without it, you really would be suffering from seizures all the time.

In fact, this balancing act between excitation and inhibition is so important for overall brain function, that our brains are perched on a kind of "knife edge", and the relatively high incidence of epilepsy in the population is likely a result of the fact that it is really hard to get this balance exactly right. And too much inhibition has very bad consequences as well (indeed, it literally "depresses" your brain and makes it difficult for you to do anything). Furthermore, the main treatments for epilepsy involve activating the GABA inhibitory system more strongly, thus altering this fundamental tug-of-war balance.

In summary, two out of the three of the core principles of this textbook, **compression** and **contrast**, emerge directly out of the basic function of neurons. As we discussed in the Introduction, we can trace the implications of these core neural properties all the way through the full scope of Psychology and behavior. The Perception chapter will provide particularly compelling demonstrations of how compression and contrast play out in our perceptual lives – the story of perception is fundamentally the story of compression and contrast.

By thinking in these terms, we have managed to dramatically simplfy our understanding of the brain, creating an almost transparent, level-spanning way of going from single neurons on up. However, none of this contradicts the emergence and complexity discussed at the outset. Instead, these principles just capture the overall general tendencies and propensities of the brain, but within that broader scope, there is a wild, complex, bubbling jungle of intertwined conversations and chatter constantly unfolding within your brain, thinking all manner of complex and ineffable thoughts. Next, we'll move up from the level of individual neurons and start to think about how all these principles might play out in terms of how different brain areas are organized to facilitate effective overall behavior and cognition.

Large-Scale Brain Organization ("Gross" Anatomy)

We will attempt to answer two closely interrelated questions about the large-scale organization of the brain, one of which is relatively easy, and the other which remains rather more murky. The easy question is: "what are the obviously separate parts of the brain, which have a distinct evolutionary and structural basis?" The hard question is: "to what extent do any of these brain parts, or regions within these brain parts, support a distinct kind of overall function?" This latter question of *functional specialization* is challenging because neurons are so massively interconnected and interdependent on each other, that it is hard

to clearly isolate any specific function. It is like any kind of team sport: we are tempted to focus on a few specific star players, but, really, the team depends on every player and the quality of their interactions (just like the gears in Figure 2-1, and any emergent system). For example, the quarterback on a football team can either look really good or bad as a function of how good the offensive line is, but nobody gives that line sufficient credit, focusing instead on the singular, more glamorous quarterback.

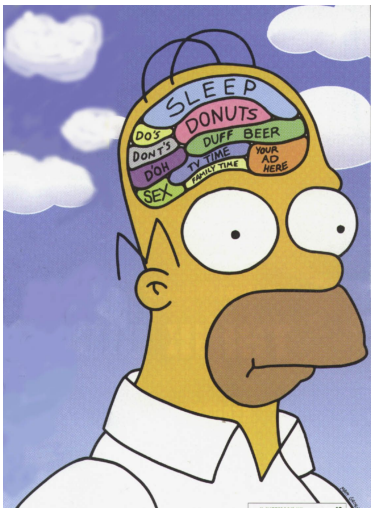


Figure 11: Fig 2-7: The brain would be a lot easier to understand if each part had an easily-labeled, distinct function.

Historically, Karl Lashley in the 1920's concluded from his extensive studies lesioning different parts of rat brains, that the brain is an *equipotential* system, operating according to some kind of *mass action* principle. That is, all areas contribute roughly equally to the overall function, and all that matters is how much overall neural tissue is intact – the more the better. This idea strongly conflicts with everything we know about mechanical systems, where each part has a specific, well-defined function, and with the overall force of *compression* in our brains: we want a simple, easy-to-understand picture about how things work. This is reflected in the ubiquitous drawings of the brain carved up into discrete functions (e.g., Figure 2-7), and in the discredited approach of *phrenology* which attempted to associate functions with different bumps on the skull. Critics have derided many recent neuroimaging studies as *neo-phrenology* because there is still this strong tendency to ascribe discrete functions to individual blobs of brain that "light up" when people are doing different tasks in the brain scanner.

As usual, the truth is somewhere in between these extremes. Even though neurons are massively interconnected and interdependent, and overall function emerges through these interactions, there is evidence that different brain areas are differentially important for different functions. But the level of functional

specialization is much more *partial* and *overlapping* than completely distinct. One way of thinking about this is in terms of the saying that "All politics is local", which is as true for the brain as it is for people, perhaps more so: neurons can't pack-up and move to a different part of the brain. Instead, they are like the 85% of Pittsburghers who live their entire lives in the same little neighborhoods. This means that different neighborhoods can develop their own special "personalities" and focus on detecting particular kinds of signals. On the other hand, the excitatory neurons in the brain also send out long-range connections. Network theorists characterize these as "small world" patterns of connectivity, such that, in the end, every neuron is only a few synapses away from every other neuron. This then limits how much "neighborhood funkiness" can develop. So, again, the brain is all shades of grey, not black-and-white: different areas are somewhat specialized, but also very interdependent.

The Big Brain Chunks

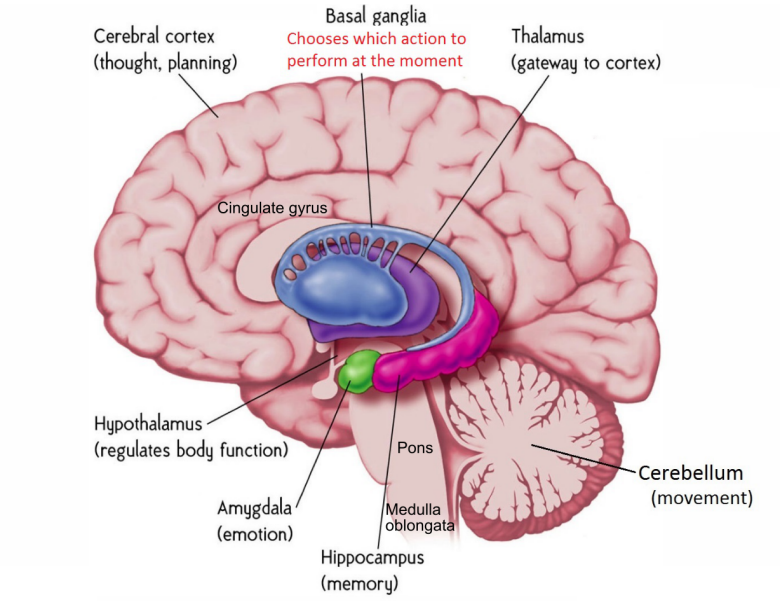


Figure 12: Fig 2-8: Large-scale ("gross") brain structures and their overall specialized functions.

Figure 2-8 shows the different brain regions that can be easily distinguished based on evolutionary history and obvious structural differences. Keeping in mind the above qualifications, the overall functions of these areas are as follows:

Cerebral cortex (Neocortex): This is the most important part of the human brain, supporting all of our special human abilities to think, read, talk, reason, plan, etc. We are exclusively conscious of activity in this part of the brain

– this is where "you" live! Most of what we discuss throughout the textbook is focused on this part of the brain, and all of what we said above about neurons is focused specifically on this part of the brain (other parts may differ in various details, but the general properties are common across the brain). Often, we'll just refer to this as "the cortex", although technically "cortex" means "sheet-like" and other brain areas also have a "cortical" kind of organization. Anatomically, this sheet-like nature of the neocortex is more evident in smaller-brained animals – in humans, the neocortex is so greatly expanded that it is all folded over on itself. It is the wrinkled sheet upon which all of our hopes and dreams rest. We'll go into more detail about the different lobes and their relative functions in the next section.

Thalamus: Functionally, the thalamus is completely intertwined with the neocortex, and neither can function without the other. There are massive, bidirectional interconnections between every part of the neocortex and the thalamus. Some parts of the thalamus "relay" information from the senses up into the primary sensory areas of the neocortex. For example, the *lateral geniculate nucleus (LGN)* receives most of the output from your eyes, and then sends that up into your *primary visual area (V1)* in the neocortex. However, it is not *just* a relay: V1 also sends massive "top-down" connections back into the LGN, and these serve to focus attention and organize all the low-level visual signals into a more coherent overall "picture". As visual information processing proceeds up to higher levels in the cortex, the thalamus continues to play a critical role through a structure known as the *pulvinar*, which again has massive bidirectional interconnections with corresponding cortical areas.

The pulvinar has been implicated in attention, and also serves to coordinate different cortical areas by synchronizing brain activity in the *alpha* frequency (10 Hz or 10 cycles per second) – we'll discuss these frequencies more later when we talk about sleep stages (the thalamus also plays an important role in sleep). Other areas of the thalamus are directly interconnected with both the frontal lobes and the basal ganglia, and are really inseparable from the overall function of those brain areas. Thus, overall, despite its anatomical separation, functionally it does not really make sense to think of the thalamus as a separate brain area from the neocortex – instead we should think of the *thalamocortical system* as a functional unit.

Basal Ganglia (Striatum): This is a collection of different brain *nuclei* (chunks of neurons) that form a complete sequential pathway or loop from the neocortex and back up into different parts of the frontal lobes. Thus, like the thalamus (which is a key part of this loop), it is hard to really separate the function of the basal ganglia from that of the frontal lobes of the neocortex, and damage to either of these areas produces very similar overall problems. Indeed, this *fronto-striatal* system is implicated in most of the major mental disorders that afflict us, including depression, anxiety disorders, ADHD, and OCD, as we'll cover in detail in the chapter on mental disorders. As noted in the Introduction, this system is the most important player in the **control** component of our three C's, and these disorders are all fundamentally disorders of control.

Anatomically, the input portion of the basal ganglia circuit is composed of the *Caudate Nucleus*, *Putamen*, and the *Nucleus Accumbens*, which collectively comprise the *Striatum*, which means "striped". The striatum receives massive input from all over the neocortex, "digests" it down into a basic "go" vs. "nogo" decision about whether to do something or not, and then sends that decision back up into the frontal lobes, by way of the thalamus. Thus, whereas the basal ganglia was previously thought to be more of a "habit learning" part of the brain, it is actually the real *decision maker* in your brain. Indeed, research shows that the basal ganglia makes a decision about what you're going to do next about 1/3 of a second before you are consciously aware of it! This reflects the fact that we're only conscious of what is going on in the neocortex, and some people find it kind of unsettling that this "other" part of your brain is "making decisions on your behalf". But really, you are *all* of your brain, not just the parts you're subjectively aware of, and again, all of these areas are massively interconnected and interdependent, so don't get too freaked out by this! Embrace your inner decision maker, which is responsible for those "gut feelings" that all so often end up being correct, even as they are often overridden by your over-analyzing conscious cortex.

The basal ganglia are unique in the brain by virtue of having by far the most *dopamine* receptors, and dopamine plays a critical role here by shaping the decision-making process according to what has worked, and not worked, in the past. Furthermore, the basal ganglia, particularly the nucleus accumbens in the *ventral* (bottom) part of the striatum, plays an essential role in controlling the firing of dopamine neurons, so that the overall dopamine signal reflects the *contrast* from your expectations, rather than raw reward or punishment itself. In the Learning chapter, we'll see in more detail how this process works, in the context of *Classical Conditioning* and *Operant Conditioning*, which largely reflect the dopamine-driven learning processes taking place in the basal ganglia.

Amygdala: The amygdala is a relatively small nucleus, which is named after the Greek word for almond (most antomical labels describe either the shape, color, or texture of the brain structure), that plays an essential role in driving our emotional life. It is extensively interconnected with both the basal ganglia and the dopamine system, and drives these systems to respond appropriately for positive and negative emotional events. For example, when a previously-neutral stimulus is associated with either a rewarding or punishing outcome in classical conditioning, the amygdala learns the association between the stimulus and this outcome, and drives dopamine firing and other behviors to anticipate and prepare for the outcome (e.g., approaching yummy food and running away from fear-inducing scary stuff). The Amygdala is also extensively bidirectionally interconnected with the neocortex, receiving sensory inputs and sending its emotional signals up to the medial and ventral regions of the frontal lobes, which are the emotion centers of your conscious world in the cortex. Thus, overall, the amygdala is a *hub* for emotional signals, interconnecting between lower-level brain stem systems such as the hypothalamus, and driving your high-level conscious emotional experiences. People with damage to this area

don't necessarily have a complete absence of emotion, but they can't connect all the pieces together in an effective way, and often behave carelessly because they fail to anticipate the potential risks of their actions.

Hippocampus: The hippocampus lives next door to the amygdala, and is essential for rapidly forming new memories of the daily events of your life (i.e., *episodic* memories). When you think of memory, mostly you're thinking of what the hippocampus does. Of all the brain areas we've considered so far, the hippocampus is the most strikingly specialized: highly selective damage to this brain structure can result in profound *amnesia* – particularly the inability to learn new episodic memories, but also the loss of at least a certain window of more recently-acquired memories. The famous patient H.M. (Henry Molaison) had his hippocampus lesioned surgically to alleviate epileptic seizures, and was unable to acquire new memories for the rest of his life. Along with our emotions, our memories are the most cherished aspect of our subjective world, and it is truly horrifying to imagine losing this ability. Therefore, you should treat your hippocampus well: it is a bit of a "canary in the coal mine", and is often the first thing to go when you lose oxygen to the brain. Likewise, heavy drinking causes this area to lose function before others, resulting in memory blackouts.

As we'll explore in greater depth in the Memory chapter, the hippocampus has several biological specializations that enable its super-memorizer abilities, but these also result in it being more sensitive. Although the hippocampus is highly specialized, it nevertheless depends entirely on extensive input from the surrounding areas of the neocortex, which convey a massively *compressed* summary of everything going on in the rest of your brain. This high level of compression makes our memories relatively inaccurate and subject to many biases, but also extremely efficient. The hippocampus then essentially takes a "snapshot" of the current state of the brain, and later, when you want to recall some prior event, it can retrieve that snapshot and cause the rest of your brain to relive that moment. During recall, the hippocampus drives those same surrounding neocortical areas in the reverse direction from when the memory was initially encoded, again demonstrating the essential interdependence of all these different brain areas. Also, the hippocampus has somewhat separable "cognitive" and "emotional" components, with the emotional one extensively interconnected with the amygdala and those frontal emotional areas that strongly interconnect with the amygdala, and it plays a critical role in making your emotional responses appropriately responsive to different situations and contexts.

Cerebellum: The cerebellum plays a critical role in learning to perform motor (muscle) movements in a smooth, efficient, and coordinated way. Anatomically, it is a kind of "mini brain" (that is what it's name means) tucked under the back of your brain, and it is also a "cortical" structure with a very distinctive sheet-like organization. In some ways, you can think of it as a kind of "hippocampus for motor learning", as suggested by the pioneering scientist David Marr in a pair of prescient papers (Marr 1969; Marr 1971) that attempted to discern the functions of both the cerebellum and hippocampus based on their unique anatomical properties. Amazingly, his ideas have largely stood the test

Area	Learning Signal: Reward	Error	Self Org
Primitive			
Basal Ganglia	+++	- - -	- - -
Cerebellum	- - -	+++	- - -
Advanced			
Hippocampus	+	+	+++
Neocortex	++	+++	++

+ = has to some extent … +++ = defining characteristic – definitely has
- = not likely to have … - - - = definitely does not have

Figure 13: Fig 2-9: Learning Rules across the brain: some of the clearest differences between brain areas are in terms of the signals that drive learning in a given area. In particular, the basal ganglia and cerebellum each specialize on two of the most important types of learning signals: reward (and punishment) vs. error signals (which are *not* the same as punishment – these are instead detailed signals with specific information about exactly what didn't go according to plan in a motor action. The cerebellum is unique in having no dopamine innervation or receptors. More evolutionarily-modern areas incorporate multiple signals, and include self-organizing learning, which means learning that happens automatically all the time, as in the hippocampus automatically taking snapshots of cortical activity.

of time, and form the core of our modern conception of these areas. These brain areas are among the most functionally specialized, and both rely on a kind of "brute force" memorization strategy to achieve their special learning abilities. This brute force strategy requires a lot of neurons: half of the total neurons in your brain live in the cerebellum! An important consequence of this strategy is that it takes lots and lots of practice to really perfect any given motor skill (e.g., gymnastics, skiing, etc), because the cerebellum has to memorize each of the many different ways to perform a motor action.

The cerebellum learns to anticipate errors, awkwardness, and inefficiency in a given motor action plan, and sends well-timed corrective signals to prevent those from actually occurring. It receives error signals from a nucleus called the *inferior olivary nucleus* (you can guess what it looks like), which drive a powerful error-correcting learning signal in the *purkinje* neurons that are one of the central actors in the cerebellar circuit. These purkinje neurons are truly amazing things, receiving over 100,000 different synaptic inputs (10 times as many as the typical neocortical neuron) – so many synapses are needed to be able to have distinct memories for each of those different motor action sequences. This form of error-driven learning is quite different from the "snapshot" memorization operating in the hippocampus, so these brain structures are also functionally distinct from each other, even though they both share the same overall brute-force memorization strategy.

Interestingly, the cerebellum and basal ganglia, which are both considered

motor control systems, have almost no direct interconnections (a rarity in the brain, as we've seen) – but this actually makes good sense, because they each perform very different functions, at different time scales (Figure 2-9). The cerebellum deals with very fast "online" motor control at the scale of 10's of milliseconds, whereas the basal ganglia is more involved in the "outer loop" of deciding which of various possible motor plans to actually execute. Thus, the basal ganglia typically acts first to select the motor plan, and then the cerebellum takes over and ensures that the selected plan is executed to the best of your ability. By analogy with the different roles in making a movie, the basal ganglia (together with the frontal cortex) is the producer, deciding what movie to make; the cerebellum is the director, who is there day-in-day-out on the set, dishing out detailed instructions to the actors to make it all look good; and motor circuits in the *pons* and other brainstem areas, on down into the spinal chord and the muscles, are the actors, actually carrying out the actions.

Hypothalamus: This tiny structure plays a huge role in controlling your basic bodily functions, including eating, drinking, sleeping, arousal, sex, stress, immune response, etc. It is the kingpin in the *HPA axis* (hypothalamic-pituitary-adrenal), which is a system of interconnected structures that release hormones including corticosteroids in response to stress. The hypothalamus has many different nuclei, each specialized for different domains, and some of these project up to the amygdala to drive emotional responses. For example, the positive reward feelings associated with eating and drinking come from the lateral hypothalamus, and these signals go into the amygdala and directly into the dopamine system, driving bursts of dopamine for (unexpected) positive events, like when a co-worker brings in leftover birthday cake to the office. The hypothalamus also receives top-down control signals from areas of ventral and medial frontal cortex, which can regulate the response to potentially stressful events, for example.

One fascinating line of research shows that rats that receive mild electric shocks, which they can turn off by moving to another part of their cage, are able to control their stress responses much better than a poor "yoked" rat that receives the exact same electric shocks, but has no control over them. Thus, the perception of control, which has been localized to those frontal cortical areas (consistent with the overall role of these areas in control more generally), is an essential factor in how the body responds to stressful situations (Steven F Maier and Watkins 2010). A clear real-world example of this is the difference between driving a car and riding along as a passenger – the driver typically experiences things as "under control" whereas the passenger is more likely to feel stress because the driver is going too fast otherwise being unsafe. More generally, chronic exposure to negative, stressful sitations over which a person has little perceived control can produce significant long-term mental heatlh problems, leading to a kind of *learned helplessness* that is associated with depression (Steven F. Maier and Seligman 1976).

Brainstem Nuclei and Medulla Oblongata: Finally, there are a number of different clumps of neurons in the brainstem that play critical roles in overall brain and body function. These are evolutionarily more ancient brain ar-

eas, like the hypothalamus, which have highly specialized functions. In computer terms, these are the core BIOS brain areas – the low-level hardware control areas. One group of such nuclei are collectively referred to as the *reticular activating system*, and include the sources of the major *neuromodulators* that "modulate" (alter) the functioning of neurons throughout the brain in various (often similar) ways:

- *ventral tegmental area* and *substantia nigra pars compacta*: dopamine – modulates learning in basal ganglia, other areas.
- *raphe nucleus (dorsal, median)*: serotonin – modulates arousal, sleep, mood.
- *locus coeruleus*: norepinepherine (noradrenaline) – modulates effort, engagement.
- *basal forebrain cholinergic nuclei*: acetylcholine (ACh) – modulates attention, arousal, learning (nicotine affects this system).

These core areas serve as master control knobs for the overall state of the brain, and are thus incredibly important and powerful. All of them receive extensive top-down projections from the frontal lobe, which thereby asserts its overall master control of these knobs, while also being subject to their effects. The mutual interdependence of all of these brain systems is evident even here at the lowest levels, and has many important implications for sleep, arousal, and other overall brain states.

Last but not least, the medulla obolongata wins the prize for the funniest name in the brain, but it is no laughing matter, providing essential low-level body control signals. Damage to this area often results in death. Enough said.

Functional Organization of the Neocortex

The neocortex is divided anatomically into four separate lobes (Figure 2-10), which can be given broad overall functional specializations that stem principally from the unique sensory inputs / motor output coming into / out of each lobe (each lobe gets one of the three major sensory input modalities, or drives motor output):

- **Occipital**: Receives the primary visual input from the LGN of the thalamus (in area V1 at the very back of the brain), and begins the processing of these inputs.
- **Temporal**: Extracts object identity information (e.g., face, pizza, laptop, etc) from visual signals coming in from the occipital lobe, and connects those with auditory signals arising from primary auditory cortex (A1), which is in the upper (superior) portion of the temporal lobe. These connections form the initial basis for *language*, in terms of the ability to name objects recognized visually, and semantically understand the meaning of spoken words. The inner (medial) part of the temporal lobe connects up with the hippocampus, and is critical for assembling the *who-what-where* elements that define the episodes (events) of our lives, that

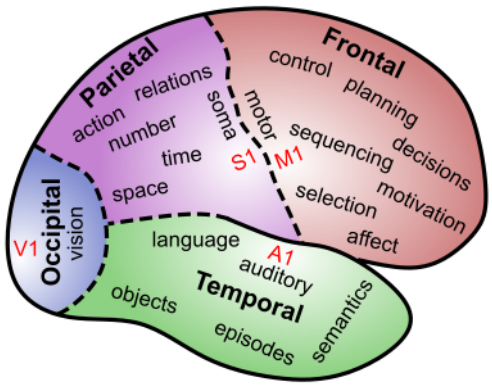


Figure 14: Fig 2-10: Functional specializations of the different lobes of the neocortex, stemming largely from their unique primary sensory / motor areas (V1 = primary visual cortex, A1 = primary auditory cortex, S1 = primary somatosensory cortex, M1 = primary motor cortex).

the hippocampus takes snapshots of. The very tip of the temporal lobe (towards the front) is important for encoding our most abstract, high-level semantic knowledge (truth, justice, etc) (Lambon-Ralph et al. 2017).

- **Parietal**: Also feeds off of the occipital visual information, but in the service of guiding motor actions, by virtue of its position betwixt the occipital and frontal lobes, and the primary somatosensory inputs in area *S1*. S1 is located just across the *central sulcus* (sulcus = groove) from the primary motor cortex, *M1* in the frontal lobe, and each has a matched *homunculus* ("little man") representation of your entire body (Figure 2-11), which is distorted in its focus on the most important areas at the expense of others (e.g., your back doesn't get a lot of neural space, whereas your fingers and mouth are very prominently represented). Because motor actions require proper positioning of your hands and body in space, the parietal lobe is where your understanding of spatial locations and relationships arises. Interestingly, these spatial representations are re-used for thinking about more abstract continuous quantities like time and number.
- **Frontal**: Is grounded by its M1 primary motor outputs, which make this lobe focused on motor control across all levels of space and time. Progressively more frontal (i.e., *prefrontal*) areas encode progressively higher-level, extended action plans, to coordinate and organize the basic motor actions encoded back in M1. These higher levels of control require things like sequencing, planning, and decision-making, and as noted above, all of these functions depend critically on interactions between frontal cortex and the basal ganglia. Overall, we think of the frontal cortex's part of this interaction in terms of generating possible action plans, which the basal ganglia then evaluates according to its dopamine-driven learning

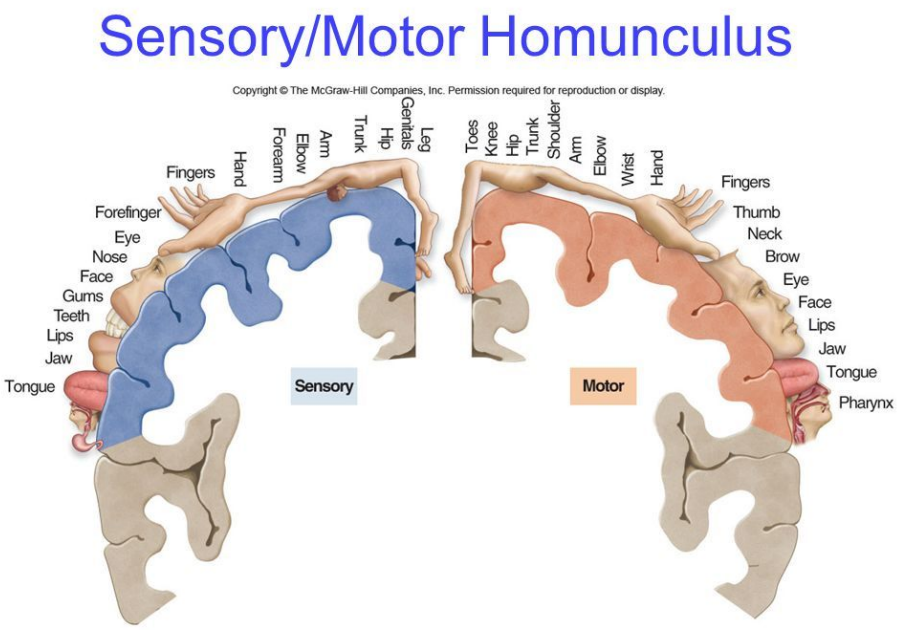


Figure 15: Fig 2-11: The coordinated *homunculus* ("little man") as represented across the primary somatosensory (S1) and motor (M1) cortex.

history, and it sends back up a strong signal activating the plan most likely to maximize future reward and minimize punishment / cost. The inner (medial) and lower (ventral) parts of the frontal lobe are anchored by the inputs from the amygdala and other core visceral areas, including primary taste areas in the *insula*. These emotional (*affective*) and overall body-state inputs are essential for guiding the overall motor control and planning processes, to focus on the things that actually matter, thus giving the frontal lobes a primary role in motivation.

Although the frontal lobe plays such an important role in control, it does not act alone. It is most densely interconnected with the parietal lobe, which provides the sensory guidance needed to inform the action planning process. For example, one important function the parietal lobe can provide is a spatial map of a sequence of actions to be taken over time, to help in figuring out the best ordering of the indvidual steps in the sequence. Furthermore, the parietal lobe can represent the likely sensory outcomes of different possible action plans, in terms of both somatosensory and visual modalities (i.e., how my arm would feel if I moved it in a particular way, and where it would end up in space), which can then feed back to refine the overall motor plan.

The lower (ventral) parts of the frontal lobe are more strongly interconnected with the temporal lobe, and these pathways can enable the frontal control system to shape and regulate all the processes taking place there, including driving what we might want to talk about, and how the hippocampus encodes memories. Furthermore, while overt action plans are most informed by the parietal spatial representations, the emotional and motivational aspects of control and planning are more directly informed by information from the temporal lobe. We typically care more about *who* and *what* when thinking about our motivational goals and emotional states, rather than the *where* information encoded in the parietal lobe. Thus, it makes sense that these temporal-lobe inputs to frontal cortex also converge on the ventral and medial affective / motivational areas.

Thus, overall, we can see in the functional organization of the neocortex this balance between different neighborhoods of neurons specializing on different kinds of information, but also depending critically on the work of other areas to get their own jobs done. In effect, the brain is just like any complex human organization (e.g., in a company, a university, the military, etc) – everyone depends to varying extents on the work that others are doing, but each person also performs some specific, specialized roles. Because neurons stay put in the brain, and neighboring neurons tend to be more strongly interconnected with each other, we can trace these networks of interaction and interdependency to help understand what each part is doing. Next, we'll briefly examine the importance of another aspect of complex organizations: a hierarchical structure.

Hierarchical Organization

Figure 2-12 shows how this combination of interdependency and specialization plays out in the case of the visual pathway going from V1 up through the object

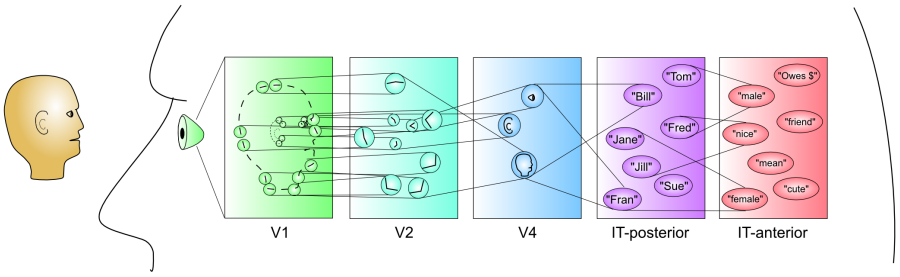


Figure 16: Fig 2-12: Hierarchical organization of detectors in the visual pathway going into the temporal lobe, supporting the ability to recognize (detect) entire objects, based on earlier levels detecting parts and features of parts. This shows the large-scale, cumulative effects of *compression* from very high-dimensional raw sensory inputs, to high-level, succinct interpretations of the world. Although a highly simplified cartoon, this roughly captures the nature of the process actually taking place in the brain.

recognition neurons in the inferior (bottom) part of the temporal lobe (IT = *inferotemporal cortex*, conveniently where *it* is recognized). There is an overall *hierarchical* organization to this pathway, such that the early stages detect simpler features (e.g., oriented edges in V1) while higher levels build on this to detect parts of objects in terms of collections of these features, and still-higher levels can then detect entire objects in terms of collections of features. Thus, by building up a cascade or hierarchy of detectors in this way, each performing their own part in a larger chain of *compression*, a very challenging overall problem can be broken down into simpler steps. Hierarchical organizations of this sort are ubiquitous and necessary for organizing, coordinating, and integrating the work of many individuals, whether it is people in the military or corporations, or neurons in the brain.

Neuromodulators and Drugs

Given their sociological, psychological and medical importance, we devote some time here to understanding the way that various well-known drugs affect the brain. We already saw above that there are specific nuclei in the brain stem *reticular activating system* that release *neuromodulators* that have broad overall effects on wide areas of the brain. Perhaps not surprisingly, these are the major targets of psychoactive drugs, because they have such major modulatory effects on the brain. By contrast, *glutamate*, which is the main *neurotransmitter* in the strict sense of transmitting detailed signals from one neuron to the next (in the neocortex and many other areas), is not directly affected by most drugs, and its effects are much more local and content-specific. Even GABA, the primary inhibitory neurotransmitter, can be considered more of a neuromodulator in

that it has broader regulatory effects and is directly affected by psychoactive drugs. To be clear, this difference between neurotransmitter and neuromodulator is strictly a functional distinction – they are all just chemicals released by the axons of neurons, but it is useful to distinguish the transmission vs. modulation roles.

Drugs can affect the brain in two opposing ways:

- **Agonists** are drugs that mimic or amplify the effect of a given neuromodulator. This term is a bit "agonizing" because it doesn't exactly sound like what it means, but you can perhaps remember it better in relation to its opposite (antagonist). Scientists typically reserve the term *agonist* to more precisely refer to chemicals that specifically bind to the same receptors as the endogenous neuromodulator, but we'll adopt a looser definition that includes anything that has a net "positive" effect on the effect of the neuromodulator. For example, *Valium* and other *benzodiazepines* are direct GABA agonists by binding to the GABA receptor and enhancing the amount of *Cl-* that enters the cell, whereas *Ritalin* (*Methylphenidate*) enhances dopamine effects by inhibiting the reuptake of dopamine after it is released, so it is a kind of agonist but acts more indirectly. There are many different biochemical mechanisms that can lead to a net agonist effect.
- **Antagonists** are drugs that suppress, inhibit, or otherwise work against a given neuromodulator. They "antagonize" that poor neuromodulator. *Curare* poison is a classic competitive antagonist for acetylcholine (ACh) at the synapses of nerve fibers onto muscles, thus acting to paralyze muscles. It acts by binding directly to the same receptors that ACh normally binds to (and it does so more effectively, i.e., with greater *affinity*), but it does *not* actually open those receptor channels. *Botulinum toxin* (*botox*) is also an overall ACh antagonist, but it works by preventing the release of ACh.

Interestingly, the neuromodulators are biologically ancient chemicals that have very different effects throughout the body, which explains why drugs often have many side-effects. For example, ACh drives the most basic function of muscle contraction throughout the body, but in the brain it is one of those high-level control knobs affecting attention, arousal, and learning. Dopamine receptors are also involved in lactation. Evolution is very pragmatic in repurposing existing technology. Furthermore, the major players of serotonin, dopamine, and norepinepherine are all chemically very similar *monoamines*, so many drugs affect all of them to varying extents. Thus, overall, understanding the full effects of any given drug can be very complicated.

- **Caffeine** is a direct antagonist for *adenosine* receptors, which in turn are antagonistic against dopamine, and overall lead to sedation (drowsiness). Thus, consistent with its widely known and appreciated subjective effects, it directly inhibits drowsiness, and also leads to a net increase in dopamine, producing pleasurable effects and leading to its addictive properties.
- **Nicotine** is an agonist for a type of ACh receptor (the *nicotinic* ACh

receptor) that drives the attention and arousal effects of ACh in the cortex. This is consistent with the stereotypical chain-smoking author or detective using nicotine to enhance their ability to focus and concentrate.

- **Alcohol** (ethanol) has complex effects on neurons, that vary with dose and over time. It acts as a GABA agonist, increasing levels of inhibition, which accounts for its psychological effects in reducing anxiety, causing sedation, and reducing "behavioral inhibition" which, paradoxically is facilitated by increasing neural inhbition (it is inhibiting your frontal control system). It also antagonizes the binding of glutamate to the NMDA receptor, which is involved in learning as well see in the Learning chapter. Both the GABA and NMDA effects combine to impair learning in the hippocampus, leading to memory blackouts.
- **Benzodiazepines** (Valium, Xanax, Midazolam, etc) are widely-used GABA agonists, which, like alcohol, reduce anxiety, cause sedation, and generally turn off the brain to varying extents. If you've ever had surgery, you've likely had Midazolam, which knocks you out and prevents you from remembering anything. In low doses, Midazolam has been used in scientific studies to produce a reversible hippocampal amnesia-like condition, due to the heightend sensitivity of the hippocampus to the effects of GABA.
- **Amphetamine** (speed, Adderall) is an agonist for both *norepinephrine (NE)* and dopamine, increasing release and actually reversing the reuptake process so that there is more of these neuromodulators in the synapse. Both of these neuromodulators affect attention and learning, consistent with the observed behavioral and cognitive effects. Adderall is used for treating people with ADHD, which is somewhat paradoxical given the "hyperactive" component of this disorder. However, it is likely that NE acts to keep people actively engaged for a longer time, "locking in" a given set of frontal control signals and preventing the characteristic distractability of ADHD (Aston-Jones and Cohen 2005).
- **Cocaine** is similar overall to amphetamine in both biochemical and psychological effects. It has a specific inibitory effect on *dopamine transporter (DAT)* that is responsible for reuptake of dopamine, thus producing an overal agonist effect on dopamine (leaving more in the synapse). These direct effects on dopamine likely play a critical role in its addictive properties, as it simulates the effects of rewarding outcomes, in a way that circumvents the natural *contrast* mechanisms that discount rewards in proportion to expectations (Redish 2004).
- **SSRI's** (Prozac et al) affect *serotonin* function by inhibiting the process of reuptake of the neurotransmitter after it has been released (i.e., *serotonin-specific reuptake inhibitors*). This allows serotonin to linger longer, and potentially have a larger overall effect. However serotonin is so incredibly complex at multiple levels, that nobody really understands exactly what is going on, and we really can't be sure if it is an agonist or an antognist. For example, serotonin (and all the other neuromodulators) have negative

feedback mechanisms that strongly regulate the amount released, and it is possible that blocking reuptake causes these feedback mechanisms to over-react, thus leading to a net reduction in serotonin release over time. Furthermore, different serotonin sub-nuclei within the raphe have contradictory effects, with some promoting positive emotional states and others having the opposite effects.

- **Psychedelics** (LSD, psilocybin, peyote, etc) all have primary effects on the serotonin system, which, among its many talents, is important for regulating sleep. The simplest explanation for the effects of these substances is that they effectively produce a waking dream state, as we'll explore in the next section.
- **Cannabis** (Marijuana) is a unique case where the drug activates receptors and associated endogenous neurotransmitter systems that were previously unknown, and have only relatively recently been discovered as a direct result of studying the effects of the drug. Thus, the receptors and endogenous neurotransmitters are known as *cannabinoid* receptors and *endocannabinoids*, and now that we have the tools to identify these things, they turn out to be found all over the body, like all the other more well-known neuromodulatory systems. However, unlike these other systems, the reason we never knew of these cannabinoid systems before is that they don't have a central nucleus that releases them – instead they are produced locally in cell membranes, and have a very localized signaling role, by sending messages *backward* across the synapse (i.e., from dendrite back to axon, instead of the usual other way around). The detailed function of these systems is still relatively unknown, and represents an exciting frontier in current research, well-timed with the recent legalization of this interesting substance in a number of different US states and other countries.
- **Narcotics** (heroin, morphine, fentanyl, opiates) are agonists for the endogenous opioid system, involved in regulating the neural response to pain stimuli. Opioid receptors are found in the amygdala, basal ganglia, hypothalamus, and thalamus, and this explains the strong emotional, euphoric effects of these drugs. These substances are widely believed to be the most addictive of all drugs.

In summary, these drugs are jacking right into those global control knobs in the brain, and provide a critical window into undertstanding how our brains function normally: your endogenous states of arousal, excitement, sedation, etc are all controlling these very same knobs. Some of these psychoactive drugs, such as caffeine, alcohol, and nicotine are very widely used (and abused), and there have been increasingly urgent discussions about the ethics of performance-enhancing drug use in schools, which is on the rise. To what extent is this like doping in sports, or should we instead consider it more like using a calculator or a computer: something that augments our native biological abilities to the general betterment of society, etc? What is your opinion?

Neuroscience Methods

Finally, we conclude this chapter with a brief overview of some of the major techniques and methods used to understand how the brain works. We covered the issues of correlation vs. causation in neuroimaging and other such techniques in Chapter 1, so here we focus more on how these techniques actually work, and what their relative strengths and limitations are from a more practical perspective.

Functional Neuroimaging: fMRI, PET, EEG, MEG

The advent of practical techniques for imaging the activity of the living, breathing human brain has truly revolutionized the field of Psychology and Neuroscience. Intial pioneering work was done in the 1980's using the positron emission tomography (PET) scanner, which requires radioactive agents to be infused into the bloodstream. The PET scanner measures the decay of these radioactive labels, which can be formulated to bind to various different substances of interest in the brain, including different neurotransmitters such as dopamine, or glucose (sugar) to measure overall metabolic activity. In 1992, several groups developed the ability to use magnetic resonance imaging (MRI) to measure the level of oxygen in the blood, known as the BOLD (blood-oxygen level dependent) signal, which varies as a function of overall neural activity within a given brain area. Interestingly, the brain over-reacts to neural activity, resulting in an over-supply of oxygen to the most active areas, rather than a depletion. This functional MRI (fMRI) technique has major advantages over PET, in not requiring an IV injection of radioactive tracers, and it has a much faster *temporal resolution* (i.e, the ability to resolve changes in activity over time). Furthermore, MRI machines are used in most clinical facilities of any reasonable size, so this technique made it possible for many scientists around the world to study how the brain responds to all manner of things inside the scanner.

In the ensuing years, fMRI techniques have improved to the point that remarkably small chunks of brain (called *voxels*, which are the volume analog of *pixels* in an image) about 1 mm on a side can be resolved, and in surprisingly many cases, these small voxels cary useful signals about what is going on in a given task. Current approaches typically focus on using the entire pattern of brain activity to understand how the brain works, which is consistent with our overall understanding about the way that many different neurons and brain areas work together to get the job done. Earlier, many scientists focused instead on identifying smallish blobs of activity that were particularly strongly activated by particular tasks (the *neo-phrenology* referred to earlier), but it has become evident that this only gives a small "porthole" view onto the full scope of brain fuction.

While fMRI can resolve relatively tiny voxels (i.e., it has good *spatial resolution*), its temporal resolution is still very limited (even though it is better than PET), because it is essentially measuring changes in blood flow, which take

a while to react to changes in neural activity (about 6 seconds or so on average). A large number of different neural activity states can come and go within that 6 seconds, and all of these end up just getting blurred together in the overall fMRI signal.

To gain more insight into the detailed timecourse of cognition, there has been a continued and increasing use of electroencephalography (EEG), which has been around since the early 1900's, which records real-time electrical signals using electrodes placed on the scalp. These signals immediately reflect changes in neural activity, providing excellent temporal resolution, but, alas, the remote recording of these signals from the scalp makes it very difficult to figure out exactly where the electrical signals are coming from within the brain. Thus, EEG has poor spatial resolution. Unfortunately, we do not yet have the perfect neuroimaging technique, which would have high resolution in both space and time. Nevertheless, advanced techniques in recording (using 100's of electrodes) and analysis have enabled EEG to achieve much better spatial resolution than before, and EEG can be combined with simultaneous fMRI recording to attempt to get the best of both worlds (though this remains challenging). You also may have heard about something called an *ERP* – this is just a way of averaging EEG signals together in a time-locked fashion, to create an *event related potential*, which has characteristic peaks and dips at different points in time, resulting from the waves of brain activation in response to a stimulus, or in preparation of a motor response.

Finally, there is a technique known as magnetoencephalography (MEG), which is the magnetic version of EEG. Recording these magnetic signals, which are much weaker than the electrical signals, requires advanced superconducting magnetometers, which in turn require complex cooling systems to get down to the superconducting realm. Thus, unlike EEG which is relatively inexpensive and portable, MEG is only available in a few labs around the world. However, it does have an advantage in spatial resolution over EEG, due to the way that the scalp distorts the electrical signal, but not the magnetic one.

Conclusions

Many scientists like to emphasize the popular sentiment that "the brain is a complete mystery" and we have barely scratched the surface in our understanding of it. However, you might get somewhat of a different impression from this chapter. In fact, we have a pretty good understanding of the large scale functional organization of the brain, which is consistent with all manner of data from neuroimaging and effects of brain damage, etc. As we'll see in the learning chapter, we have a remarkably good understanding of the details about how neurons learn at the synaptic level, and certainly we know a great deal about all the basic mechanisms underlying spiking. Detailed computer models incorporating all this data have been able to reproduce, at least at a coarse, approximate level, much of the actual human behavior observed in well-controlled laboratory studies, in domains such as perception, learning, memory, language, and cognitive control

(R. C. O'Reilly et al. 2012).

Thus, while there certainly are a few deep mysteries and major discoveries yet to be had, one could reasonbly argue that we are at that stage in solving a jigsaw puzzle where a lot of the edges and key regions have been filled in. Future editions of this textbook may not differ as much as you might think, as we start to fill in the rest of the picture. Only time will tell for sure, but there is at least room for optimism that we really are on the precipice of having a solid *science* of the brain and mind, and the goal of this textbook is to provide a coherent, comprehensive, and possibly a bit premature account of it.

Summary of Key Terms

This is a checklist of key terms / concepts that you should know about from this chapter.

- Neuron: synapse, dendrite, cell body, axon
  - neurotransmitter, reuptake, receptor, channel, ion, membrane potential, threshold, spike / action potential
  - 10,000 inputs to 1 output = *compression*
  - excitatory: glutamate, AMPA, Na+; inhibitory: GABA, Cl-
  - Tug-of-war creates *contrast* – neurons respond *relative to average*
- Brain:
  - Cerebral cortex / Neocortex:
    - Occipital lobe: vision, V1
    - Temporal lobe: objects, auditory, A1, language, episodes, semantics
    - Parietal lobe: action, somatosensory, S1, homunculus, number, space, time, relations
    - Frontal lobe: motor, M1, control, planning, sequencing, decisions, motivation, affect
    - Hierarchy of compression
  - Thalamus: relay, thalamocortical system
  - Basal ganglia: decision making, control, dopamine
  - Amygdala: emotion
  - Hippocampus: episodic memory
  - Cerebellum: error-driven motor learning
  - Hypothalamus: core body functions, HPA axis, stress
  - Brainstem: reticular activating system, dopamine, serotonin, norepinepherine, acetylcholine
- Drugs:
  - Agonist: activates, enhances neurotransmitter / receptor function
  - Antagonist: inhibits, suppresses neurotransmitter / receptor function
  - Caffeiene: adenosine
  - Nicotine: ACh
  - Alcohol: GABA
  - Benzodiazepines: GABA

  - Amphetamine: norepinepherine
  - Cocaine: dopamine
  - SSRI: serotonin
  - Psychedelics: serotonin
  - Cannabis: cannabanoid
  - Narcotics: endogenous opioid
- Methods:
  - PET: radioactivity, slow (bad temporal resolution)
  - fMRI: blood oxygen (BOLD), faster than PET but still slow, good spatial resolution
  - EEG: electric signals, fast (real time, good temporal resolution), but poor spatial resolution
  - MEG: magnetic signals, fast, better spatial resolution, expensive

Chapter 3: Consciousness, Sleep, and Arousal

Intro

Why are we only conscious of the cortex?

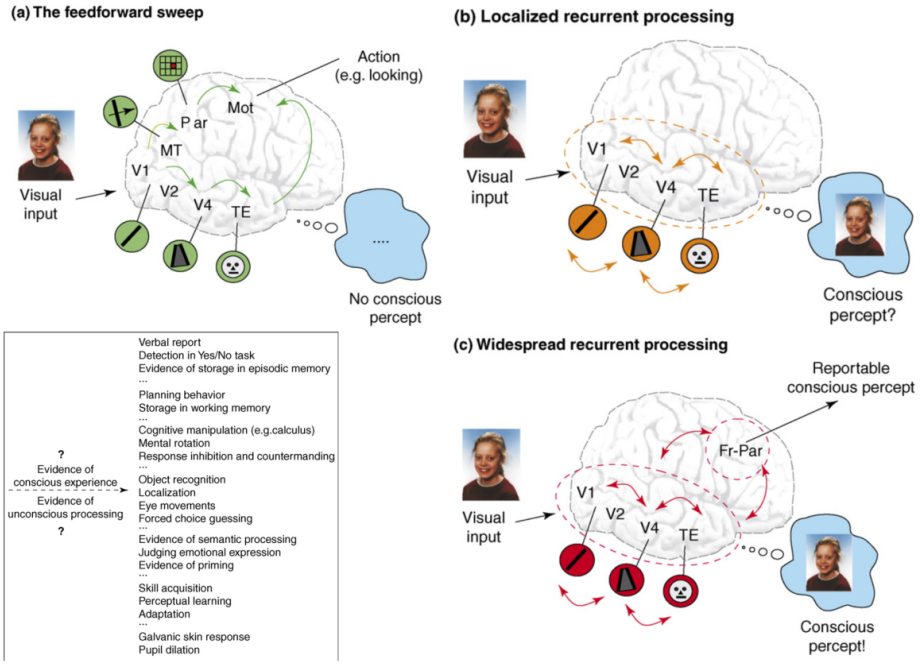


Figure 17: Fig 3-1: Consciousness is associated with robust recurrent processing across wide areas of the neocortex (Lamme, 2006).

todo: metacognition figure too.

Sleep

- basic stages of sleep / arousal and critical roles of brainstem systems and thalamus
- insomnia and the interdependence of frontal cortex with brainstem neuromodulatory systems
- neural basis of input consciousness in terms of bidirectional interactions: consciousness is a reflection of this mutual interdependence and interaction among neurons. Neocortex is unique in having bidirectional excitatory connections, which come at cost of epilepsy.

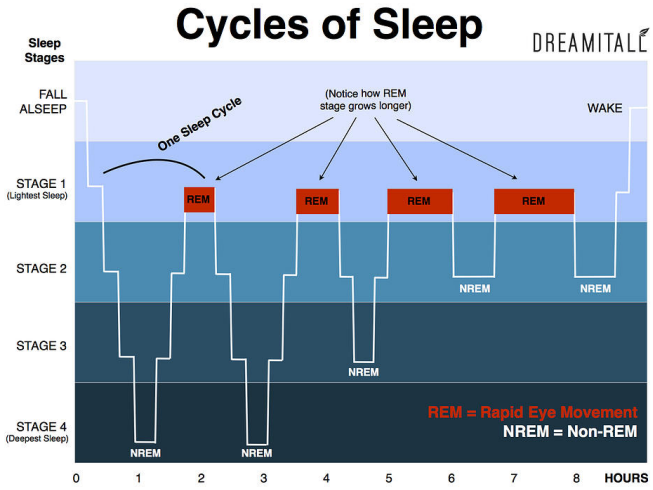


Figure 18: Fig 3-2: The four stages of sleep and their timecourse over a typical night. Deeper stages are characterized by slower brain waves (slow wave sleep), and brain waves during REM are similar to waking. Progressively less deep sleep and progressively more REM sleep occur over the course of a night.

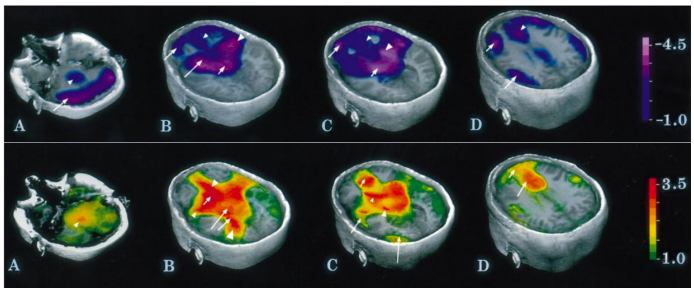


Figure 19: Fig 3-3: The frontal cortex goes deeply to sleep when you're dreaming, while the amygdala and hippocampus are highly active. This explains why your dreams are so incoherent and disorganized, and it is impossible to ever catch an airplane or show up for a test on time, etc. Also, dreams are emotionally charged and incorporate recent memories, thanks to the amygdala and hippocampus.

Chapter 4: Sensation, Perception, and Attention

Sensory Systems

Vision

Audition

etc

Perception

Attention

Key points: Compression, Contrast, Top-down

Chapter 5: Learning, Motivation, and Emotion

Learning is the single most important process taking place in the brain. Without learning, nothing else is possible. All of our focus on the three-C's of compression, contrast, and control presumes a brain with sensible patterns of synaptic connectivity, that produce *useful* forms of each of these phenomena. Without learning, neurons would randomly compress incoming sensory information, detecting irrelevant, bizarre features that don't have any behavioral relevance. Contrast would compare these random things against each other, producing equally meaningless relative comparisons. Control would drive us toward random goals, and our behavior would be just a jumble of strange impulses.

Learning is essential because there are *way* (way, way, way...) too many synapses for any kind of genetic process to shape in a detailed way. There are only about 20,000 different protein-coding genes in the human genome, which is only 2 times the number of synaptic inputs on a *single* neuron. It is inconceivable that genes could code for any sensible fraction of the 100 *billion* times that amount of information that would be required to configure the full human brain. This genetic argument accords with the obvious fact that we learn the vast majority of our abilities over an extremely protracted developmental window, in a way that depends critically on the experiences and education that we are exposed to.

Thus, the brain (specifically the neocortex) is fundamentally a *self-organizing* system, which somehow magically transforms raw sensory inputs into *knowledge* encoded in its billions of synaptic connections. The mystery of this process has long perplexed philosophers, who have explored the opposing ideas of *empiricism* vs. *rationalism* and positions in between. Empiricists embrace the idea that learning proceeds directly from sensory experience, while rationalists argue that there is no way that raw experience by itself is sufficient to create the sophisticated level of knowledge an adult human (philosopher) has. Modern scientific approaches to this question retain much of this ancient debate, with some favoring a generous amount of innate knowledge, and others arguing that almost everything is learned.

We'll return to these issues in the Development chapter, but the quick summary is that neither extreme view is likely to be correct, with genetic and experiential factors each playing critical roles. In particular, there is ample evidence that genes establish broad patterns of initial connectivity and orchestrate developmental transitions, such as synaptic pruning, which in turn strongly influence an experience-driven learning process operating at synapses throughout the neocortex.

Our objective in this chapter is to first understand the nature of these synaptic learning processes, which have been figured out in spectacular detail at this point, and explore some broader ideas about how they might result in this magical self-organization of knowledge over development. Then, we turn to the forms of learning that were the focus of behaviorism: *classical and operant conditioning*. These both depend on similar dopamine-driven learning

mechanisms operating in the basal ganglia, amygdala, and related areas, which are now very well understood. These forms of learning shape our core decision-making process to select actions that are likely to be rewarding, and not punishing.

Finally, we broaden our perspective beyond the limited world-view of the behaviorists, and consider the possibility that *internal* factors such as *goals*, *drives*, and, ultimately, *emotions*, might play a central role in driving both our learning and decision-making behavior. This perspective, long embraced by social psychologists, is only recently beginning to be explored from the neuroscientific angle, which has been perhaps overly-enamored with the remarkable alignment between the classic externally-driven behaviorist conditioning processes and the function of dopamine in the basal ganglia.

Synaptic Plasticity

If learning is the most important thing in the brain, then the most important thing about learning is that it takes place in the synapses interconnecting neurons. This idea goes back at least to *Santiago Ramon y Cajal* in the late 1800's, the pioneering Spanish neuroscientist who advanced the idea that interconnected networks of neurons are doing most of the work in the brain. Logically, the strength of the connections between neurons should alter the patterns of information flow through these networks, and thus makes sense as the primary locus of learning, and knowledge. *Donald Hebb* cemented this idea with a compelling, well-specified proposal that memories are formed when neurons that are active at the same time increase the strength of their synaptic connections, so that they are then more likely to co-activate each other in the future (Hebb 1949). In effect, learning is "gluing together" the different elements of a memory. This idea has been captured with the pithy statement that "neurons that fire together, wire together".

However, it was not until 1966 that this **Hebbian** form of learning was actually demonstrated in the brain, by Bliss and Lomo (Bliss and Lomo 1973). They described a form of **Long Term Potentiation** (**LTP**) of the synaptic strengths between well-defined groups of neurons, where potentiation means "getting stronger" and the "long-term" aspect of it was critical to distinguish from earlier discoveries of synaptic potentiation that only lasted for a few minutes. If synaptic changes are really the basis for learning and knowledge in the brain, they had better last for more than a few minutes, because clearly our memories and knowlege can last a very long time.

The field of LTP research expanded rapidly from that point onward, and progressively more detailed questions were addressed about the exact nature of what is changing in the synapses, and what specific factors in the activity of the sending and receiving neurons on either side of the synapse were critical for causing it to change. After many controversies and twists and turns in this amazing story of scientific discovery, we now have a very solid and detailed understanding of how this process works, at least in terms of all the underlying biochemical mechanisms. It is a fabulous success story for the power of the

scientific method, to drill down and figure out exactly how some complex system actually works. Perhaps most remarkably, Hebb's original idea seems to have been nicely supported, by a remarkable interaction of different moving parts: changing the strength of the synapse requires *both* the sending and receiving neurons to be active.

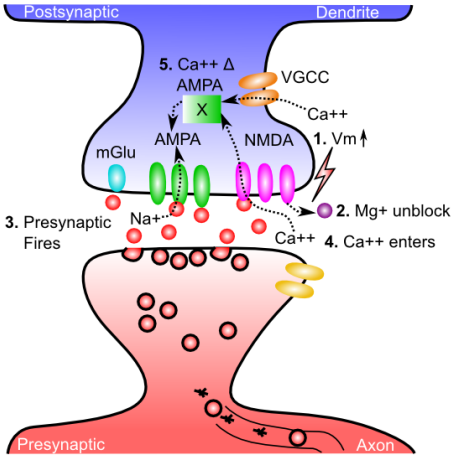


Figure 20: Fig 5-1: Mechanisms of synaptic plasticity, resulting in changes in the overall strength of the synaptic connection between the sending axon and the receiving dendrite. 1. The receiving neuron must be active, so that its elevated membrane potential (Vm) kicks out the positively-charged Mg+ ions from the NMDA receptors (2). 3. The sending neuron must fire and release glutamate, which then binds to the NNDA receptors, causing them to open and Ca++ ions to enter (4). Ca++ then triggers complex chemical pathways that ultimately result in changes in the numbers of AMPA receptors poking out across the membrane, which thus changes the overall amount of Na+ that can enter for any given firing of the sending neuron.

Figure 5-1 shows the major steps in the process of synaptic change. The receiving neuron must be active enough so that its elevated membrane potential pushes out positively-charged magnesium ions (*Mg+*), which are otherwise blocking the opening of the *NMDA* receptors. And the sending neuron must be actively releasing glutamate neurotransmitter, as a result of spiking, because glutamate binding to the NMDA receptors (in addition to the AMPA receptors) is necessary to cause them to open. Whereas AMPA receptors allow Na+ ions to flow into the cell, NMDA allows *calcium* (*Ca++*) ions to enter, and these Ca++ ions then trigger a cascade of chemical reactions that ultimately leads to the change in synaptic plasticity. This critical role for Ca++ is consistent with many other similar such biochemical processes throughout the body – again, evolution often reuses existing mechanisms.

The main consequence of Ca++ entry is a change in the number of AMPA

receptors in the synapse, which then changes the overall amount of Na+ that can enter when the sending neuron spikes. Much more can be said about the details of these Ca++ driven chemical pathways (Rudy 2013), and the other associated changes that take place in the synapse, but the core logic remains the same as Hebb envisioned it: both neurons must be active for the synapse to change.

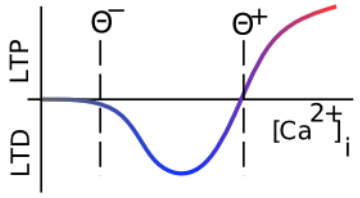


Figure 21: Fig 5-2: Direction of synaptic change as a function of the amount of calcium entering the dendritic spine. Lower amounts of Ca++ result in LTD = long-term *depression* or decrease in synaptic strength, whereas higher amounts result in LTP = potentiation or increase in synaptic strength.

However, Hebb overlooked one *essential* aspect of learning, which was also neglected in the early days of research on LTP. This is the fact that you can't always only increase the strength of synapses. Eventually, all the synapses would get ever-stronger, and the brain would blow up in a huge epileptic seizure. Instead, it is equally if not more important that synapses also *decrease* in synaptic strength, which has been named **Long Term Depression** or **LTD**. Decreases may be more important than increases, from the perspective of the *compression* function of neurons: each neuron has to essentially throw away a huge amount of information in order to compress its 10,000 inputs into a single output signal, and LTD makes synapses weaker and thus facilitates this information filtering process. In any case, Figure 5-2 shows that the balance between LTP and LTD is a function of the overall amount of calcium entering the dendrite – lower amounts result in LTD, while higher amounts result in LTP. This behavior emerges from a competition between two different chemical pathways, one which drives LTP and the other LTD, and their relative dependence on Ca++ levels. This is yet another tug-of-war taking place within neurons – this competitive dynamic is a very commonly-used mechanism at all levels of the brain.

One intriguing finding that makes sense in terms of this balance between LTP and LTD, is that weak activation of perceptual inputs seems to make those things harder to see, while strong activation makes them easier to see (Newman and Norman 2010). Thus, the weak activation leads to the lower levels of Ca++, and causes LTD, whereas the stronger activation drives higher levels and LTP.

Neocortical Learning

Now that we know in detail how learning operates at the synaptic level, you might think that all of the mysteries of brain function should be solved, given what we said about the essential role of learning. Unfortunately, this is not the case. There are a number of challenges here, but chief among them is that there are so many synapses and neurons involved in learning any given bit of knowledge, that it is essentially impossible to go directly from behavior of the individual synapse up to this *emergent* behavior of learning in the larger neural network. The major tool that can be used to bridge this gap are computer simulations of neural networks, with equations capturing things like the function shown in Figure 5-2, and the overall firing activity of neurons in response to stimulus inputs, etc.

Extensive work with such models has repeatedly shown that the known Hebbian-like learning mechanisms described above does *not* result in the kinds of larger-scale learning that people are clearly capable of. The reasons for this are well understood, but beyond the scope of this discussion. Furthermore, the kind of learning that *does* work reliably in these neural models, and is used in the recent powerful AI (artificial intelligence) models currently powering the speech recognition and other advanced capabilities in your cell phone and other gadgets, is called **error backpropagation** (Rumelhart, Hinton, and Williams 1986), and it makes some additional demands on the biology that some influential people have argued are implausible (Crick 1989).

This problem has been my specific area of research for over 20 years, and my colleagues and I have developed progressively more biologically plausible models of how this error-driven learning process could work within the neocortex (R. C. O'Reilly 1996; R. C. O'Reilly et al. 2012). Our latest idea is that the brain is constantly making predictions about what will be seen next, at a rate of about 10 times per second (i.e., the *alpha* frequency), and very specific patterns of neural connectivity in the neocortex and thalamus provide a "ground truth" correct answer against which those predictions are compared. Thus, the difference between these predictions and what actually happens provides the error signals driving learning, and we have shown how these error signals, which exist as differences in the activity states of neurons over time, could drive learning in synapses throughout the neocortex. Furthermore, our computer models show that this form of learning can indeed acquire the kinds of sophisticated knowledge that people do, for example the ability to recognize different categories of objects.

There is just one problem with all this: while our proposed synaptic plasticity mechanisms are consistent with the existing body of detailed knowledge, they also make a few extra demands that have not been tested empirically. So we do not yet know if this theory all goes through or not. Furthermore, there are various other different theories about how all of this could work, which make different, testable predictions. Thus, hopefully we'll get some answers in the not-too-distant future, and then we can potentially connect the dots all the way from the beautifully detailed biochemical level up to the high-level effects of

these mechanisms in forming new knowlege representations within the neural networks of the neocortex. For now, we have to live with a glaring hole in our overall understanding of this most important process of learning in the brain.

One important phenomenon in neocortical learning is **imitation learning**, where somehow we are able to observe other people's behavior and turn that into some approximation of that behavior ourselves. Although this may sound relatively simple, no existing AI models have been able to achieve this feat, and upon closer examination, the process of turning the perception of behavior into your own motor program requires a highly sophisticated perceptual and motor control system. Thus, the fact that even young infants appear to be capable of this is quite remarkable (Meltzoff and Moore 1994). An important neural substrate for this form of learning has been found, in the form of **mirror neurons** that appear to achieve this feat of mapping observed behavior into the same patterns of neural firing that are active when you perform the same behavior (Iacoboni, Woods, and Rizzolatti 1999). However, it is not known how these neurons learn this mapping in the first place, so it remains a phenomenon in search of a deeper explanation. Nevertheless, there is an intriguing suggestion that these mirror neurons might be affected in autism spectrum disorders, which could potentially account for the difficulties in empathy in this population (Gallese, Keysers, and Rizzolatti 2004).

Dopamine-modulated Learning

Most introductory textbooks do not address any of the above topics in learning, and focus exclusively on the relatively well-understood domain of conditioning, which has been studied since the days of Pavlov and the behaviorist school in the early 1900's. This has become an area of renewed interest in neuroscience, since the discovery that dopamine activity almost perfectly accounts for the nature of these conditioning phenomena (P. R. Montague, Dayan, and Sejnowski 1996; Schultz, Dayan, and Montague 1997).

Classical (Pavlovian) Conditioning

The classical conditioning paradigm (Figure 5-3) centers around learning the connection between a previously *neutral* stimulus (the **conditioned stimulus** or **CS**) and a biologically-established, affectively significant *outcome*, known as the **unconditioned stimulus** or **US**. In the classic experiments by Pavlov, the ringing of a bell served as the CS, and food reward as the US, and the subjects were dogs, who learned over a few repetitions of the CS followed by the US to salivate after hearing the bell, in antipation of receiving the food. The salivation is somewhat confusingly labeled the **un/conditioned response** (U/CR), where it is *un*-conditioned (UCR) prior to learning in response to the food US, and *conditioned* (CR) after learning in response to the CS. So, the same response has two different labels depending on what is driving it. Ecologically, this simplified lab experiment is thought to capture the real-world learning about different

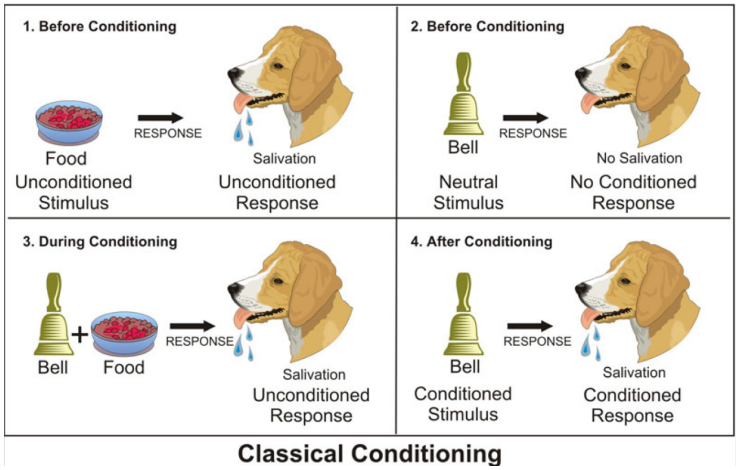


Figure 22: Fig 5-3: The classical conditioning paradigm.

stimuli that help us anticipate and prepare for important upcoming outcomes. For example, when you are hungry and driving down the highway on a road trip, the sight of a McDonald's sign alerts you to the availability of food there. Thus, the McDondald's sign is effectively a CS, and indeed this conditioning paradigm applies well to the goal of advertising, which is to establish a solid connection between a brand logo and desireable US outcomes.

Although Pavlov and the behaviorists where exclusively concerned with overt behavior such as salivation, we now know the internal biology that drives this form of learning. Figure 5-4 shows how dopamine neurons in the *ventral tegmental area* (*VTA*) of the brainstem reticular activating system respond in a classical conditioning experiment (Schultz 1986; Schultz, Dayan, and Montague 1997). When the US (labeled R = reward, a juice drop) is presented without any prior CS, dopamine responds with robust firing above its "tonic" steady base rate of firing. This is consistent with the naive idea that dopamine encodes raw reward signals. However, when the very same reward is presented after a CS (which has been reliably paired with the reward in prior conditioning trials), *dopamine no longer fires to the reward!* Furthermore, when the CS is presented and the reward is *withheld*, dopamine neurons show a suppression or *dip* in firing below their tonic baseline. Psychologically, you would feel disappointed if you didn't get the reward you expected, and indeed that is exactly what the dopamine neurons are signaling.

These results, from the pioneering work of *Wolfram Schultz* and colleagues, have profound, far-reaching implications, and represent one of the most exciting and important findings in neuroscience. They are also one of the most important examples of the *contrast* principle, as we emphasized in the Introduction. Specifically, these results show that dopamine neurons respond to the contrast or difference between an expectation or prediction of reward, and what is actually

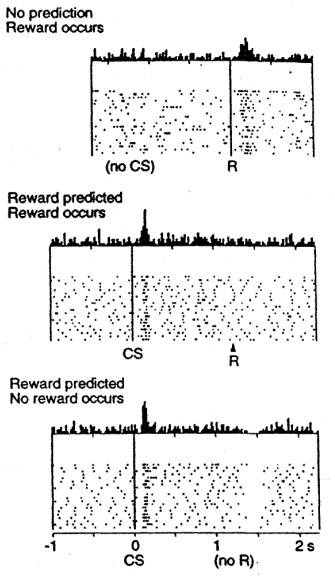


Figure 23: Fig 5-4: Dopamine neuron firing in a classical conditioning paradigm of CS followed by US (labeled R for reward – it was actually a drop of juice). Top: Unexpected rewards (at time point R) drive dopamine firing. Middle: Trained CS followed by R shows dopamine firing at the CS, but *not* for the reward. Bottom: Trained CS followed by *omission* of R shows reduction of firing at R. Each row of dots shows when a dopamine neuron fired a spike on a given recording trial, and the bars at the top show the accumulated histogram of all the spikes at that corresponding point in time across all such trials. Dopamine does *not* respond to raw reward input, because it fails to fire in when the reward is accurately predicted by the CS, in the middle panel. Furthermore, it directly signals "disappointment" by reducing doapmine firing when an expected reward is not received. These and many similar results show that dopamine responds to the *contrast* or difference between predicted and actual rewards. From Schultz et al, 1997.

received, *not* to the raw reward input itself. As we suggested in the Introduction, this contrast property of dopamine is what drives insatiable greed, dissatisfaction, and apathy, because once we learn to expect any given positive reward-like outcome, we no longer receive dopamine for it! This property of dopamine is what causes kids to be so entitled and spoiled: they come to expect all that coddling from their overprotective parents, and all the excitement they get from playing video games, so that when they finally get out into the "real world", it is all so difficult and filled with disappointing drudgery. It is also why so many famous people, especially rock stars it seems, turn to dopamine-activating drugs of abuse – once they adapt to their new amazing famous lifestyle, their dopamine system no longer gives them that amazing feeling of unexpected reward. Drugs like cocaine artificially bypass the expectation-driven contrast mechanisms of the dopamine system, producing more reliable bursts. However, even here the system slowly adapts and more and more drug is required to achieve the same effects, so really there is no escape from the evil maw of the dopamine contrast effect!

From a hard-nosed learning theory perspective, there is a very good reason why the dopamine system must work in this contrast-based way: *learning is most efficient when it is focused on what is not yet learned.* Learning something you already know simply doesn't make much sense. Thus, in the case of classical conditioning, continuing to learn about the fact that the CS predicts the reward after the system has already acquired this association doesn't make any sense. And this logic shows that dopamine is fundamentally a *learning* signal, not a reward signal. In particular, as we'll see in the next section, dopamine directly affects learning in the basal ganglia and other brain areas, including the areas that are learning about the CS – US association in the first place. Thus, as dopamine stops firing at the time of the US (R, reward), it stops the further learning of this association.

This basic theory of how learning should function was systematized by *Robert Rescorla* and *Allan Wagner* in a seminal paper (Rescorla and Wagner 1972), where they proposed a very simple mathematical "learning rule" that says that the amount of new learning should be proportional to this contrast or difference between what you already expect the reward should be, and the actual reward you receive. This is also known as a **reward prediction error** (**RPE**). Roughly a decade later, *Rich Sutton* and *Andy Barto* published an important extension to this idea (Sutton and Barto 1981), known as the *temporal differences* (*TD*) learning rule, which can also account for the fact that dopamine learns to fire at the onset of the CS, even as it stops firing for the expected US. Furthermore, this work led to the development of many advanced mathematical techniques in a field collectively known as *reinforcement learning* (*RL*), which is a branch of *machine learning* that deals specifically with learning from overall reward / punishment signals. These RL techniques have been used in many different AI technologies, and play a central role in the recent advances from the Google DeepMind group, in their models that learn to play Go and challenging video games (Silver et al. 2017; Mnih et al. 2015).

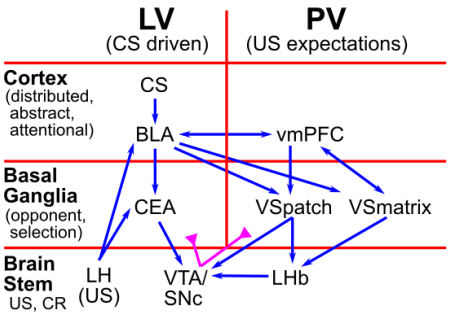


Figure 24: Fig 5-5: Biological systems involved in classical conditioning. The vertical organization reflects the separable contributions of the amygdala to forming CS – US associations (labeled "LV" for "learned value"), and the ventral striatum in driving contrast-with-expectation firing in the dopamine system (VTA / SNc) (labeled PV for "primary value" – together these constitute the PVLV system named in honor of Pavlov). Horizontally, the BLA (basolateral amygdala) within the LV system learns to associate CS's with corresponding US's, while the CEA (central amygdala) reduces these higher-level associations down to specific go / nogo signals in a basal-ganglia-like fashion, and directly drives dopamine firing and core behavioral responses (*conditioned responses*) appropriate for different US's. The PV system likewise has a cortical component in the ventral and medial areas of the prefrontal cortex (vmPFC), and a basal-ganglia component in the ventral striatum (VS). Dopamine firing in the VTA / SNc drives learning throughout all of these areas.

Although the TD learning rule provides an elegant and powerful mathematical description of classical conditioning, the brain networks actually involved in this form of learning are considerably more complex. Figure 5-5 (Mollick et al. 2018, submitted; Hazy, Frank, and O'Reilly 2010) shows a summary diagram of these networks. Conditioning learning involves interactions between two "vertically" organized sub-systems, one involved in forming associations between CS's and US's, which depends on different areas in the *amygdala*, and another that drives the contrast-with-expectations differences in the dopamine system, which depends on the ventral (bottom) areas of the striatum in the basal ganglia. Furthermore, within each of these systems, there are separable contributions from cortex and cortex-like processing in the amygdala, versus the go vs. nogo kind of decision-making that the basal ganglia is specialized for (which we'll learn more about in the operant conditioning section below). This framework can account for a wide range of data about the biology and function of dopamine-driven learning in the brain, and, given the overall complexity of the system, the ability to simulate it all in a computer model is essential for understanding how it all works.

Extinction and Context in Conditioning

As was the case with LTD (long-term-depression), figuring out how the associations between CS and US are *unlearned* is just as important as figuring out how they are learned. This involves the phenomenon of **extinction**, where the CS is repeatedly presented while withholding the US. From Figure 5-4, this should produce repeated *dips* in dopamine levels, which in turn should drive LTD in synaptic connections, causing the association between the CS and US to be unlearned. While this all does occur, the situation turns out to be considerably more complicated, in ways that make sense ecologically. To make a long story short, the brain actually learns *new associations* during these extinction events, in addition to weakening (somewhat) the existing ones. These new associations effectively encode **context-specific exceptions** to the original association – e.g., "in this particular situation, you're not going to get the food, but you might still get it in other situations". Furthermore, the nature of this new learning is under top-down control from the ventral / medial frontal cortex, which can play a critical role in interpreting the nature of what is going on: has the world really changed, or is it just kind of random? (Quirk and Mueller 2008; Gershman, Blei, and Niv 2010)

The advantage of all this is that the initial CS – US association is relatively preserved, and expecially if this was something learned through a painful, dangerous experience, it is probably a good idea to keep these memories around. Better safe than sorry. The disadvantage is evident in the phenomenon of PTSD (post-traumatic stress disorder), where traumatic memories cannot be extinguished, and keep intruding into normal life. There are significant individual differences in the extent of PTSD, and a major factor reflects the ability to exert top-down control and establish a strong new context to override the traumatic situation.

In the lab, these extinction phenomena are observed in the phenomena of **spontaneous recovery**, **reinstatement**, and **renewal**, which are typically observed in aversive conditioning situations (i.e., when the US is a negative outcome, like getting shocked). Spontaneous recovery refers to the re-emergence of the CS – US association after extinction (typically after a break), without any further training, clearly showing that extinction learning did not erase this original memory. Reinstatement occurs after a single US presentation without the prior CS, after which the CS – US association is reinstated. The US reactivates the associated memories and this is enough to overcome the extinction learning. Renewal is particularly revealing of the important role of context. In this case, the subject is conditioned in one environment (A) and extinguished in a second, novel context (B). When put back into the original context (A), the original CS – US association is *immediately* effective without any further learning. In other words, the subject learned a context-specific exception ("when in context B, I won't get shocked") instead of unlearning the original association.

Operant / Instrumental Conditioning

Classical conditioning is the sensory front-end to the other major form of learning studied by the behaviorists: *operant* or *instrumental* conditioning. This form of learning occurs through the reinforcement or punishment of *actions*, instead of stimuli. The central idea is captured in **Thorndike's law of effect**: *actions that lead to good outcomes are more likely to be taken, while those that lead to bad outcomes are less likely* (Thorndike 1911). This is so intuitive that it is difficult to imagine it being otherwise, but nevertheless, it captures a considerable amount of behavior in humans and animals.

We now know how this type of learning works, in terms of dopamine's effect on the basal ganglia (Figure 5-6) (M. J. Frank 2005; Gerfen and Surmeier 2011). Unexpected positive outcomes following a given action result in a burst of dopamine (as we saw in Figure 5-4), and this dopamine burst acts on D1 receptors located on the "Go" neurons of the basal ganglia to drive LTP of the synapses into the neurons that decided to trigger that action. Thus, these stronger synaptic inputs make it more likely that the same action will be triggered again in the future, when similar inputs are driving the basal ganglia (i.e., in similar situations), thereby achieving Thorndike's law of effect. The opposite pattern of changes occurs when unexpectedly bad outcomes arise, which drive dips in dopamine firing, and end up strengthening inputs to the "NoGo" neurons that compete against the Go pathway and prevent an action from being triggered. Thus, actions that lead to bad outcomes are less likely to be triggered, consistent with the other half of Thorndike's law of effect.

The overall relationship between dopamine and the basal ganglia is summarized in Figure 5-7, where classical conditioning processes train the **critic** what kinds of rewards or punishments to expect, and the resulting differences between these expectations and actual rewards / punishments, reflected in the dopamine signal, then drives learning in the **actor** (basal ganglia). This image

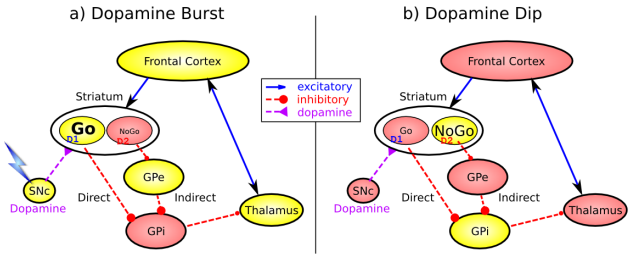


Figure 25: Fig 5-6: How increases in dopamine (bursts) and decreases in dopamine (dips) drive learning in opposing Go vs. NoGo pathways in the basal ganglia. Through the complicated basal ganglia circuitry, the firing of Go (aka direct pathway) neurons leads to a net excitation of motor plans in the frontal cortex. The NoGo pathway has the opposite effect, preventing the frontal activation that would otherwise occur from Go activation. When an action leads to an unexpected positive outcome, the resulting dopamine burst activates a special type of dopamine receptor (the D1 receptor), which drives LTP learning in the input synapes to the Go neurons. This makes those neurons more likely to fire again under similar circumstances, achieving Thorndike's law of effect. The opposite happens when dopamine dips occur for unexpectedly bad outcomes, which interesingly has a net LTP effect on the NoGo neurons via D2 receptors, and an LTD effect on the Go neurons.

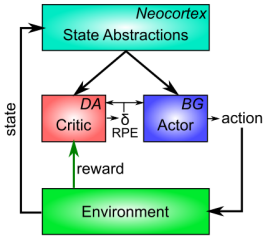


Figure 26: Fig 5-7: *Actor-Critic* schematic for the relationship between the dopamine (DA) signal driven by principles of classical conditioning (the critic), and action decisions triggered in the basal ganglia (BG) (the actor). The environment *state* is represented by various abstract, higher-level *compressed* representations in neocortex, which feeds into both the critic and actor. The actor decides on actions to take as a function of these neocortical inputs, and the critic generates predictions about the kinds of US outcomes that are likely to result. Learning in both the critic and actor is a function of the dopamine signal, which is symbolized as a delta, or *reward prediction error* (RPE). Thus, classical and operant / instrumental conditioning are connected through this actor – critic relationship.

of dopamine as a critic fits with our overall conception of the *contrast* nature of this signal: it is never satisfied and quick to criticize, just like a critic. Of course, the poor actor has to do all the hard work of coming up with stuff for the critic to critique, but, as you may have experienced, it is often hard to be properly critical of your own behavior, whereas it is much easier to see what is wrong with other people. Thus, separating the critic and actor components in the brain makes sense, and is another example where two fully interdependent systems can nevertheless be seen as performing distinct functions.

Partial Reinforcement, Gambling, and Shaping

One of the topics that the behaviorists explored extensively was *reinforcement schedules* – different rates and patterns of delivering rewards. The most interesting and relevant finding from all this work is that **partial reinforcement** can have surprisingly strong effects compared more reliable reinforcement schedules. In a partial reinforcement schedule, rewards are only delivered randomly on a fraction of successful action trials. In effect, it is just like gambling, where there is a relatively infrequent, random payout. The net effect of this is to confound the critic system, which can no longer accurately predict what kind of outcome to expect. Therefore, when a positive reward is received, it is not *discounted* like would have been if it was perfectly predictive. You will get that burst of dopamine for the reward! This is why gambling can be so addictive – it works just like addictive drugs in disabling the stingy, harsh dopamine critic.

Another important discovery in instrumental conditioning was that more complex behaviors can be built up from simpler elements through the process of **shaping**. This is the technique used to get circus animals to perform their complex tricks, for example, and is often used in scientific research with animals to study more difficult cognitive tasks.

Finally, it is important to recognize the difference between a **primary** vs. **secondary reinforcer**. A primary reinforcer directly satisfies a biological need (e.g., food or water), while a secondary reinforcer is indirect, and must be learned. Money, points, and gold stars are common examples of secondary reinforcers, which are effective for motivating people to do things. Interestingly, animals typically require primary reinforcers, but people readily learn to value secondary reinforcers. This ability to value initially arbitrary stimuli is essential for modern economic life – it would be rather inconvenient to have to directly exchange food, water, or other items of direct value.

Motivation

Despite the satisfying modern synthesis between dopamine and the behaviorist-era conditioning phenomena, this overall view of behavior focuses almost entirely on **external / extrinsic** factors (reward / punishment) to the exclusion of **internal / intrinsic** factors such as goals, drives, desires, etc. This is consistent with the behaviorist-era prohibition on considering internal factors more generally,

but we should have no such constraints on our modern thinking about this topic. Nevertheless, the current research still carries some of this extrinsic bias, with the central role of internal factors having been somewhat less emphasized. By contrast, researchers in the field of social psychology have a long tradition of thinking about the central role of goals, desires, emotions and mood on behavior.

Before exploring some of these ideas, it is interesting to ponder the state of mind of a behaviorist from the 1920's: did they really think that their *own* personal behavior was fully determined by external rewards and punishments? Were they not aware of having internal goals that drove them to torture rats for long hours, day after day, in pursuit of such ineffable, remote rewards as scientific understanding and a chance of prestige and fame? The tangible rewards associated with scientific research are sufficiently distant and improbable, while the immediate working conditions involve relative poverty and extreme hard work, that it is really hard to understand why people would do such a thing without invoking some significant long-term internal state variables.

Figure 27: Fig 5-8: Drive reduction theory according to Hull, 1943. Basic needs create drives when those needs are not satisfied, and behavior is then recruited to satisfy those drives.

The one form of internal state that behaviorist's did consider was the notion of a *drive* or state of internal discomfort (e.g., due to lack of food or water) that then motivates behavior toward reducing that discomforting state (Hull 1943) (Figure 5-8). But this **drive reduction** theory has trouble accounting for motivations such as our desire to learn and work, which don't really seem to be associated with discomfort-reduction processes.

A more comprehensive theory of motivation was developed by *Abraham Maslow*, at around the same time as Hull (Maslow 1943). Maslow's **hierarchy of needs** (Figure 5-9) captures the intuitive idea that higher-level needs are not relevant unless the more basic needs essential for survival are satisfied. The two lowest levels in the hierarchy are physiological needs (breathing, food, water, etc) and safety. Once those are satisfied, then higher-level needs such as love and belonging and esteem become relevant. Finally, at the highest level, Maslow put *self actualization*, which includes things like morality, creativity, and lack of prejudice. Interestingly, this highest level resembles the Bhuddist notions of enlightenment, where one transcends lower-level attachments and needs, and can act in a more principled, rational, and yet spontaneous manner. These frameworks capture the subjective feeling that we are slaves to our basic needs, and we yearn to be free from these low-level demands.

One problem with Maslow's theory, shared with any theory that attempts to articulate universal features of human behavior, is that people are rarely so

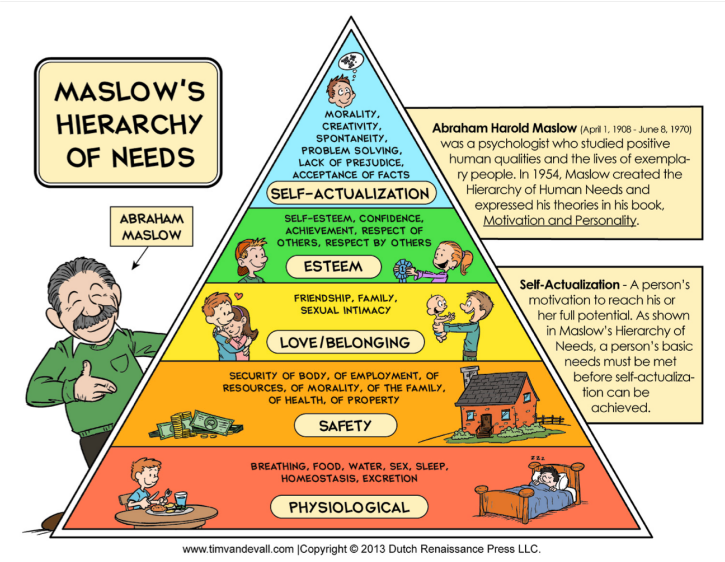


Figure 28: Fig 5-9: Maslow's hierarchy of needs. Higher-level needs are only considered once lower-level ones are satisfied.

compliant, and regularly violate his strict hierarchy. For example, people have been known to literally work themselves to death, including recent cases of video gamers playing to death as a result of neglecting basic bodily needs. Furthermore, teenagers routinely risk their personal safety in order to show off and otherwise enhance their social belonging and perceived self-esteem. Nevertheless, as a general tendency, the hierarchy makes sense, and certainly the numerous cases of cannibalism in the face of extreme hunger suggest the power of these more basic physiological needs.

Goal-driven Behavior

A more general motivational framework is based on the notion of *goals* and the idea that people are specifically motivated to achieve their goals. These goals can be highly diverse in their specifics, but they share the common property of delivering positive reward signals upon goal completion (e.g., "the satisfaction of a job well done"), or even progress toward goal completion ("almost there, just around the corner.."), and corresponding negative states associated with failure (disappointment, embarassment, lack of self-esteem). Many aspects of goal-driven behavior have been studied over the years (Tolman 1948; G. A. Miller, Galanter, and Pribram 1960; Powers 1973; Klinger 1975; Gollwitzer 1993; Carver and Scheier 1990).

In the animal behavioral tradition, goal-driven behavior has been studied in the context of paradigms such as **satiety** and **devaluation** (Balleine and

Dickinson 1998). In these cases, an animal is instrumentally conditioned to press one lever for food while in a state of hunger, and is then given as much food as they want. They are then put back into the box with the lever – if behavior is driven by purely *habitual* stimulus – response associations, they should push the lever even if they are no longer hungry. However, if they are actually thinking about the outcome produced by pressing the level, and recognizing that they don't want that outcome, then they should not press the lever. Interestingly, results show that damage to the ventral and medial areas of prefrontal cortex (**vmPFC**) cause rats to press the lever even when they are full. The same kinds of results have been shown in the devaluation studies, where the food is subsequently paired with a bitter taste outside of the lever-pressing context, so that it is no longer desireable. If the animal still presses the lever, then they aren't clearly representing the outcome of the lever press.

The importance of the vmPFC brain areas for goal-driven cognition is consistent with neural data showing that these areas (in particular the *orbital frontal cortex*, *OFC* and *anterior cingulate cortex*, *ACC*) have many neurons that anticipate the possible US outcomes associated with a given situation and actions taken within that context, and impair goal-directed behavior when damaged (Rudebeck et al. 2006; Jonathan D. Wallis and Kennerley 2011). More generally, this is consistent with the overall role of the prefrontal cortex in driving goal-driven controlled behavior, which requires the tight coordination between plans and their potential outcomes in order to decide on the plans that will lead to the most desireable potential outcomes. As discussed in the Neuroscience chapter, these vmPFC areas are directly interconnected with the basal ganglia, amygdala, and dopamine system, forming the overall *control* and *decision-making* system of the brain, and each of these areas plays a critical role in supporting the overall emergent ability to behave in a goal-driven, controlled manner. Furthermore, as we'll see in the clinical disorders chapter, these are the brain systems that are implicated in most of the major clinical disorders.

Finally, one of the most fascinating and important demonstrations of the importance of intrinsic motivation comes from studies showing that giving people extrinsic rewards can actually *undermine* intrinsic motivation (Deci, Koestner, and Ryan 2000)! For example, giving kids awards for drawing actually caused them to draw less than kids who did not receive these awards. These results are controversial, however, and systematic reviews of the literature have reached opposite conclusions (Cameron, Banko, and Pierce 2001). One of the most important factors appears to be whether the task in question is actually reasonably strongly intrinsically motivating in the first place: there is stronger evidence of the undermining effect when the task has stronger intrinsic interest, compared to more "boring" tasks, for which external rewards might be useful.

Emotion and Arousal

The fact that the same brain areas involved in goal-driven motivated behavior are also the primary areas associated with emotion raises the important question as

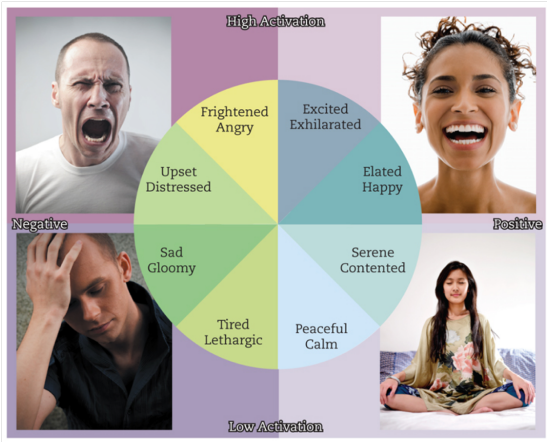


Figure 29: Fig 5-11: Valence vs. arousal *circumplex* model.

to the relationship between emotion and motivation. It is somewhat difficult to provide a crisp, principled definition of *emotion*, which thus makes it difficult to arive at a clear understanding of its relationship with motivational states. Some widely-recognized properties of emotion are that it has some kind of distinctive, characteristic subjective feeling, is associated with physiological arousal at least to some extent, and that it drives associated behavioral responses. It is also generally agreed that emotion should be biologically grounded, at least in the more "primitive" or basic level of emotions.

All of these properties are consistent with the idea that emotional states and motivations have a strong connection (Cardinal et al. 2002). For example, one's overall *happiness* is most strongly associated with feelings of personal self-efficacy and control (along with interpersonal connectedness and belonging). Likewise, feelings of *sadness* are strongly associated with dissappointment, failure, and lack of control. Thus, a simple overall hypothesis is that emotions are the subjective states associated with our core motivational systems. Let's see how far this idea can take us, in understanding the full spectrum of emotional states.

Figure 5-11 shows the simplest standard model of emotion, known as the **circumplex model**, which distinguishes between two separate dimensions of *valence* vs. *arousal*. Valence refers to the "sign" of the emotion, positive vs. negative, while arousal refers to the intensity of the emotion. Anger and exhilaration are opposite valence but the same high level of arousal. These two valences are also associated with opposing *approach* vs. *avoid* behavioral orientations, which have been identified as core opponent aspects of emotional / motivational states and corresponding personality dimensions (Carver and White 1994; Read et al. 2010).

While this simple framework captures the most essential dimensions of affective / emotional states, it is likely that the valence aspect of emotion

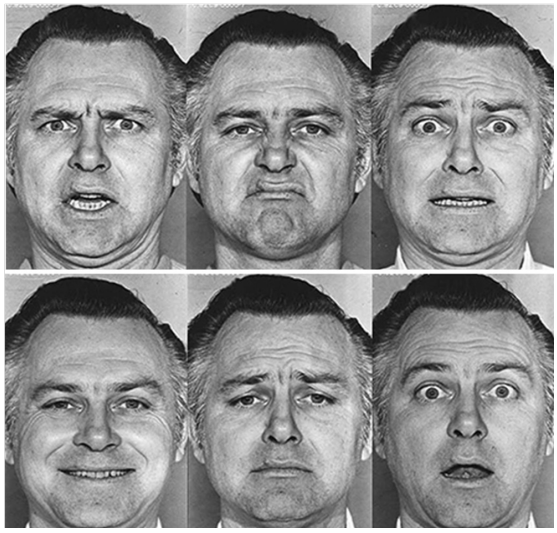


Figure 30: Fig 5-12: Six different basic emotions as represented by facial expressions: anger, disgust, fear, happiness, sadness, and surprise.

is considerably more complex than the *bivalent* (two valences) nature of the circumplex model. For example, *Paul Ekman* found that there are 6 **basic emotions** that have clearly recognizable facial expressions, which are universal across cultures: anger, disgust, fear, happiness, sadness, and surprise (Ekman and Friesen 1976). Later work added other emotions based on vocal and facial expressions, and Plutchik proposed a systematic wheel of emotions based on 8 emotion categories arranged in opponent pairs, with an arousal dimension as well (Figure 5-13).

In addition to the basic happy / sad elements of emotion, which may be more closely related to goal-driven motivational states, some of these other emotional states are more clearly social in nature, and the role of facial expressions and vocalization clearly implicates a strong social communication role for emotions. Thus, we can potentially organize emotional states in terms of a set of distinct functional domains, where these states can serve to motivate people toward appropriate patterns of behavior. We roughly organize these according to Maslow's hierarchy of needs, with slightly different groupings.

- Physiological States: Hunger, thirst, pain, tiredness, lust, and the need to excrete are all basic motivational states associated with core body functions necessary for survival, and correspond with the physiological level in Maslow's hierarchy. These may not be considered "emotional" states per se, but they share the same properties of being strongly biologically determined, varying in level of intensity or arousal, and capable of driving appropriate behaviors to mitigate negative states and approach positive ones. While hunger and thirst may not typicaly need to be communicated

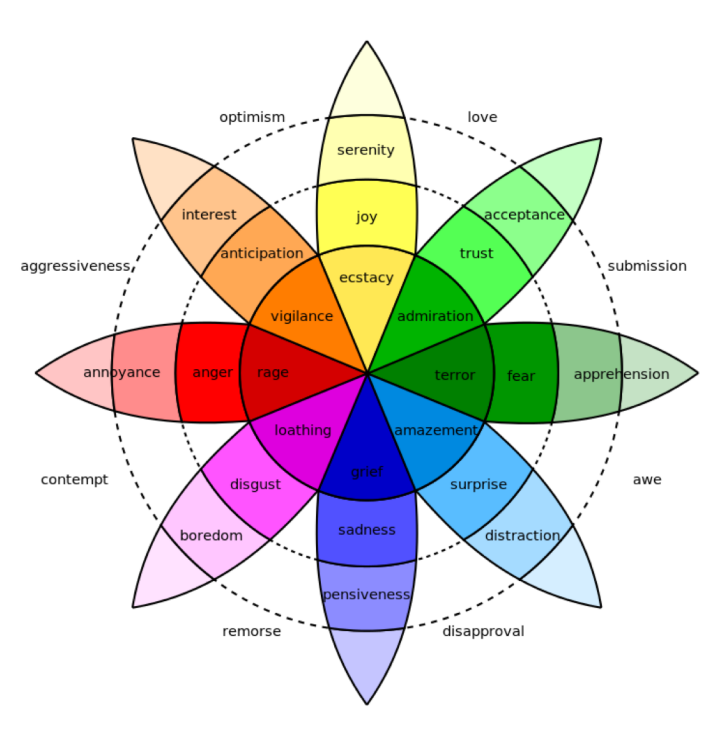


Figure 31: Fig 5-13: Plutchik's wheel of emotions, with arousal (intensity) represented as distance from the center along any of 8 different categories of opponent emotions.

socially using basic facial expressions, tiredness and lust likely do, and have clear social cues in the form of yawning and flirting behavior.

- Safety states: Fear is the emotional correlate of Maslow's safety level, and has both a direct internal motivational role (driving you to avoid scary situations), and an important social communication role for alerting others of potentially dangerous situations, which is facilitated by the presence of the unique fear facial expression. Disgust is an interesting case which may be more strongly social in nature: it is important to communicate to others that food might be rotten and disgusting, and it seems likely that this original function has been extended to apply to labeling the behaviors of others in the group as dangerous or otherwise something to be avoided. Hate is an emotional state that is also clearly negative and social in nature, and associated with disgust: it is the emotional state and social communication associated with labeling others as belonging to the out-group.
- Social states: Love is the opposite of hate, and is the positive social affective state associated with members of the in-group. It obviously corresponds with Maslow's 3rd level of love and belonging. Other important social states include dominance and submission dynamics, along with trust and admiration, which have to do with establishing and perceiving relative status within the social order. These correspond with Maslow's esteem level, and are very strong and often-overlooked motivational states for social beings, from dogs to monkeys to humans. We do not necessarily have clear terms for these as emotional states (e.g., the feeling of being dominated by, or of dominating, a social other), but there is evidence that they are important factors in personality and interpersonal interactions (Hopwood et al. 2013), and certainly we have terms such as "diss" = disrespect and "pissing contest" that refer to such interactions.
- Goal-associated states: many of the remaining states are associated with goal-driven behavior, including happiness and sadness (and their varying levels of intensity or arousal) as noted above, but also anger and frustration which are associated with impediments to progress toward achieving one's goals, and curiosity, interest, and surprise which are associated with recognition of new interesting avenues to pursue. Boredom, distraction, optimism, and anticipation are also other states that clearly seem to be goal-related. Grief and loss are perhaps not so obviously goal-related, but in some ways they reflect a profound disruption of one's sense of overall control and order in the universe, in addition to the basic feelings of missing a loved one.

Thus, overall, it does seem that emotional states can generally be understood as corresponding to biologically-determined motivational states, which provides a clear functional story for why we have emotional states in the first place (Cardinal et al. 2002). As such, this raises important questions about the standard "Hollywood" story about the special status of emotion as a unique aspect of human beings. Under this motivational framework, many of our

emotional states are common across all mammals at least, and represent a genetically-coded, low-level aspect of our brains, not something special and unique about humans. On the other hand, because our emotional states are so strongly felt, and provide dramatic color to our lives, we regard them as special.

Also, emotion is what keeps us from harming each other (except when it is what drives us to harm each other, in the case of hate and anger), and the lack of basic emotional connections in psychopaths enables them to do horrible things that "normal" people would never do. So, from a survival perspective, we really depend on everyone sharing these protective emotional responses, and anything that doesn't is immediately scary and foriegn. Furthermore, we do have a large portion of our vmPFC devoted to emotional processing, and these emotional representations are likely novel combinations of more basic, lower-level emotional states, shaped over our personal histories, and thus likely provide a much richer and elaborated emotional tapestry than found in other animals.

Emotional / Motivational Encoding in vmPFC

Figure 5-14 shows a map of what the vmPFC emotional / motivational tapestry might look like, based on tracing the inputs and outputs of these areas relative to lower-level emotional and motivational areas in subcortical areas (Ongür and Price 2000). Consistent with the circumplex model (Figure 5-11), there are separable areas for positive vs. negative valence, and arousal. Interestingly, the negative valence area, known as area 25 or subgenual ACC, has been implicated in major depressive disorder through the work of *Helen Mayberg* and colleagues, and electrical stimulation in this area is a promising treatment (Riva-Posse et al. 2014).

Also, consistent with the anatomical principles from the Neuroscience chapter, these areas relate to nearby areas in terms of the ACC areas at the top relating to motor plans coded in surrounding PFC areas, and OFC areas toward the bottom being driven by visual, olfactory, taste, and viceral inputs coded in nearby areas. Thus, we can thing of ACC as being more associated with action planning, including things like effort and difficulty costs, while OFC is more important for representing outcomes in terms of their relevant sensory features (taste, appearance, etc).

Biological Grounding of Emotion and Arousal

Finally, there is a somewhat strange history of thinking about emotion that is typically emphasized in introductory textbooks, and seems to reflect the desire to understand emotional states as special, biologically-grounded, important states. Specifically, *William James* and *Carl Lange* each independently proposed that emotion arises first in our bodily responses such as sweating, heart racing, etc, and is only later recognized as an emotional response as a direct result of these initial physiological responses. In contrast to this James-Lange theory, *Walter Cannon* and *Phillip Bard* proposed that higher-level processes in the brain play a

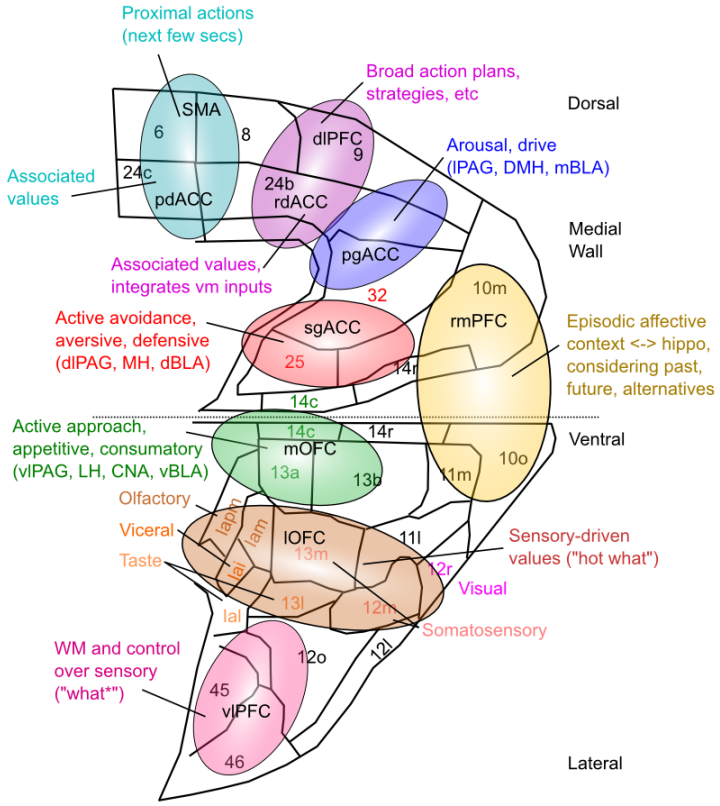


Figure 32: Fig 5-14: Map of ventral / medial frontal cortex (vmPFC) areas and their associated roles in emotional / motivational states, as a function of connectivity with subcortical areas that have established emotional / motivational valences. The broad organization is consistent with the circumplex model, with separate positive (appetitive) and negative (aversive) areas, and a separate arousal area, along with other forms of specialization. BLA = basolateral amygdala; CNA = central amygdala; PAG = periaqueductal grey; LH / MH / DMH = lateral / medial / dorsomedial hypothalamus.

critical role in driving our emotional experiences. Finally, *Stanley Schacter* and *Jerome Singer* argued that both physiological and higher-level interpretational processes were both essential, with their *two-factor theory*.

Ultimately, all of these theories still emphasize that emotional states have both physiological and higher-level interpretational aspects, and the unique, interesting aspect of emotion is that it can activate the body in ways that purely abstract mental states do not. From a modern perspective, it is clear that many different brain and body responses occur essentially in parallel, producing our rich, complex, and fascinating subjective experiences of emotion.

One final issue concerns the optimal level of arousal for driving motivated behavior. The **Yerkes-Dodson law** (Yerkes and Dodson 1908) established the principle that there is an optimal level of arousal somewhere in the middle between low and high levels, following an **inverse-U-shape** curve. This same curve has been found for levels of dopamine as well. You may have experienced this in experimenting with different levels of caffeine – too much is actually not productive, as you get too hyper and unable to focus.

Summary of Key Terms

This is a checklist of key terms / concepts that you should know about from this chapter.

- Synaptic Plasticity
  - Hebbian Learning
  - Long Term Potentiation (LTP)
  - Long Term Depression (LTD)
  - 5 Steps of NMDA / Ca++ synaptic plasticity
  - What determines LTP vs. LTD direction of synaptic plasticity?
  - Error backpropagation
- Classical Conditioning
  - Conditioned stimulus (CS)
  - Unconditioned stimulus (US)
  - Un/conditioned response (U/CR)
  - Dopamine responses to CS, R, no-R in conditioning expt
  - Rescorla-Wagner learning rule / reward prediction error model of dopamine
  - Extinction, and its context sensitivity: spontaneous recovery, reinstatement, renewal
- Operant / Instrumental Conditioning
  - Thorndike's law of effect
  - Implementation thereof in terms of dopamine effects on Go / NoGo
  - Actor / Critic model
  - Partial reinforcement and gambling
  - Shaping to build up complex behaviors
  - Primary and secondary reinforcers
- Motivation

  - External (extrinsic) vs. internal (intrinsic) motivation
  - Drive reduction
  - Maslow's hierarchy of needs (levels of hierarchy)
  - role of vmPFC in satiety / devaluation effects
- Emotion and Arousal
  - Circumplex model
  - Six basic emotions according to Ekman's original faces
  - Relationship between emotion and motivation
  - Importance of both physiological and higher-level interpretations for emotion
  - Yerkes-Dodson law

Chapter 6: Memory

Memory is the direct product of learning, so everything we learned in the previous chapter will help us understand how memory works. If you can remember it, of course. Some of the major questions that have been the focus of memory research include:

- What different kinds of memory are there?
  - Are there specialized brain areas for different kinds of memory?
- How long does memory last (for each different type)?
- What factors determine how well memories are encoded and recalled?

Thus, the study of memory has been focused on fairly practical and descriptive questions, befitting the essential role that memory plays in everyday life (and especially for students). Our memories are also a core aspect of our sense of self, and movies such as *Total Recall* (based on a Philip K. Dick short story, as so many good movies are) have explored this function of memories in provocative and interesting ways. Furthermore, by now most people have heard about the profound amnesia caused by damage to the *hippocampus*, e.g., from the famous case of *Henry Molaison* (*H.M.*) who had his hippocampus surgically removed and lost the ability to form new memories for the rest of his long, exceptionally well-studied life. The movie *Memento* artistically and accurately captures the subjective nature of this condition, and is required viewing for anyone interested in memory (I personally have *two* copies, each a gift to my wife – memory certainly can be fallible). What makes the hippocampus so important for memory? What kinds of memory do *not* depend on the hippocampus? These are some of the important questions we will address in this chapter.

From Synapses to Memory

If memory is the direct product of learning, and learning is the direct product of synaptic plasticity as we learned in the previous chapter, then in principle *memory should be found in every synapse in the brain.* In fact, this is *true*, but it is also true that some synapses are more important than others. A deep understanding of memory requires reconciling these two perspectives on memory, and integrating some additional properties of neurons beyond their synapses.

First, it is useful to contrast the nature of memory in the brain with memory in a computer – it is tempting to try to use the computer as a simple analogy (as was especially popular in the early days of cognitive psychology, which relied extensively on computer analogies), but actually it doesn't work anything like that. In a computer, there are two major types of memory: RAM (random access memory), and a "hard disk" (which these days is typically more like RAM than the physical hard drives that were used for many years, but it still plays the same functional role). RAM is where *active* memories reside – the stuff the computer is currently working on. Elements from RAM are read into the central processing unit (CPU), processed, and then written back into RAM, often many times (e.g., when you are editing a document in your word processor, those

words live in RAM, and are accessed many times to redraw the screen as you edit). When you are done working on it, you save the memories from RAM to the hard drive, where they can reside essentially permanently. If the power goes off before you save, the RAM is lost – it is active (fast) and *temporary*, whereas the hard disk is slower but permanent.

In contrast, there is no fundamental separation between memory and processing in the brain. As we emphasized in the Neuroscience chapter, processing in the brain is distributed across all of the billions of neurons in the brain, with each neuron playing a small role within larger networks. Each neuron is detecting some particular patterns as a function of its synaptic connections, helping to compress and simplify the vast stream of information flowing through the brain. Thus, the direct effects of any given synaptic changes depend critically on where in the brain those synapses are, and what kinds of information processing those neurons are doing. For example, synaptic changes in occipital cortex should affect visual processing much more than decision making, while synaptic changes in the frontal cortex and basal ganglia should affect decision making much more than vision. Below, we'll see how synaptic changes in the hippocampus come to be so important for so much of what we generally think of as "memory".

The functional equivalent of RAM for a computer is not so obvious in the brain: without a CPU, there is no need for quickly reading and writing information from a RAM-like memory system. Instead, everything the neuron needs to carry out its detection and compression function is right there in its synapses, and learning directly modifies these synaptic connections. Furthermore, as we saw in the last chapter, this learning is long-term (i.e., long-term potentiation, LTP, and long-term depression, LTD), so it seems that memory in the brain is typically long-lasting, unlike RAM. And yet, we have all had the experience of suddenly forgetting what we were just talking about, or entering a room with a clear purpose, which has just vanished into thin air. These seem like distinctly RAM-like properties. What kind of stuff in the brain are these experiences made out of?

The answer is: *neural firing* – the ongoing spiking activity of the vast numbers of neurons in your brain that are currently above-threshold and sending their signals to other neurons. Indeed, this **neural activity** is an essential additional contributor to memory, and even though it is fundamentally different from RAM in terms of the underlying function of the brain, it nevertheless has some properties in common with RAM, in terms of being relatively *transient* and, well, *active*. Furthermore, we'll see that the frontal cortex / basal ganglia system is uniquely capable of sustaining patterns of neural activity over longer durations, and thus corresponds to a kind of RAM-like system called **working memory**. Thus, as often happens in biology, the same functional properties (active, transient memory vs. slower, permanent memory in this case) can emerge from very different underlying mechanisms. Furthermore, we'll see in the next chapter that even though the brain is nothing like a computer at the level of individual neurons, it does behave somewhat like a computer at the larger-scale systems level, where the RAM-like working memory system plays a critical role,

but we'll postpone consideration of this level until then.

In summary, we'll start our investigation of memory with the following principles derived from neuroscience:

- Memory can be broadly defined as *any* form of persistence of information over time in the brain – any trace of some prior event can be considered a type of memory.
- Neurons have two primary sources of such persistent information:
  - **Activity** in the form of ongoing spiking, electrical potentials underlying that spiking, and the chemical states of other parts of the neuron, which are *transient* – once a neuron stops firing and its other electrical and chemical states dissipate, a memory trace is no longer actively present in that neuron.
  - **Synaptic changes** from learning, which are relatively *long-lasting*, and change what kinds of input signals will activate the neuron in the future (i.e., what it *detects*).
  - These two aspects of neural memory directly influence each other, because learning is driven by neural activity, and changes in synapses result in different patterns of neural activity, but despite this interdependence, we can see how each plays more of an essential role in different types of memory.
- The specific *content* of the memory supported by any given neuron and its synapses is a direct function of its role within the larger neural networks of the brain – memory happens everywhere in the brain at all times, directly within content-specific processing areas (e.g., visual memories in visual cortex, etc).

Finally, there is one more critical constraint from neuroscience, having to do with the widely-used concept of *transferring* information from one part of the brain to another. As noted above, this is how everything works in a computer (information is constantly being transferred among the different components of RAM, CPU, and hard disk), but information in the brain is not encoded *symbolically* as it is in a computer, and therefore cannot be so easily moved around. Instead, as we've emphasized repeatedly, each neuron has learned to detect patterns of activity in its inputs, and thus information can only be transferred by neurons in another brain area detecting their own version of the information encoded in a given brain area. In other words, information transfer in the brain is much more like the game of *telephone*, where a given message is passed from one person to another, often resulting in amusing misunderstandings (Figure 6-1). The same thing happens in the brain: information transfer is *always* accompanied by fundamental transformations of the content, with each area adding its own *spin* or interpretation, with important consequences for understanding the relative veracity of memory.

Figure 33: Fig 6-1: The *telephone* game, which is an apt metaphor for how information is transferred in the brain. Neurons, like people, take a given signal, interpret it in their own particular way (as a function of their synapses), and send out their own interpretation. The idea of direct symbolic information transfer as in a digital computer does not apply in the brain.

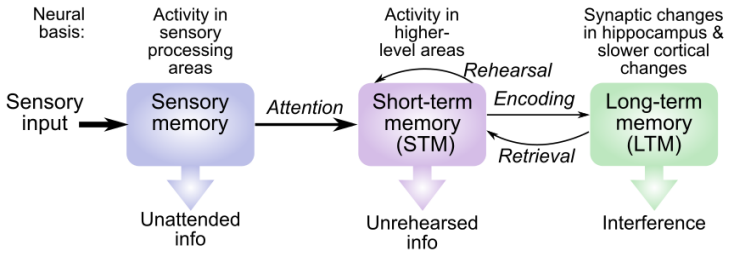


Figure 34: Fig 6-2: The modal model of memory and its neural basis, in terms of neural activity and synaptic changes. At each step, processing is required to transition to the next: only attended sensory items enter STM (and the rest is lost), and actively encoded STM information enters LTM. Active rehearsal sustains information in STM. Information in LTM can be retrieved back into STM, and is lost primarily via interference.

The Modal Model of Memory

Figure 6-2 summarizes the **modal model** of memory, which is so-named because it summarizes the common elements of many different models of human memory that had been developed in the early part of the cognitive revolution (Atkinson and Shiffrin 1968). It does a good job of capturing many different phenomenological aspects of memory, and we can use it to see how the neural principles play out in practice. It involves three separable components, *sensory memory*, *short-term memory (STM)*, and *long-term memory (LTM)*, with information flowing from one to the next dependent on relevant active processes including *attention* (sensory memory to STM) and *encoding* (STM to LTM).

First, sensory input activates **sensory memory**, which is characterized as a transient, high-capacity memory system that represents the sensory input at various levels of abstraction. Sensory memory corresponds largely to the *activity* of neurons that have been stimulated by the sensory input, at various levels along the kinds of hierarchically-organized sensory processing pathways discussed in the Neuroscience chapter.

There are different names for this activity within each modality, including **iconic** memory in the visual pathways, and **echoic** memory in the auditory pathway. Iconic memory generally persists for less than a second, whereas echoic memory lasts longer, up to about 4 seconds. These differences reflect the extent to which the neural activity in associated visual or auditory brain areas can persist. Because auditory information is inherently transient and evolving over time, the brain has extensive subcortical mechanisms that integrate and preserve these auditory signals over time, resulting in its longer persistence.

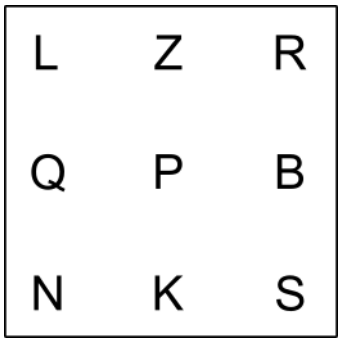


Figure 35: Fig 6-3: Sperling's sensory memory task. In the *full report* condition, participants attempted to retrieve all items, and typically only recalled about 4.5 on average. In the *partial report* condition, an auditory cue presented after a variable delay indicated which of the three rows to recall. For delays less than a second, they could accurately recall the letters within the cued row, indicating the presence of a high-capacity sensory memory trace (iconic memory) that decays within a second if not activated into STM via attention.

Classic experiments by *George Sperling* and others established these duration values, by flashing a display with 3 rows of 3 letters each (Figure 6-3), and probing people to report a particular row from the display after variable delays (Sperling 1960). In this *partial report* condition, people were generally able to report the information within about a second, but not longer. Critically, the relatively large amount of information in the full display was above people's capacity to encode in its entirety (as established through other *full report* conditions where they had to try to recall all of the letters), so the partial report cue allowed them to focus attention on one row, resulting in the activation of corresponding representations in STM. However, once the sensory memory trace fades, it is gone, and cannot be "transferred" to STM.

Experiments such as these also established the next step of the modal model, which is more strongly capacity-limited, but longer-lasting, and is referred to as **short-term memory** (**STM**). Only information within the focus of **attention** makes the jump from iconic or echoic sensory memory into STM, and given the capacity constraints, attention can only grab about 3-4 "items" into STM from sensory memory (corresponding to a single row from the Sperling task). From a neural perspective, STM corresponds to neural activity in higher levels of the neocortex (in temporal and parietal lobes) that have more highly compressed encodings of the sensory input. Thus, as noted above, the "transfer" of information from sensory memory to STM results in a significant compression and abstraction of the original signal. The ability to uniquely activate these compressed, abstract detector neurons in higher brain areas requires attention to filter the lower-level sensory input, thus explaining both the need for attention and the lower capacity of STM relative to sensory memory.

Furthermore, the smaller capacity of STM enables it to persist for longer periods of time, because more neurons across multiple of these higher-level areas can participate in representing this information, resulting in a more redundant and robust collection of such neurons. In the terminology from the Consciousness... chapter, STM corresponds to the fully recurrent activated state, which is highly likely to be the subject of conscious awareness. Indeed, one of the defining characteristics of STM is that you are consciously aware of it. Thus, the overall picture of STM is that the underlying neurons are mutually activating each other via bidirectional excitatory connections, causing a bit of an "echo chamber" as these spiking signals pass back and forth among these neurons, resulting in a longer-lasting activation trace compared to sensory memory. Rough estimates of the duration of STM extend up to about 30 seconds, but this is strongly dependent on the process of **maintenance rehearsal**, which involves the deliberate attempt to keep those neurons firing robustly by continuously focusing attention on them.

Interestingly, up to this point, the modal model only includes memory mechanisms based on neural activity. This reflects the fact that the synapses in the sensory pathways have been very well-trained by the time anyone is participating in Sperling-style experiments, so the synaptic changes there typically don't make much of a noticeable difference. However, there is an extensive

literature on *perceptual learning* which can reveal the effects of these ongoing synaptic changes. Thus, as noted above, memory really is happening at every synapse in the brain, whenever activity is sufficient to drive synaptic changes. However, you sometimes have to try pretty hard to see the effects of these changes, and the modal model only covers the most obvious forms of memory.

Finally, the last component of the modal model introduces a form of memory that does depend on synaptic changes, in the form of **long-term memory** (**LTM**). In the terms of the modal model, memories are "transferred" into LTM from STM through the process of **encoding**. They can also be recovered back from LTM into STM via **retrieval** processes. This model was developed during the 1960's, when the computer metaphor was at its height, and this encoding process was typically envisioned as transferring "files" between the RAM-like STM and the hard-disk of LTM. But what does this correspond to in the brain, given that we don't think the concepts of RAM, hard-disk, or transfer really apply in the brain?

The Hippocampus

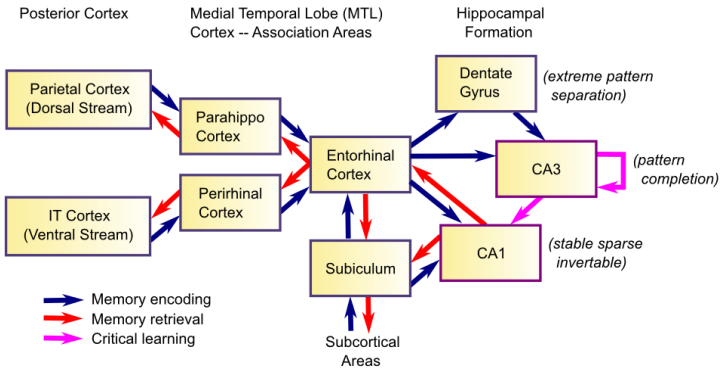


Figure 36: Fig 6-4: Connectivity and structure of the hippocampus. Sensory memory and STM are supported by activity in the posterior cortex areas, which then feed into two cortical areas in the *medial* (middle) region of the temporal lobe, the *parahippocampal* and *perirhinal* cortex. These then feed into the *entorhinal* cortex, which thus has a maximally *compressed* encoding of everything active in the rest of the brain. The areas of the hippocampal formation then effectively take a snapshot of this cortical activity.

This is where the *hippocampus* makes its grand entrance on the memory scene: in most cases, the initial encoding of information from the active state of the cortex (i.e., STM) into a form that can be later retrieved (i.e., LTM) depends on the hippocampus. As noted in the Neuroscience chapter, the hippocampus sits "on top" of the neocortical hierarchy of areas, and can quickly take a "snapshot" of the current pattern of activity across the upper layers of the cortex (Figure

6-4). Thus, the unique anatomical position of the hippocampus, plus some important special properties of the hippocampus itself, enable it to play such a critical role in the encoding and retrieval of memories.

In brief, you can think of the hippocampal neurons as *detecting* the elements of a memory (e.g., the *who, what, where* elements of an event or *episode*). Synaptic changes in these neurons then enable even a subset of those elements (e.g., the query "what did you have for dinner last night?") to re-activate these same neurons in the hippocampus. When these neurons fire, they act in turn to re-activate the memory out in the neocortex (i.e., the *retrieval* arrow between LTM and STM in the modal model, Figure 6-2). Thus, whereas neurons in the visual pathways are detecting objects and object features, neurons in the hippocampus are detecting *memories*, and that is why they play such a central role in our mnemonic life.

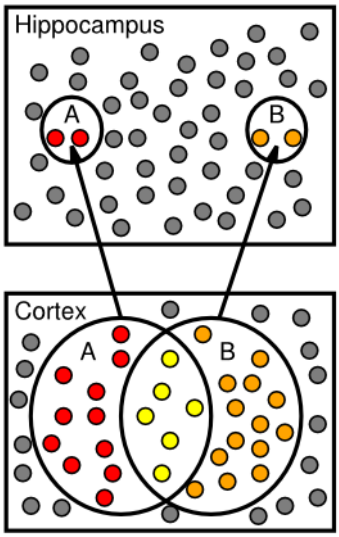


Figure 37: Fig 6-5: Pattern separation in the hippocampus: overlapping patterns of neural activity in the cortex result in separate, non-overlapping patterns in the hippocampus, because it has *sparse* activity (i.e., very few neurons active).

Figure 6-5 shows one of the magic tricks used by the hippocampus to be able to rapidly encode new memories without overwriting other existing memories, known as **pattern separation**. This is the key idea developed by *David Marr* as mentioned in the Neuroscience chapter, which applies to both the hippocampus and the cerebellum (Marr 1969; Marr 1971). The idea is that if you simply reduce the number of neurons firing in the hippocampus compared to the cortex (i.e., make them *sparse*), then the patterns of activity in the hippocampus will overlap much less than those in the cortex, and therefore, there will be less overlap or interference in the synaptic changes involved in

memory encoding. Mathematically, this derives from the fact that squaring a small number, such as .01, results in a *much* smaller number (.0001) – the small number is the probability of a neuron getting active, and the square is the resulting probability that it would be active in two different memories. More realistic, detailed simulations of the hippocampal circuit confirm this basic principle (R. C. O'Reilly and McClelland 1994).

There are many important implications of this pattern separation property. First, as we noted in the Neuroscience chapter, this results in a kind of *brute force* memorization strategy in the hippocampus. It doesn't try to make any direct connections between related memories – instead it just effectively sticks each memory in its own separate "box". This is great for quickly finding a place to stick a new memory, but it means that the hippocampal version of those memories is a completely disorganized, haphazard pile of these separate boxes. Thus, a major further process in memory involves a much slower process of trying to organize and systematize all those memories, known as **memory consolidation**. Specifically, memories that are initially encoded in the hippocampus are gradually incorporated into synaptic changes among neurons in the neocortex, resulting in the formation of more systematic, well-organized **semantic knowledge** (McClelland, McNaughton, and O'Reilly 1995). Some of this consolidation may take place during sleep, as memories are replayed during dreams (Wilson and McNaughton 1994; Buzsáki 1989; Roumis and Frank 2015), and much of it certainly depends on the usual retelling and ruminative replaying of memories throughout the course of daily life.

Hippocampal pattern separation and memory consolidation have major implications for educational learning and expertise. Everything you learn in class is initially encoded through hippocampal brute-force memorization, and only over a relatively long period of repeated learning and practice does a systematic an *productive* form of semantic knowledge emerge. This is consistent with how much experience it takes to become an expert in a given domain: roughly 10,000 hours or 10 years (Ericsson and Lehmann 1996). Thus, if you really want to master something, be prepared to spend a long time slowly shaping your neocortical synapses to develop the necessary systematic knowledge base.

Another important implication of pattern separation is the canary-in-a-coal-mine nature of the hippocampus. Driving down the activity level of the hippocampus requires an extensive amount of GABA inhibition, and thus the hippocampus is extra sensitive to the effects of alcohol and benzodiazapines (e.g., valium, midazolam), which are GABA agnonists as discussed in the Neuroscience chapter. Furthermore, the rapid rate of learning in the hippocampus requires high levels of NMDA receptors, which makes this system susceptible to epileptic seizures due to the development of over-strong excitatory synaptic connections (recall that H.M. had his hippocampus removed due to epilepsy, which often has a hippocampal source). Both of these factors may contribute to a heightened sensitivity to oxygen depravation.

Pattern separation also has important implications for the retrieval of memories from the hippocampus. To the extent that it is always trying to

keep different patterns separate, it is then hard to take a partial retrieval cue (e.g., the "what did you have for dinner?" question) and have that re-activate the original pattern of neural activity that was present when the memory was originally encoded. This retrieval process is called **pattern completion**, as it involves filling-in or completing the partial cue pattern. Instead of doing pattern completion, the hippocampus might just end up encoding a retrieval attempt as a brand new experience, and activate entirely new neurons as a result of pattern separation. Thus, pattern separation and pattern completion are essentially opposing forces within the hippocampus. Pattern completion is supported by special connections within one of the main areas of the hippocampus (the *CA3*), which effectively "glue" together the different elements of a memory. Detailed analyses of the battle between pattern separation and pattern completion suggest that the specific anatomy of the hippocampus is particularly well-suited for balancing between these competing demands (R. C. O'Reilly and McClelland 1994).

Taxonomy of Long-Term Memory

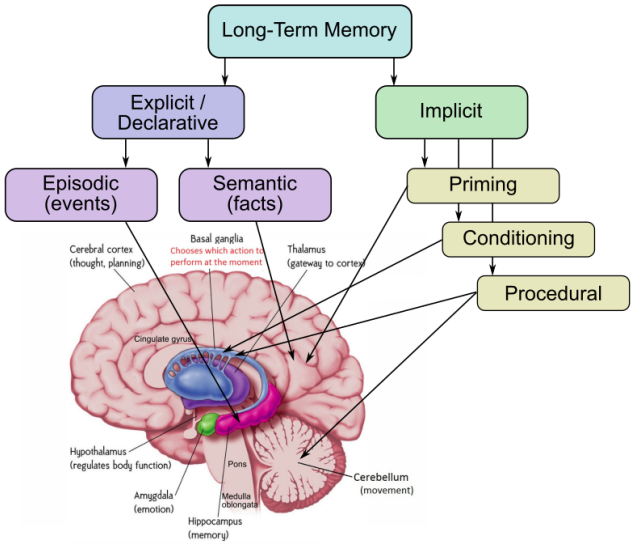


Figure 38: Fig 6-6: A standard Long-Term Memory taxonomy, and associated neural substrates. The broadest distinction is between consciously accessible memories supported by the cortex and hippocampus, versus non-conscious memories largely in subcortical areas. Priming is an interesting case of a cortical memory effect that is not directly accessible to consciousness.

Although the hippocampus plays a dominant role in the process of encoding new memories into LTM from STM (in the terms of the modal model), synaptic

changes occur everywhere there is activity in the brain. Thus, a major focus of memory research has been attempting to document and organize all these different "types" of long-term memory within overall memory **taxonomies** (akin to the taxonomies used to organize different species of animals, for example). Because memory and processing are both occurring within each and every neuron, these taxonomies are really just descriptions of the different kinds of processing taking place in the brain, which we reviewed in detail in the Neuroscience chapter. Nevertheless, we will briefly review the most popular taxonomy here, and the research that went into its construction.

Figure 6-6 shows perhaps the most widely adopted LTM taxonomy, proposed initially by *Endel Tulving* (Tulving 1972) and refined by *Larry Squire* (Squire 1992). It features a top-level division between **explicit** or **declarative** memory, as contrasted with **implicit** or non-declarative memory. Explicit memories are those we can have direct conscious access to (and declarative means you can declare it verbally), while implicit memories are not consciously accessible. Given what we know about consciousness from Chapter 3, explicit memories are therefore those in the neocortex. Interestingly, we likely are not directly conscious of hippocampal memories, given the requirement of recurrent / bidirectional connectivity for consciousness, which is only partially present in the hippocampus. Instead, we become conscious of hippocampally-supported memories when they are recalled back into cortex.

The two major subtypes of explicit memory are hippocampal **episodic** memories (i.e., the memories of all the daily events and episodes in our lives, and those we read about or watch in movies or on TV), and **semantic** memory, which is a summary term for all of the facts and knowledge we have, which has been integrated into our cortical synapses over many years of memory consolidation. During this consolidation process, the episodic character of the knowledge gets winnowed away, leaving only the bare knowledge devoid of the **source** or **context** information about where we learned these facts. Newly learned facts (e.g., much of what you are learning in this course) still retain their episodic trace – you can probably recall when you heard about something interesting for the first time in lecture, or read about it in a book. Sometimes, people feel like they have a particularly clear sense of where on the page they read something, but in my experience this has proven illusory more often than not.

Within the much more diverse umbrella of implicit memories, there are *procedural*, *conditioning*, and *priming* memory traces. The separability of **procedural** memory from hippocampal episodic memory was vividly demonstrated by H.M., who was able to learn a challenging new procedural task such as learning to trace a picture when looking in a mirror (try it – it is hard!) at the same rate as neurologically intact control participants. This is because procedural learning depends on the cerebellum and basal ganglia, not on the hippocampus. Likewise, as we reviewed in the Learning chapter, **conditioning** depends on the amygdala, basal ganglia, and dopamine systems, and is thus separable from hippocampal and cortical memories (and was also intact in H.M.).

The value of this memory taxonomy is debatable. Really, it is just assigning new labels for the functions of brain areas, which can be much more richly and accurately described (e.g., as in the Neuroscience chapter) than in such a broad taxonomy. Furthermore, it is missing many important parts of the brain. Perhaps most importantly, the central division according to the criterion of conscious access is problematic at many levels. Consciousness is inherently subjective, and putting a subjective construct at the center of a major theoretical framework jeopardizes the entire enterprise. Furthermore, it immediately eliminates application to animal memory (R. G. M. Morris 2001), as the notion of consciousness in animals is certainly fraught with controversy. It also unnecessarily complicates any kind of straightforward understanding of memory in terms of underlying neural mechanisms.

For example, given our detailed understanding of how the hippocampus works, it is highly likely that even rats (which have a large hippocampus relative to the rest of their brain) encode something like episodic memories of all the different experiences in their lives. Rats likely don't sit around idly reminiscing as people do, but that doesn't mean they don't re-activate their episodic memories in response to relevant stimulus cues – indeed, this has been demonstrated in many experiments recording from hippocampal neurons in rats. Thus, it is more productive to find the many parallels in brain systems across species, so that we can integrate a much broader scope of data into our theories of memory.

The case of **priming** is particularly illustrative of the limitations of a consciousness-based framework. Priming is the measurable facilitation in processing information that was previously processed (e.g., "priming the pump"). It results from the small synaptic changes throughout the neocortex, driven by neural activity. Thus, although we are not directly conscious of priming itself (e.g., we don't know that our responses are faster by about 10 msec), we *are* typically conscious of much of the activity that drives priming. And these are the very same synaptic changes that add up over time to produce new semantic memory learning. So does it really make sense to put this in the implicit memory category? Another example is the considerable contributions of the frontal and parietal cortex to procedural tasks: we can certainly be aware of activity in these brain areas, and yet they are put in the implicit category.

Another interesting case similar to priming is the difference between **recognition** and **recall**, which has been studied extensively in the memory literature (Jacoby, Toth, and Yonelinas 1993). Recognition memory is characterized as using the overall feeling of *familiarity* with a given stimulus to decide whether it was on a given memory list, whereas recall involves the explicit, conscious *recollection* of episodic details from the time of study. Recollection generally depends on the hippocampal pattern completion process to re-activate those episodic details, whereas familiarity can be supported by differences in neocortical activity patterns reflecting synaptic weight changes, which are not strong enough to drive full recollection (K. A. Norman and O'Reilly 2003). Thus, familiarity is similar to priming, but interestingly, H.M. and some other severe amnesics were impaired at familiarity-based recognition memory, but their priming was

intact. This is because the familiarity signal is likely driven by the neocortical areas surrounding the hippocampus (e.g., perirhinal and entorhinal cortex) that were damaged along with the hippocampus proper, whereas most priming tests probe lower-level semantic or visual cortical areas.

Amnesia

Patients with hippocampal damage such as H.M. have also shown us that two different types of **amnesia** (loss of memory function) can be *dissociated* (i.e., separated, do not always co-occur): **retrograde** vs. **anterograde** amnesia. Retrograde refers to memories of the past (like "retro" styles etc), while anterograde refers to the ability to form new memories. H.M. was profoundly impaired in his ability to form new memories, and thus suffered from severe antergrade amnesia. However, many of his memories from his more distant past were largely intact, meaning that he had comparatively mild retrograde amnesia. Furthermore, his basic semantic knowledge of facts etc was largely intact.

We can understand this dissociation in terms of the basic explanation of hippocampal function given above. The hippocampus is critical for rapidly learning new episodic memories, because of its unique position at the top of the cortical hierarchy, and its special properties including pattern separation and a relatively fast learning rate. Thus, damage to the hippocampus almost always produces significant impairments in encoding new episodic memories. However, because of the gradual incorporation of episodic memories into the neocortex through the consolidation process, older memories from the past can still be recalled even without the help of the hippocampus.

Interestingly, consolidation predicts that more recent memories leading up to the point of hippocampal damage should be most impaired, as they have had less time to be consolidated into the neocortex. This *gradient* of retrograde amnesia is often observed at least to some extent in human amnesics. Interestingly, extensive investigations of retrograde gradients and memory consolidation in rats have produced inconsistent results, likely reflecting the variability in the extent to which rats actually recall prior episodes, across different experimental paradigms (Sutherland, O'Brien, and Lehmann xx 2008; Anagnostaras, Maren, and Fanselow 1999).

Another fascinating form of amnesia is **childhood amnesia**, which is the widely-documented phenomenon that people cannot generally remember anything before about 3 years of age. Go ahead, give it a try – can you? Many studies have attempted to understand the reasons for this amnesia (Hayne 2004). Overall, it is likely a result of fact that the neocortex is not sufficiently well organized before that age, to support the ability of earlier hippocampal snapshots to be translated back out into the cortex. In effect, the "language" that the earlier snapshots were recorded in is no longer something that the more mature brain can understand (and indeed language learning itself likely plays a significant role).

Although the neocortex continues to develop and learn in significant ways

beyond the age of 3, there is presumably just enough stability for those earliest memories to persist. And those early memories that you can still recall have likely been recalled, reinforced, and elaborated many times in the ensuing years, so they are well-consolidated and may not actually be very accurate anymore. Nevertheless, in my own case, I feel like I do have vivid, first-person memories of my 3-year-old-self living for 6 months on the island of Grenada in the Caribbean, including a scary encounter with a large crab behind the house. Thus, as is the case with memory in general, emotional arousal and the relative novelty of experiences play a large role in one's ability to later recall them.

Memory Capacity and the Importance of *Chunks*

One of the dimensions along which memory systems vary is in terms of their capacity, with sensory memory being high capacity, STM having a strongly limited capacity of around 3-4 items, and LTM being essentially unlimited in its overall storage capacity. But any consideration of capacity raises the central question of *what counts as an* item *for the purposes of measuring capacity?* In the Sperling experiments (Figure 6-3), items were individual letters, but what if we instead put words where the letters were in the 3x3 grid display? Memory capacity will be about the same, now measured in words instead of letters, but that represents a considerable increase in overall *letter* memory capacity!

The answer to this puzzle is to introduce the concept of a **chunk**, which is somewhat circularly defined as an element that acts like a single item with respect to memory capacity measurements. If the stimuli are *random* letters, then each letter is a chunk, but if the letters can be formed into words, then the word becomes the chunk. Likewise, if words can be combined into sensible sentences, then those sentences become the chunks. In short, a non-circular definition of a chunk is *anything that we have an existing stable neocortical semantic representation of.* This is still not very precise, but it will do for now.

Based on an influential and provocative article by *George Miller* (G. Miller 1956), many textbooks incorrectly cite the capacity of STM as *7 plus or minus 2*. However, Sperling's original data, and data from many other tasks and domains, strongly suggests that it is actually **the magic number 4** (Cowan 2001; Luck and Vogel 1997). Although overly simplistic, one way of thinking about this is that each cerebral hemisphere can hold 2 items when pushed to the limit (2x2=4), with 1 being much more comfortable (Buschman et al. 2011). The higher capacity of 7 applies only to verbal memory that can be sustained by a rehearsal mechanism known as the *phonological loop*, where you repeatedly verbalise (in your mind, but also using your actual vocal muscles at a subthreshold level) the to-be-remembered material (A. Baddeley, Gathercole, and Papagno 1998). Our extensive experience with verbal material presumably produces this larger capacity beyond what is generally available with the "default" neural mechanisms.

Encoding and Retrieval Strategies (i.e., How to Study!)

Because memory capacity is determined by the availability of appropriate chunks, one major category of memory-enhancement tricks involves creating new chunks, and efficiently leveraging the ones you already have. This is the main trick employed by contestants in the memory olympics competitions, and was well-documented in the case of an individual, S.F., who developed chunking strategies that allowed him to remember over 100 random digits (Ericcson, Chase, and Faloon 1980). In this case, he turned 3-digit numbers into times to run a mile or other standard distances, as S.F. was an avid runner. Another common example of chunking is the creation of acronyms. For example, the "big five" personality dimensions that we'll encounter later can be organized into the acronym *OCEAN*, which then makes it much easier to remember them all.

Another effective encoding / chunking strategy (i.e., **mnemonics**) is to associate different words with different familiar spatial locations, known from the days of ancient Greece as the **method of loci**. An even more **elaborative encoding** strategy is to create stories involving these locations and familiar people (e.g., "my mom went from the living room to the kitchen, to get a popcorn snack"), where each of the words then stands for something that you're trying to remember (e.g., mom = *memory*, living room = *hippocampus*, kitchen = *neocortex*, and popcorn = *semantic memory*). Because the hippocampus really loves to encode episodic memories, these episodic chunks are particularly effective and memorable.

Several other related principles of effective memory encoding have been developed. For example, the influential **levels-of-processing theory** (Craik and Lockhart 1972) postulates that more *deeply* encoded information will be better remembered. The notion of levels or depth here corresponds to the levels of processing in the neocortex, going from raw sensory information up to higher-level semantic information. For example, many studies have found that encouraging people to think about the meaning of a word, as compared to noticing the case or font of the letters, results in better memory.

An interesting example of the benefits of deeper, more elaborative encoding comes from the notion of **desirable difficulties** (Bjork 1994) – memory is often better if you have to work harder to process the information, even in sometimes fairly strange ways. For example, making information harder to read can improve subsequent memory, and a font was recently created called *Sans Forgetica* to leverage this finding (see: `http://www.sansforgetica.rmit/`). Another example is the **testing effect**, where taking a test improves subsequent memory. Thus, you should always test yourself on the key words listed at the end of each chapter, and hopefully your class includes weekly quizzes that give you an opportunity to test yourself.

One of the best ways to really learn something is by teaching it to others, which is a version of the **generation effect** – having to produce a sensible explanation of something greatly improves comprehension. Try cornering a friend and give them a mini-lecture on how memory works in the brain – you'll

soon find the gaps in your understanding, and strongly reinforce the parts you already do understand. Seriously, if you want to learn, teach! This is one of the most important synergies in academia: by having to explain what we've learned in our research through teaching, professors then understand it all much better.

One of the most fascinating encoding principles is the **encoding specificity principle** (Tulving 1983), which reflects the fact that episodic memories tend to bind together all of the different elements present when a memory is encoded, and thus recall of those memories will be best when those original elements are present at the time of recall. This is a direct result of the pattern completion vs. pattern separation battle operating in the hippocampus – if too many elements are different from the original event, the hippocampus tends to perform pattern separation instead of the pattern completion required for recall. All those random elements present at the time of encoding are typically summarized with the term **context**, and thus lead to the **context-dependent memory** phenomenon. A classic example of this phenomenon is that people are better able to recall information when tested in the same physical context as it was originally learned – for example, if you study in a library, then taking a test in that same library will generally result in better performance.

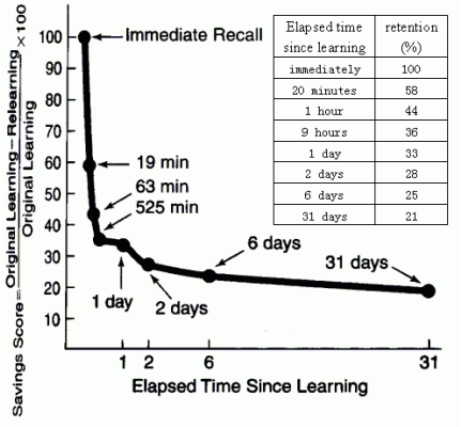
The most famous demonstration of this encoding specificity / context-dependent memory principle was conducted in a study where items were learned either on a beach or underwater using scuba equipment, and then tested either in the same or different context (Godden and Baddeley 1975). Participants in the same-context conditions (either on land or under water) performed significantly better than in the cross-context conditions. Another notorious demonstration involved study and test either drunk or sober, which again found that, surprisingly, testing while drunk was better than sober *if* initial learning was drunk (Goodwin et al. 1969). This has been labeled the **state-dependent memory** effect, and presumably reflects the same encoding specificity principle. It is important to emphasize that memory was much worse when learned drunk, even when tested drunk, so that is *not* a good strategy overall. Other demonstrations of state-dependent memory involve mood states, and we'll see later that this **mood-dependent memory** effect creates an unfortunate feedback loop in depression, where you're much more likely to remember all the bad memories in your life when you're depressed, making everything seem that much more bleak. On the bright side, this also works for positive memories in positive mood states.

Another critical way to improve encoding is to use **spaced** instead of **massed** practice – i.e., to space out your studying over multiple separate study sessions, instead of *cramming* at the last minute. This is beneficial for the same reason that gives rise to context-dependent effects: spacing out study causes the information to be learned across multiple different contexts, and thus helps to make the knowledge more independent of that context. A critical point here is that *context* includes a significant contribution from *time* – your internal mental state is constantly evolving over time, and will be significantly different a week or two from now (Howard and Kahana 1999). Thus, even if you study in the same physical context, your internal mental context will be different, shaped by

all those synaptic changes taking place between the two study sessions. Indeed, a critical aspect of the memory consolidation effect involves exactly this process of thinking about the same issues from the very different perspectives that emerge as your brain changes over the period of years.

In short, you should study by engaging in deep, elaborative encoding of the material, connecting it in multiple different ways with your existing knowledge chunks, and testing yourself as much as possible, ideally by trying to teach material to others. Furthermore, you should do this in a spaced fashion, across multiple different days, ideally in different physical and mental contexts.

Memory Retention and Interference



Elapsed time since learning	retention (%)
immediately	100
20 minutes	58
1 hour	44
9 hours	36
1 day	33
2 days	28
6 days	25
31 days	21

Figure 39: Fig 6-7: The forgetting curve from Hermann Ebbinghaus's data. The initial steep dropoff is likely due to synaptic-level stabilization processes, and the longer plateau reflects essentially permanent long-term memory, with loss due largely to interference.

Even once you've successfully encoded some new information into LTM, it is still not safe! Memory is often a fleeting thing, as you have almost certainly experienced. Figure 6-7 shows the data from *Hermann Ebbinghaus* who pioneered the study of memory retention in the late 1800's (Ebbinghaus (1885) 2013). This curve is striking in its steep initial dropoff, followed by a relatively stable plateau. We can understand the nature of this curve in terms of two different processes, which have long been the subject of debate in the field: **decay** and **interference**. It is surprisingly difficult to distinguish these two on purely behavioral grounds, as it is generally impossible to prevent interference from happening, and decay is defined as an automatic, continuous process. However, detailed studies of the molecular processes following the synaptic plasticity events described in the Learning chapter allow for some resolution of this debate.

It is likely that the steep initial dropoff in memory is due to synapse-level

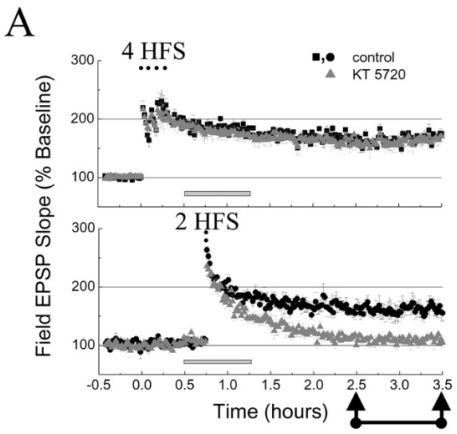


Figure 40: Fig 6-8: Forgetting curve from synapse-level stabilization effects, which shows a steep dropoff in synaptic strength for more weakly potentiated synapses (2 HFS curves in bottom graph; HFS = high frequency stimulation) compared to more strongly potentiated synapses (4 HFS, top graph). The KT 5720 curve shows the contribution of protein synthesis, which emerges over the period of an hour or so, and if these proteins are not available, the weaker memory decays back to baseline. From Alarcon, Barco & Kandel (2006).

processes, which can be considered a form of decay, but that after roughly a day or two, synaptic changes have stabilized to the point that subsequent forgetting is likely due primarily to interference effects. These synapse-level processes are collectively known as *synaptic consolidation*, which is distinct from the *systems consolidation* processes described above, where memory initially encoded in the hippocampus is learned in the neocortex as well. As shown in Figure 6-8, there is a roughly 15-20 minute period when synaptic changes can decay rapidly if they were not sufficiently strong in the first place, or reinforced by subsequent plasticity events (Alarcon, Barco, and Kandel 2006; Frey and Morris 1998). Over the course of an hour, synaptic changes are reinforced by processes that depend on new proteins being synthesized, including muscle-like *actin* fibers. Further stabilization occurs during sleep, over the next day or two (for all the details, see (Rudy 2013)).

After all of this synaptic consolidation has taken place, it is likely that further loss of memory is due to interference effects, which occur when new synaptic changes move the synaptic strengths in a different direction than was needed for an existing memory. This is known as **retroactive interference**, because it is interfering with older ("retro") memories. The extreme pattern separation in the hippocampus can help to minimize the amount of retroactive interference, by encouraging the use of distinct sets of synapses to encode different memories, but it is impossible to completely eliminate interference. An example of retroactive interference would be when you encode where you parked your car

today, versus yesterday. Because of the large amount of overlap in the overall context, you likely re-activate many of the same neurons involved in encoding these two memories. Thus, the synaptic changes that are made today will help you recall that it was parked in the South-West corner of the lot, but these changes will likely overwrite many of the synapses that encoded the "South-East" location from yesterday.

There is another form of inteference called **proactive interference** which is somewhat strange compared to the more intuitive nature of retroactive interference. In proactive interference, prior learning interferes with your ability to form *new* memories. This can happen if you are trying to learn new information about the same items over time. For example, if you use distinctive new items on every trial of a memory task (shapes, colors, letters, words, animals, etc), then it is easier to remember those items compared to re-using the same items repeatedly (Hasselmo and Stern 2006).

The Fallibility of Memory

In addition to failures of basic encoding and forgetting, there are other pitfalls in the domain of memory, which arise largely from the fact that the hippocampus only receives a highly *compressed* view of the outside world, filtered through many layers of cortical processing and compression as shown in Figure 6-4. Figure 6-1 of the telephone game also captures the kind of compounding effects that emerge from information propagation through the cortex. From this perspective it is a wonder that we can accurately remember anything at all! The current US president provides a teachable moment about the falliability of memory – as improbable as it might seem, it really does appear that his many difficulties with *facts* and the *truth* represent genuine delusions, and not just a Machiavellian level of strategic disinformation. Critically, people's delusions are often strongly shaped by their motivations and desires, so the practical difference between these two interpretations may not be that far apart.

In any case, the bottom line is that we *all* encode our memories through spectacles of one shade or another – we cannot see the world as it truly is, because perception is fundamentally a creative, active construction, as we discussed in the Perception chapter. Thus we are all susceptible to having **false memories**. One of the first demonstrations of this point in the memory literature was due to *Frederic Bartlett*, who tested people's ability to remember a story known as the "War of the Ghosts", over an extended period of time (Bartlett 1932). This story was based on Canadian Indian folklore, and contained many concepts and events that were entirely unfamiliar to the English participants in his experiment. As a result, the participants had great difficulty remembering the story, and ended up reshaping it to fit their own conceptual structures. These conceptual structures are called **schema**, and we'll revisit them again in the next chapter.

An important real-world implication of this strong tendency to *schematize* memory is in **eyewitness testimony**, where people are likely to encode the events of a crime according to their existing *stereotypes* and biases. Furthermore,

these biases can be activated by leading questions. For example, in one seminal study, the experimenters manipulated the use of leading terms like "smashed" in a car crash scenario, and this had large effects on participant's memory of things like the speed and damage involved (Loftus and Palmer 1974). Interestingly, as was the case in the Bartlett study, participant's confidence in their *false* memories was often higher than for their accurate ones.

The other major issue that has received considerable media attention is recovered memories of childhood sexual abuse. Unfortunately, abuse is all too common, but it is also the case that memory in young children is even more unreliable than in adults. Studies have shown that children can report having actually experienced events that they only imagined (S. J. Ceci et al. 1994), and some forms of therapy designed to uncover repressed memories may have used leading questions that could have created false memories.

In the experimental literature, false memory has been extensively explored using the *Deese, Roediger, McDermott (DRM)* paradigm (Deese 1959; Roediger and McDermott 1995). In this paradigm, a number of words that overlap strongly with a given target word (e.g., pillow, dream, night, etc) are studied, with the result that the target word ("sleep" in this case) is often confidently endorsed as having been on the study list. This is vidid demonstration that memory operates on high-level compressed semantic representations.

Working Memory and the Prefrontal Cortex

Finally, we conclude with one more important distinction between different types of memory, in this case between short-term memory (STM) and **working memory** (*WM*), which was proposed by *Alan Baddeley* and *Graham Hitch* (Baddeley and Hitch 1974). The notion of working memory resembles the functional properties of RAM in a standard computer: information that is currently being processed, maintained in an active, directly accessible state. Furthermore, this framework includes a *central executive* that functions much like a CPU in a computer. Working memory is distinguished from "regular" STM, where the latter includes just basic maintenance of information, whereas working memory is specifically about the information used for ongoing processing, which is particularly strongly maintained, even in the face of potential distractors.

As is often the case, the biology may provide a more precise definition of the difference between STM and working memory, in the form of robust sustained firing of neurons in the prefrontal cortex, which was discovered in the early 1970's (Fuster and Alexander 1971; Kubota and Niki 1971). This sustained neural activity was postulated as the neural basis of working memory (Goldman-Rakic 1995), and studies showed that this form of neural activity is indeed more robust and resistant to distraction than activity in posterior cortical areas (E. K. Miller and Desimone 1994). Computational models have shown that the basal ganglia can play a critical role in supporting this robust active memory in frontal areas, by dynamically switching the system between maintenance and rapid updating modes (R. C. O'Reilly and Frank 2006). Thus, overall there is ample biological

evidence that sustained neural activity in the frontal cortex is different from that in posterior cortical areas, in ways that accord with the overall distinction between working memory versus STM. We'll focus more on this frontal / basal ganglia working memory system in the next chapter.

Summary of Key Terms

This is a checklist of key terms / concepts that you should know about from this chapter.

- Neural mechanisms of memory:
  - Activity (spiking etc): fast, active and transient
  - Synaptic changes: slower, long-lasting
- Modal model
  - Sensory memory: iconic, echoic (high capacity, short-lived) – neural activity in sensory cortex
  - Short-term memory (STM): requires attention, limited capacity (magic number 4) – neural activity in higher cortical areas
    - Sperling task
  - Long-term memory (LTM): requires encoding – synaptic changes in hippocampus and cortex.
- Hippocampus
  - Antomical location on top of cortex
  - Pattern separation from sparse activity
  - Pattern completion to recall memories
  - Memory consolidation: semantic knowledge forms slowly in neocortex
- LTM Taxonomy
  - Explicit / Implicit
  - Episodic / Semantic / Priming / Conditioning / Procedural
  - Issues with consciousness
- Amnesia
  - Anterograde
  - Retrograde
  - Childhood amnesia
- Encoding / Retrieval Strategies
  - Chunk
  - Mnemonic
  - Method of loci
  - Elaborative encoding
  - Levels of processing
  - Desirable difficulties: Testing effect, generation effect
  - Encoding specificity principle
    - Context-dependent memory
    - State-dependent memory
    - Mood-dependent memory
  - Massed vs. Spaced practice (cramming is bad)

- Memory Retention and Interference
  - Decay: synaptic stabilization
  - Interference: Retroactive vs. Proactive
- Fallibility of Memory
  - False memories: War of the Ghosts
  - Schema
  - Eyewitness testimony & leading questions
- Working memory vs. STM
  - Robust firing in prefrontal cortex

Chapter 7: Thinking, Control and Intelligence

What is *smart*? This is the fundamental question for this chapter, with many profound personal and societal implications. Is there just one kind of smart, or are there multiple different forms of intelligence? How can we reconcile any form of *general* intelligence with everything we've learned up to this point, about how the brain works at a biological level? The brain is composed of billions of neurons, interconnected by vast networks of synapses, wherein all of our knowledge, and, presumably, intelligence, must lie. Do "smart" people have more neurons or synapses? Or, perhaps, *fewer* synapses? Are their neurons somehow fundamentally different from other people who measure as less smart according to standard intelligence tests? And what are those intelligence tests measuring anyway? Are they really some kind of "pure" measure of intelligence, or do they just reflect the degree of western-style education (and health and wealth) that a person has? What does your IQ score really tell us about you as a thinker, and about your prospects for future success in school and the real world? So many important questions!

If our brains were more like digital computers, these questions would have much simpler answers. It is relatively easy to measure the power and speed of a computer, and many people tend to think of human intelligence in these terms. As we discussed in the previous chapter, a computer has discrete parts (the CPU, RAM, and hard drive), and each of these parts can be directly quantified in terms of capacity and speed. If you're at all savvy about these things, you can obsess about getting the best value for your money along each of these dimensions, and, generally speaking, the faster the CPU and the more RAM and hard-drive storage, the more you can achieve with your computer. Computers really do come in obvious degrees of "smartness".

But our brains are nothing like that of a digital computer. Instead, cognition emerges out of the interactions of billions of chattering neurons, which are fundamentally shaped by learning processes over an extended period of time. As we will explore in the development chapter, we start out with virtually no discernible intelligence (despite how cute and special our parents think we are), and it takes most people a few *years* to even learn how to control their own bowels! Wow. The rest of the animal kingdom must think we are complete idiots, which comports with an amusing *Onion* headline to that effect.

Given that we clearly don't start out with much in the way of intelligence, it seems hard to escape the conclusion that intelligence is fundamentally a product of learning (in concert with other developmental / maturational changes). And this view is also hard to avoid when you think about all those synapses that need to get wired up in just the right way to produce whatever cognitive abilities we end up with.

So are "smart" people just better learners then? If so, what makes some people better at learning than others? When we explored this question in the Learning chapter, one of the major conclusions is that learning is driven fundamentally by *motivation*, and all that dopamine and related machinery that

gets us up in the morning and ready to pursue our daily goals, etc.

Indeed, we will review various sources of evidence that are consistent with the overall idea that motivational differences play an outsized role in determining measured level of intelligence. Of course, there are many, many complex factors that shape an individual's trajectory of learning and development, and motivation is itself a multi-faceted thing, so perhaps we aren't explaining too much when we say that motivation plays an important role.

But understanding the major factors shaping intelligence may affect how we think about ourselves, and others, in important ways. If we view intelligence as a product of learning and motivation, then it is more obviously malleable. This is the critical difference between a **fixed mindset** about intelligence, versus a **growth mindset**, as emphasized by *Carol Dweck* and colleagues (Dweck 2008), in an increasingly influential body of work. The growth mindset emphasizes that intelligence is not something that people "have", but rather, something they have to cultivate – something that grows over time. Increasingly, schools and teachers are recognizing that motivational factors have a huge impact on educational success, and they are developing innovative ways of motivating students to learn, and making the material more obviously self-relevant.

Fundamentally, the idea that intelligence is largely the product of time spent learning means that **anyone can learn anything**, if they only have sufficient motivation and time to invest into it. This open-ended, ambitious view of intelligence surely has the effect of opening up your individual horizons and sense of what is possible. Personally, I have always had this belief, and I have learned lots of complicated things, often slowly and with great difficulty. Eventually, things that once seemed impenetrable become just another familiar part of my mental toolkit. I have a very salient early memory of spending far longer than my peers figuring out how to simply connect a battery to some gadget in a summer school class as a kid. I felt like an idiot. But eventually, I figured it out, and learned this valuable lesson that, with sufficient effort, I could succeed.

Hopefully, you are now motivated to learn more about the history and current state of understanding about the nature of human intelligence, and the thinking processes that underlie it! We'll start off by exploring the core questions of what "thinking" is, and what kinds of brain mechanisms are particularly important for it. The conclusion from this may seem to contradict what was just said above: maybe we *do* have something like a CPU in our heads after all – except it is a CPU made out of neurons and brain systems, and it runs on dopamine! This is an important example of an *emergent* system, like the gears we talked about in the neuroscience chapter: the overall function of a CPU can be supported by various different "substances", just like the gears can be made of many different materials, and yet still function more-or-less the same.

Nevertheless, our neural CPU has major differences from a computer CPU, and the fact that it is made of neurons does have important implications for how it works. Indeed, one can understand a lot about the particular strengths and

limitations of human cognitive function, in terms of the overall idea that we can do both neuron-like computation, *and* something that approximates the function of a digital CPU. We have yet to develop powerful AI (artificial intelligence) systems that capture this unique combination of both forms of computation, and perhaps once we do, we will unlock the real magic of our brains!

After gaining a better understanding of the "mechanics" of intelligence, we'll review the history of thought about the nature of intelligence, and how it has been measured. Furthermore, we'll examine the data about the real-world implications of IQ test scores, and circle back to these big questions about the relationship between intelligence and motivation.

Another way of thinking about all of these issues, is in terms of the *control* component of our three-C's. Our neural CPU serves as a kind of overall control system for the rest of our brain, and, as we have emphasized, this is fundamentally a *motivated* form of control, focused on getting us the things we need and want, and avoiding all the bad stuff. Thus, the idea that motivation and intelligence are inextricably intertwined makes perfect sense from this perspective: the brain systems supporting our control systems (in the prefrontal cortex and basal ganglia) are the very same ones that directly interface with lower-level motivational and emotional pathways in the amygdala and dopamine system.

The Neural CPU in the Prefrontal Cortex and Basal Ganglia

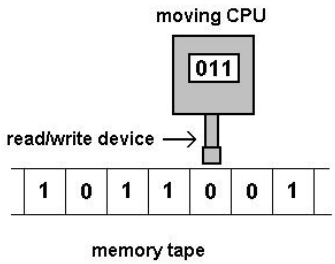


Figure 41: Fig 7-1: The components of a Turing machine: with just three basic components, any computation can be performed!

To understand what kind of neural machinery it would take to support CPU-like functionality in the brain, we start with the surprisingly simple mechanisms needed to make a computer work. At the most abstract level, *Alan Turing* and *John Von Neumann* worked out the basic principles of a *universal* computational device (something that could in principle do *anything*) in the 1930's and 40's (Turing 1936; von Neumann 1945). Amazingly, this device only requires three

essential components (Figure 7-1): 1. A way of reading and writing information from a memory system (conceptualized as a *tape* by Turing); 2. A *program* that determines how this information is transformed in between being read and written; and 3. Some *active* memory where things can be temporarily cached, for the program to refer to. These elements were elaborated by Von Neumann, in one of the most important unpublished papers of all time (von Neumann 1945), creating the foundation for modern digital computers. Now days, we take it for granted that computers can do almost anything, but this was just theory not so long ago.

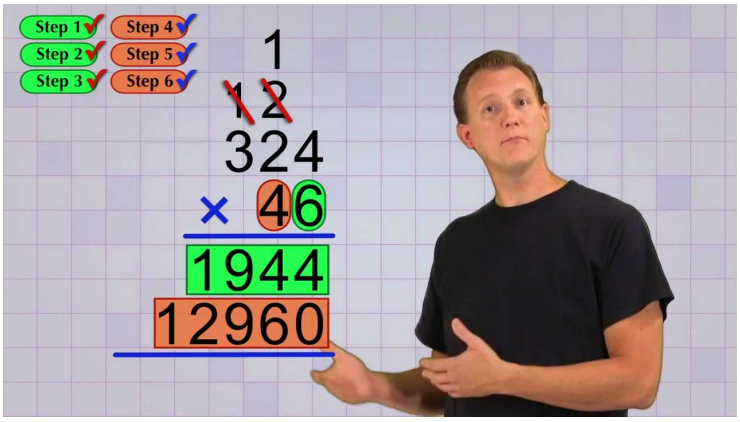


Figure 42: Fig 7-2: Computers solve problems by breaking them down into many small *sequential* steps, each one involving a specific, well-defined operation such as adding numbers, writing them down somewhere, and reading them back in for use later. Just like you do when performing multi-digit arithmetic. Alan Turing showed that these basic processes can be used to solve *any* problem.

You can get a good feel for how a computer works, and why it can do anything, by considering the traditional strategies for performing multi-digit arithmetic (Figure 7-2). Instead of just staring at those big numbers, you break the problem down into a sequence of simple, discrete steps. That sequence of steps is the *program* or **algorithm**, and each individual *operation* involves one of a small set of different processes, such as adding or multiplying single-digit numbers, writing down some numbers for later use (i.e., storing onto the tape in a Turing machine), and reading those numbers back in at the appropriate time (as you move to the next column of digits).

This kind of sequential, discrete, step-wise processing is entirely different from how our neurons work. Neurons also break down a problem in to simpler components, but a critical difference is that they all work together in *parallel* instead of the fundamentally sequential, *serial* processing required for a universal computer. The major advantage of serial processing is that it is much more flexible – any arbitrary collection of operations can be sequenced one after the other over time, but the same is *not* true for parallel computation. Some

operations are mutually incompatible with each other, or depend one on the other, and simply cannot be performed simultaneously in parallel. Indeed, one of the great challenges of modern computer science is trying to come up with even moderately usable parallel computing frameworks, and it is very clear that the universal flexibility of traditional serial computation does not extend into the parallel realm: parallel computation must generally be setup on a case-by-case basis. For example, in the case of multi-digit multiplication, you have to do the tens-place part of the problem first, before you know how much to carry over to the higher digits, etc – you can't just do everything all in one step.

More generally, parallel systems are really good at doing the same kind of thing over and over again really fast (e.g., detecting patterns via networks if interacting neurons in our brain), but they are not so good at doing random, arbitrary, *different* things, which is precisely where serial computation excels. However, serial computation is inherently much slower (one step at a time). These fundamental tradeoffs between parallel and serial computation mean that a system that can do both will be able to achieve the best of both worlds – that is the magic recipe that the human brain has achieved. Our brains are parallel at the level of individual neurons and networks of neurons, but at the larger *systems* level of the brain, we can achieve a form of flexible, serial computation.

Before turning to the biology of the brain systems supporting this latter form of computation, we can see strong evidence for the presence of these two different forms of computation at the psychological level. For example, we would predict that you need to use your "mental CPU"-like capacity whenever you take on a novel task. For example, when you first learned to drive a car, you relied on a sequential, deliberate process that consumed all of your attention – at each point in time, you had to keep reminding yourself of what you were supposed to be doing. However, with sufficient practice over time, these slow, effortful processes gradually become **automated**, and you may now find yourself driving down the freeway with very little awareness of any of the underlying steps you're effortlessly performing. This difference between the initial effortful **controlled processing** and the subsequent **automatic processing** was captured in a highly influential pair of papers by *Walter Schneider* and *Richard Shiffrin* (the same one who published the famous paper on the modal model of memory from the previous chapter) (Schneider and Shiffrin 1977; Shiffrin and Schneider 1977).

A widely-studied example of this difference between controlled vs. automatic processing is shown in Figure 7-3 – the *Stroop* task (Stroop 1935; MacLeod 1991). The participant is instructed to either read the word or name the color of the ink the word is written in. Because word reading is so overpracticed, it is an automatic process for most adults (to the point that you often can't stop yourself from re-reading the annoying text on your cereal box every morning). Therefore, when confronted with the diabolical *Conflict* condition in the Stroop task, where a color word (e.g., "Green") is written in a different ink color (e.g., red), it takes extra cognitive control to prevent yourself from just blurting out the word ("Green"), when the task is to name the color (for which you have much less practice). This shows up as a significant delay (and overt errors) in this

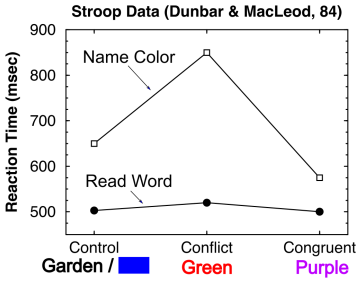


Figure 43: Fig 7-3: The Stroop task (Stroop, 1935; MacLeod, 1991) demonstrates the difference between controlled and automatic processing, in the context of reading words in different ink colors. If you try to name the ink color of the word "Green" when it is written in red ink (the *Conflict* condition), the automatic process of reading dominates over the relatively rare process of naming ink color, and you have to deploy controlled processing, which takes extra time as shown in the plotted data. Reading words (bottom line) isn't affected much by the ink color – the well-trained brain networks supporting this process proceed in parallel without any supervision required. Interestingly, even when the response is identical in the *Congruent* case, the task of color naming is still slower than word reading, reflecting the extra control being exerted. The *Control* condition involves either non-color-word reading, or pure color naming.

condition. Interestingly, even when the ink color and the word are *Congruent*, there is still a delay associated with naming the color – there is extra control being exerted to support this relatively unfamiliar color naming process.

If you run the Stroop task on kids, they don't show the same effects – reading has yet to become automatic in them. Furthermore, some Stroop researchers spent enough time color naming so that *it* become more automatic than reading, and they showed a kind of reverse-Stroop effect! Thus, this process of *automatization* is dependent on learning and practice – over time, our brains naturally turn deliberate, sequential, controlled processing into more parallel, automatic fast processing. This is the same phenomenon that occurs for driving and so many other tasks that you once found difficult and mentally all-consuming. Automatization is analogous to the *chunking* process discussed in the memory chapter – we have a very limited active memory capacity, but once we learn new concepts, we can greatly expand our capacity by using this limited capacity on chunks of information that used to be separate. Automatization is the process of forming *procedural* chunks – combining sequential steps into faster parallel processing that becomes relatively independent of our mental CPU – we no longer need to exert detailed conscious effort keeping the process moving along.

What it takes to be a Computer

The human brain is likely unique in having the ability to function like a Turing machine – other animals have plenty of automatic parallel processing skills, but they just don't seem to be capable of solving novel, complex tasks by performing a sequence of mental processing steps. The reason we can function like a computer is that we have some special capacities lacking in other types of brains, supplying the key ingredients of a Turing machine:

- *Program:* we use our *natural language* (e.g., English) as a kind of programming language. There is abundant evidence that we routinely use verbal self-instruction to remind ourselves of what we're supposed to do next in a complex, novel task (Miyake et al. 2004). We literally talk ourselves through the problem, and this capacity for stringing together different such verbal programs is an essential element of flexible, universal computation. It is unclear how far we might be able to get at flexible controlled processing without language, but likely not very far.
- *Active Memory* (registers, cache memory): special properties of our *frontal cortex* and *basal ganglia* give us the ability to maintain a small amount of information in active, **working memory**, as mentioned in the previous chapter. This is what you use when solving a mental arithmetic problem, by constantly juggling the digits around in your working memory. Working memory replaces the piece of paper you would otherwise use in keeping track of all the *partial products* and *control state* needed to keep progressing through a complex problem. It is also essential for maintaining the program itself, and in this way it much resembles the function of RAM in a computer, which maintains both the program and the *stack* and *heap* forms of active memory needed to carry out the program.
- *Controlled Memory Storage and Retrieval*: we also have the ability to take control over our hippopcampal episodic memory system, to deliberately encode and retrieve task-relevant information as needed, playing the role of the memory tape system in the Turing machine, and a hard drive in a modern computer.

Consistent with the central role of the frontal cortex in making our mental computer work, this brain area (particularly the part of it in front of the primary motor area – the **prefrontal cortex**) is differentially expanded in humans relative to other primates and mammals more generally (Semendeferi et al. 2002). Thus, a simple story is that our unique mental-computer skills are due to this expanded brain area, but this cannot be entirely correct, because prefrontal cortex is also similarly expanded in other great apes. Despite its expanded size, the capacity of our prefrontal working memory system (around 4 chunks) is dramatically smaller than that of even the most primitive digital computer. Thus, we are left with this rather startling conclusion: our super huge brains packed with neurons are often no match for a simple serial computational device composed of just a few basic parts – while our brains are impressive in many ways, they pale in comparison to a dime-store calculator for doing basic arithmetic!

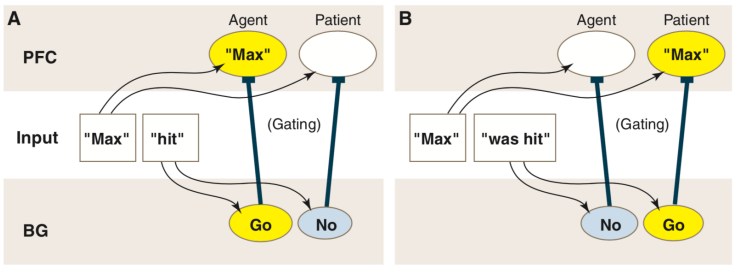


Figure 44: Fig 7-4: How the basal ganglia (BG) can "gate" information into different parts of prefrontal cortex (PFC) to achieve a flexible kind of variable binding like that needed for digital computers. In this case, the BG can flexibly control whether a given word / concept ("Max") is encoded as the agent or patient in a given sentence / scenario, based on other available cues or context (from O'Reilly, 2006).

The **basal ganglia** also play a critical role in making our mental CPU function, by orchestrating the serial, sequential steps of cognitive operations that we take in solving a problem – it turns the parallel brain into a serial system. As noted in the neuroscience chapter, the basal ganglia are critical for making a Go / NoGo decision about what to do next, and this decision-making bottleneck forces us to take discrete, sequential cogntive steps. Interestingly, the early cognitive models of human reasoning and problem solving incorporated something called a **production system** (Newell and Simon 1972), which plays the same role as the basal ganglia in the brain (Stocco, Lebiere, and Anderson 2010; Jilk et al. 2008). Figure 7-4 also shows that the basal ganglia can enable a form of flexible *variable binding* (R. O'Reilly 2006), which is another important property of computer systems that is otherwise hard for neurons to achieve. There are still important mysteries about how exactly the prefrontal cortex and basal ganglia learn to become a Turing-machine like system, but we do understand many of the basic principles at work already.

In summary, although we are unique in having *some* ability to perform flexible, serial, controlled processing to solve novel tasks, our brain is still running in automatic, parallel processing mode under the hood, and that greatly limits our computer-like abilities. What we lack in serial computing abilities, we try to make up for with all the amazing mental skills that we have automatized through learning and practice. This is consistent with the idea that motivation, which drives this learning, plays such an important role in human intelligence. And it also makes sense of the many ways in which our cognition differs from that of an optimal, fully rational computer system, as we discuss in a later section. Before turning to some of those issues, we first consider an alternative possibility for at least some individual differences in intelligence.

Individual Differences in Prefrontal Cortex / Basal Ganglia?

The critical role for working memory, cognitive control, and the prefrontal cortex / basal ganglia system in supporting our flexible computer-like cognitive abilities does introduce another possible explanation for individual differences in intelligence, however. It *could* be the case that different people somehow have different capacities / speeds / functionality in these particular brain systems, and that is what explains overall differences in intelligence. Furthermore, *if* this were the case, then because this flexible controlled-processing system is used as the first step in learning new skills and cognitive abilities (especially in math and other school-based learning), there could be a kind of snowballing effect where small initial differences in these brain areas could multiply over time, leading to larger overall differences in measured IQ scores. This kind of scenario is most similar to the "traditional" notions of intelligence as a fixed thing that you either have or don't have, and obviously fits well with intuitions based on the computer metaphor of the mind.

But how well does it fit with the available data? First of all, it is well-established that major disruption to the prefrontal cortex results in impairments to controlled processing, for example on the classic Stroop task (J. D. Cohen and Servan-Schreiber 1992; Stuss et al. 2001), and on many aspects of social, moral and other forms of reasoning (Eslinger, Flaherty-Craig, and Benton 2004). However, the latter paper, and various other related findings, have shown that measured IQ scores can be relatively intact even with significant early frontal brain damage, suggesting that the relationship may be somewhat more complicated.

The most relevant question, however, is whether *normal* variation in prefrontal cortex / basal ganglia function accounts for much of the measured individual differences in intelligence? One early attempt to answer this question relied on establishing correlations between measures of working memory and intelligence, and came up with a strong positive correlation on the order of .6 to .8 (Engle 2002). However, subsequent work largely undermined that conclusion, instead suggesting that there is a separate factor for general fluid intelligence, independent of working memory capacity (Engle 2018).

Another angle on this question found that much of the measured differences in working memory capacity were actually due to motivational factors in the first place (Adam and Vogel 2016). Specifically, participants who scored lower on their working memory scale did so because they had a higher probability of "lapsing" – just failing to engage in the task on a given trial. However, when they did engage, their measured working memory capacity was essentially the same as those who had a high working memory capacity score (due to a high overall level of task engagement). Thus, consistent with the overarching importance of motivation, it may be the critical "third variable" that drives the relationship between measured working memory and intelligence scores.

A recent attempt to more directly find the neural correlates of individual differences in intelligence found that motivational and emotional areas of the

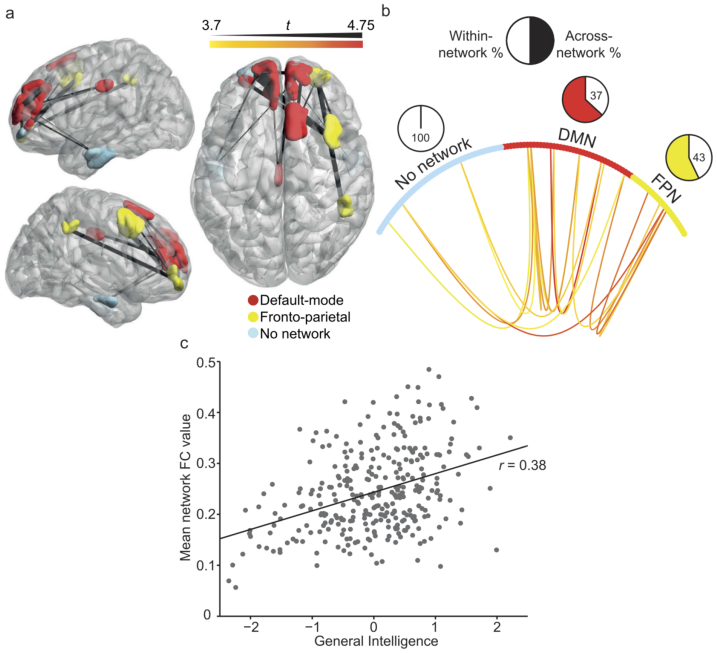


Figure 45: Fig 7-5: Correlation between brain activity & functional connectivity, and measured general intelligence (Hearne et al, 2016). The "default-mode network" areas in red are primary emotional / motivational areas in frontal cortex (see Figure 5-14) that are active when people are left to their own thoughts in the scanner. The fronto-parietal network in yellow are areas associated with cognitive control and working memory. Interestingly, the emotional / motivational areas make a major contribution to the overall correlation, both in terms of within-network interconnectivity and cross-network connections to the control areas.

prefrontal cortex are among the most strongly correlated with measured general intelligence (Hearne, Mattingley, and Cocchi 2016) (Figure 7-5). Thus, overall, the same answer keeps coming back up across all of these different studies: individual differences in intelligence seem to be more strongly driven by motivational factors than by the raw capacity or other properties of the prefrontal cortex / basal ganglia system. We can make sense of this result by considering that the relatively measly capacity of prefrontal cortex (only around 4 items can be maintained at a time) seems completely unrelated to the massive numbers of neurons in this brain area (which has maybe 10 billion neurons overall). Thus, the overall functional properties of this system are unlikely to be due to normal variation in the numbers of neurons or other basic biological properties of these areas. Instead, they are much more likely to be due to the degree of learning and experience that has shaped these networks to perform like a serial computer.

Strengths, Weaknesses, and Biases of our Neural Computer

The picture of human intelligence that has emerged from the above considerations has important implications for understanding how we think and reason in real-world situations, with life-and-death level consequences. To summarize: **Our brains operate best on familiar, concrete problems via prior learning and automatization, but we do have a limited ability to tackle novel problems through slow, serial, controlled processing. Also we are generally lazy and motivational factors are paramount in everything we do and learn.** This particular combination of strengths and limitations results in a systematic set of **cognitive biases** and other cognitive properties, many of which were first characterized by the nobel-prize winning psychologist *Daniel Kahneman* and his late longtime collaborator *Amos Tversky*.

These biases are described in terms of the reliance on **heuristics**, which are short-cut, "rule of thumb" kinds of solutions to problems, in contrast to the typically more labor-intensive (and often intractable) *optimal*, *rational* solutions. For example, the **representativeness heuristic** characterizes our reliance on overall similarity judgments of a given situation to a stereotype or prototypical situation that we have previously learned about. In other words, instead of going through all the mental effort of trying to truly understand a novel problem or situation, we just lazily rely on our previously-learned parallel neural pathways that *compress* something down into a seemingly well-understood high-level summary. Indeed, the representativeness heuristic is really just another name for our principle of compression.

The classic demonstration of this heuristic involves the fictional *Linda*, who is described as follows:

> Linda is 31 years old, single, outspoken, and very bright. She majored in philosophy. As a student, she was deeply concerned with issues of discrimination and social justice, and also participated in anti-nuclear demonstrations.

You are then asked which of the following alternatives is more probable:

A. Linda is a bank teller.

B. Linda is a bank teller and is active in the feminist movement.

What do you think? Most people answer B, because it is more similar to the stereotype activated by the description of Linda. Again, we lazily rely on our basic neural processing mechanisms whenever possible, and these naturally produce a stronger match for the second characterization of Linda. But, from a rational, mathematical perspective, it is *impossible* for B to be more probable. That word *and* is mathematically equivalent to multiplying probabilities, and the probability of anyone being active in the feminist movement is, objectively, less than 1. Thus, the joint probability of being a bank teller *and* a feminist must necessarily be less than just being a bank teller!

The **availability heuristic** is similar to the representativeness heuristic, in also relying on an overall sense of familiarity, instead of doing the hard cold math. In this case, the familiarity comes from the emotional (motivational) salience of different events, instead of our compiled stereotypes (though both are very similar in relying on learned synaptic pathways – it is not clear that these are really distinguishable at a neural level). To demonstrate this heuristic, please estimate which is more likely:

A. Dying in an airplane crash

B. Dying in from nephritis, nephrotic syndrome, and nephrosis

Many people avoid flying in airplanes due to fear of crashes, which inevitably get extensive media coverage, and are highly salient and gruesome. It is not clear if there has ever been any major media coverage of the latter cause of death, or if you even know what it is. Yet it is the 9th highest cause of death, at over 50 thousand in the USA in 2016, according to the CDC: `https://www.cdc.gov/nchs/fastats/leading-causes-of-death.htm`. In contrast, there were only *59* total deaths *worldwide* due to airplane crashes in 2017!

The real-world life-and-death implications of our reliance on the availability heuristic were starkly illustrated in an analysis by *Gerd Gigerenzer* of people's avoidance of flying after the September 11, 2001 terrorist attacks in the US (Gigerenzer 2006). He found that there were roughly 1,500 additional deaths caused by people choosing to drive instead of fly – driving is much, much riskier than flying, and yet because it is so familiar and commonplace, people vastly underestimate the relative risks.

Gigerenzer and colleagues have many other studies showing our general incompetance in dealing with statistical information, again with real-world implications. For example, even highly-trained doctors make significant mistakes interpreting the statistical results of medical studies on health risks and probabilities. One of the most important errors we make is known as **base rate neglect**. For example, if you hear that drinking alcohol increases your risk of death by 12%, that sounds like a big effect. But people routinely neglect to take into account that the base rate against which

this percentage is measured is extremely low. In practice, the difference amounts to just *4* additional deaths per 100,000 people per year: NYT article: `https://www.nytimes.com/2018/08/28/upshot/alcohol-health-risks-study-worry.html`. This is really an instance of the *contrast* effect – we focus on differences but not on the raw values.

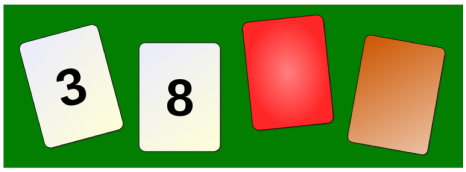


Figure 46: Fig 7-6: Wason card selection task. Your job is to verify whether the following rule holds for these cards: If there is an even number on one side, then the other side is red. Which cards would you turn over to test whether this rule holds?

Another important demonstration of our strong preference for reasoning about familiar, concrete situations comes from the **Wason card selection task** (Wason 1968) (Figure 7-6). Here, your job is to verify the application of a particular rule about what is on different sides of the cards. Look at the figure and formulate your answer. Now, consider this alternative problem:

> You are a bartender. If you are going to serve alcohol to someone, they need to be over 21. Who do you card?

Critically, you card someone who is *under* 21. Did you likewise decide to flip the *brown* card in the Wason card selection task? Probably not. But think about it: if a card has brown on one side, and it has an even number on the other side, then the rule is incorrect. The brown card is the "cheater". You probably said that the red card should be flipped. But the rule says nothing about the other side of a red card – it could be either even or odd. Likewise, people over 21 can either drink or not drink – it doesn't matter. You don't card people who are obviously over 21.

This task again demonstrates that our ability to perform abstract logical reasoning is extremely limited, and we do much better with familiar, concrete cases like the bartender example. By contrast, a standard digital computer programmed with the basic rules of logic can easily get the abstract form of this problem correct, and in fact would struggle understanding a question like "who do you card?" without much more explict specification of what that means. Thus, people are really good at "common sense" reasoning, and pretty bad at abstract reasoning, and computers are generally the opposite.

Task Transfer and Education

Another very important consequence of the concrete nature of our brains is that things we learn in one context often do not **transfer** very well to another

context. For example, classic studies of problem solving tasks such as the *Tower of Hanoi* have found that, having figured out a solution to that problem, people are generally not very good at applying that very same solution to another version of the exact same problem, portrayed in a different "skin" (i.e., differing in only superficial, task-irrelevant factors) (Kotovsky, Hayes, and Simon 1985). Although transfer can occur, and there are reliable ways to make it more likely to occur (J. R. Anderson, Reder, and Simon 1996), it is clearly not the natural state of the system. From a neural perspective, the brain learns everything in terms of the specific patterns of neural activity present during the learning episode, so the only way to get knowledge to transfer to a novel situation is to ensure that these patterns of neural activity are shared across situations.

These and other related findings have led to the development of the **situated learning** theory (Lave and Wenger 1991; Greeno, Moore, and Smith 1993), which emphasizes the concrete, specific nature of human learning. Although there have been some debates about the generality of this framework (J. R. Anderson, Reder, and Simon 1996) – people *do* have some ability to engage in abstract reasoning and *some* knowledge does transfer – the basic principles remain solid and are increasingly incorporated into educational strategies.

A recent controversy about transfer of learning has arisen in the context of the increasingly popular brain training programs, such as that offered by `lumosity.com`. These programs are based on the idea that you can train up your brain like a muscle, and increase your overall intellegence as a result. However, consistent with the overall difficulty in transfering learning, (Simons et al. 2016) review a number of studies showing that there is very little transfer of this brain training beyond the specific tasks that you practice. So, you can get better at the specific arcane puzzles that you practice, but, unfortunately, it doesn't really transfer to make you generally more intelligent. And lumosity was successfully sued for making false, misleading claims to the contrary – that's psychology in the real world!

Programs in the Mind: Problem Solving and Reasoning

Despite all the denigration of our capacity for abstract reasoning, and the relatively modest power of the neural CPU in our brains, we nevertheless can do some impressive feats of cognition. At least, some of us can! After all, Turing was able to prove the universal nature of his computational machine, using his own relatively modest neural CPU (he didn't have a real computer to work with yet). In this section, we review some of the kinds of things we can do with our neural CPU's, and how these have been studied in psychology.

As noted earlier, some of the pioneering early work using the computer as a model for the human mind was conducted by *Allan Newell* and *Herbert Simon* at Carnegie Mellon University (CMU) in Pittsburgh, PA (Newell and Simon 1972). These scientists and their colleagues focused on how people solve various kinds of challenging puzzles and games, including the *Tower of Hanoi* (Figure 7-7), and chess. When you first start doing any kind of puzzle, the initial strategy is

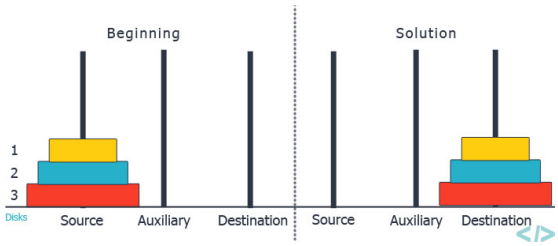


Figure 47: Fig 7-7: Tower of Hanoi task: Your job is to move all the discs from the source to the destination, one at a time, without ever putting a larger disc on top of a smaller one. Go for it!

often a semi-random **trial-and-error** process of trying out different actions, and seeing how it goes. Another name for this is **hill-climbing** or **gradient ascent**, where you measure success in terms of the visual similarity of the current state to the target state. In the case of Tower of Hanoi, this results in trying to put discs on the destination peg as early as possible, which turns out to not be such a great idea.

Interestingly, good puzzles often have this characteristic of requiring you to move further *away* from something that looks like the final solution, in order to solve the problem. In other words, they specifically thwart the natural hill-climbing solution. This ability to move away from the goal has been studied in babies and animals using a *transparent barrier detour* task, where a desired object (food or toy) is hidden behind a transparent barrier – the direct reach solution must be rejected in favor of the indirect reach-around behind the barrier (Diamond 1990; J D Wallis et al. 2001). These studies show that the prefrontal cortex is important for this ability to overcome the direct approach in favor of the indirect reach, consistent with the idea that these kinds of challenging puzzles more generally tap our higher-level flexible cognitive processing.

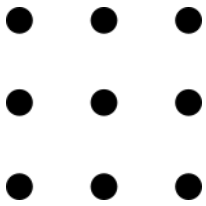


Figure 48: Fig 7-8: Classic "thinking outside the box" insight problem. Connect all of the dots using four straight lines, without lifting up your pen.

This same requirement for overcoming the initial "obvious" solutions to a problem is featured in **insight problems**, where the surface features of the problem strongly suggest one type of solution, but another, different kind of solution is actually required (Figure 7-8). The relative difficulty that we have breaking out of a given **mental set**, known as **functional fixedness**, is nicely

illustrated by these problems. But really, it is just a more advanced version of the same problem in all of these puzzles: the first ideas don't work, and you need to hunt around for an entirely different approach. This has entered the popular lexicon in terms of "thinking outside the box", likely in reference to the puzzle in Figure 7-8.

Games such as chess provide the opportunity to explore more advanced strategic thinking. Early AI approaches to solving these games involved the "obvious" strategy of figuring out what your opponent might do if you make a given move, and so on, to pick the move with the best potential future outcome. This is known as a *look-ahead* search algorithm, and is essentially what the Deep Blue chess playing computer employed when it beat the famous chess champion, Gary Kasparov. The main challenge with such an algorithm is the *combinatorial explosion* problem associated with the very large *state space* of a complex game like chess – there are so many different possible future move strategies, that searching all of them quickly becomes computationally prohibitive. The Deep Blue computer basically used massive numbers of computer chips to search this space in parallel, executing a pure "brute force" solution.

The limited capacity of our neural CPU prevents us from using such a brute force strategy. Instead, research from Simon and colleagues at CMU showed that people use a much more perceptually-based strategy (Gobet and Simon 1996). Specifically, chess experts leverage their massively-parallel neural networks to recognize good and bad board positions, based on extensive learning experience with the game. The recent advances by the Google DeepMind team on AI systems that play the ancient game of *Go* essentially combine this perceptual expertise strategy with the brute-force look-ahead strategy, to create a human-crushing Go master machine (Silver et al. 2017). This system provides a good demonstration of the power of the combination of parallel neural-like computation, with more flexible CPU-like processing. However, unlike people, the DeepMind model did not learn everything from the ground up – it was programmed and otherwise designed strictly for playing Go, with the rules of the game coded directly into the look-ahead part of the system. Nevertheless, it did discover novel strategies by playing millions of games with itself.

So do we think that the DeepMind system exhibited human-like insight and creativity? Or is it really just another example of brute-force searching of the state space of the game, which happened to turn up novel strategies? And how do we really differentiate this from what people do when they come up with novel insights to challenging problems? One critical difference is that people have an awareness of the strategies they are trying, and can explicitly formulate particular aspects of the overall puzzle, to guide their insight process along. Thus, we are using something more like a **means-ends** analysis of the problem – reasoning backward from target solutions (the "ends") to the means that should lead to those ends. Also, we can deploy these creative reasoning skill across many different domains. We'll see how long it takes for machines to match or exceed our abilities in these respects!

Measuring Intelligence and its Implications

Now that we have some understanding of the nature and scope of human intelligence, we turn to the controversial history of IQ testing. This history is controversial because the central question of the genetic versus environmental basis for intelligence has been at the center of these tests from their beginning, and because of the major question of the potentially biased nature of these tests. Briefly, one of the earliest tests was developed by *Alfred Binet* in France in the early 1900's, and translated and further developed by *Lewis Terman* at Stanford University, resulting in the **Stanford-Binet** IQ test. This test was developed based directly on academic skills, and includes factors such as knowledge, quantitative reasoning, visual-spatial processing, working memory, and fluid intelligence. The test was standardized for children at different grades, and was initially focused on identifying children who were outside of the normal range for each grade.

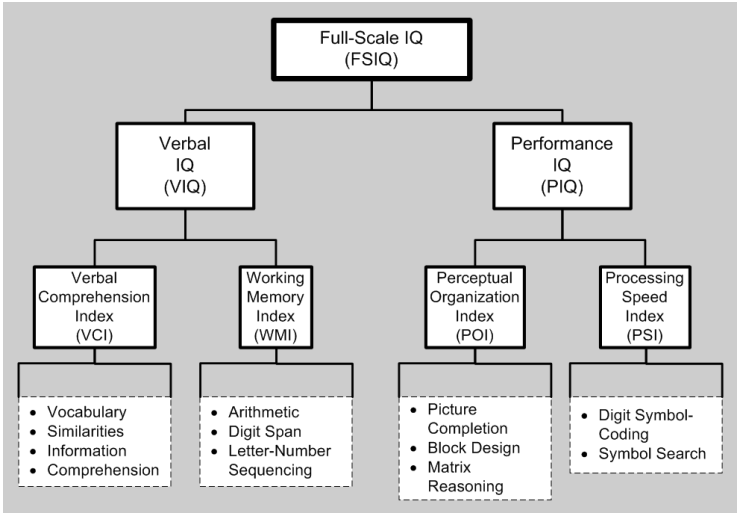


Figure 49: Fig 7-9: Sub-scales of the WAIS intelligence scale, divided into verbal vs. non-verbal ("performance") components.

The later **Wechsler Adult Intelligence Scale (WAIS)**, first developed by *David Wechsler* in the 1930's, is still widely used to this day, and is by far the most popular IQ test. It attempted to improve over the Stanford-Binet by separately testing verbal and non-verbal ("performance") intelligence (Figure 7-9), and relying less on raw speed as a factor (though it is retained as a specific sub-factor). The WAIS is similar to the Stanford-Binet test in using more of a "scattershot" approach to measuring intelligence – many different specific tests are combined together to paint an overall picture. And, consistent with the essential role for working memory in supporting flexible CPU-like processing, working memory constitutes one of four major subscales in the WAIS.

An important feature of the WAIS measure is that it has been **standardized** so that a score of 100 is *defined* as having a precisely average level of intelligence, and intervals of 15 points represent a standard deviation according to the normal (gaussian) distribution. Thus, someone with an IQ of 115 is one standard deviation above the mean, and is thus measured as being "smarter" than 84.13% of the population, and someone with an IQ of 130 is "smarter" than 97.72% of the population. This standardization applies to "adults" aged 16 or older, and there is a different scale for children.

One of the major sources of controversy about IQ tests such as the WAIS is the extent to which it might be systematically biased against different populations, such as women and minorities. This is a difficult question to answer, because there is no other accepted standard to compare against. In this case, we can at least address three important features of the test: its **reliability**, **construct validity**, and **predictive validity**. The reliability is assessed by repeatedly testing the same people, ideally with different versions of the same overall test, and measuring the consistency of their scores across tests. Construct validity is much more subjective and difficult to assess – basically it is a judgement call on the part of scientists that the individual test components are actually measuring something useful about the construct of intelligence. In any case, it is somewhat circular: the test measures those specific abilities that it actually tests, and your score reflects your ability to perform those specific tasks. No specific test can ever measure anything other than what it specifically tests.

Thus, the most important property of the test is its *predictive validity*: how well does it predict *other* things about people? This is the most relevant and directly measurable property of an IQ test – if it is really measuring intelligence, then it should be able to predict how well people do at things that we generally agree require intelligence. Here, the evidence is a bit mixed. For example, your IQ test score predicts subsequent school performance (i.e., grades) with a correlation factor of about .5 (Neisser et al. 1996). This is actually not a very strong correlation – even though it sounds like a "half strong" correlation, you have to square this number to determine how much total variance in grade outcomes are accounted for by IQ – so only 25% of the total variance. This means 75% of the variance is *not* attributable to IQ – by far the majority of it. Nevertheless, it is probably the best *single* predictor of academic performance. Interestingly, in discussing the strength of this factor, (Neisser et al. 1996) point largely to motivational factors as likely additional determiners of academic performance.

The amount of variance predicted by IQ scores on factors such as job performance, income, socio-economic status (SES), health, and social status are all lower than for school performance, and it is very difficult to accurately factor out the contributions of parental factors (e.g., parent SES), which affect both IQ and most of these other factors (this point also applies to the educational predictiveness).

Importantly, the predictive validity of IQ scores does *not* differ across different groups, such as women or minorities (Neisser et al. 1996). This

provides one way of assessing whether IQ tests are inherently biased, and the results suggest that they aren't biased at this level. However, just looking at the test from a construct validity perspective, it definitely does test lots of knowledge that is typically learned in school, and must be expressed verbally. Thus, an individual's prior schooling environment is undoubtedly going to have a significant impact on their IQ score. Furthermore, as noted above, there are significant correlations between parental SES and IQ, at about .33 (Neisser et al. 1996). Thus, there is no doubt that your raw IQ score reflects a strong contribution from various environmental factors that likely systematically vary among different groups. Interestingly, the predictive validity factor is not directly affected by such overall mean differences – it is only affected by the *variance* in scores across individuals.

In summary, IQ is both a biased and fair test of intelligence! It is biased at the mean level, and by the very fact that *any* test of complex task performance is likely to be affected by relevant SES-level factors. But at the level of predictive validity, it is fair in that there aren't significant differences across groups. We'll consider the data on the genetics and heritability of intelligence in the chapter on genetics and development.

Multiple intelligences

One of the major debates about IQ measures is the extent to which there are really distinct aspects of intelligence, or rather a single underlying **general intelligence factor**, typically denoted by the letter *g*. *Charles Spearman* was the main original advocate of the importance of this *g* factor – he noted that performance across the different subscales of IQ tests is positively correlated, and attributed this common factor (*g*) as being the "true" underlying meaning of intelligence. Arguing the opposite case was *L. L. Thurstone*, who advocated a multi-factorial model of intelligence. Interestingly, both were looking at the same data, and just interpreting them differently – there are definitely multiple separable intelligence factors in addition to the common variance across all of them associated with *g*.

A distinction that is less subject to these whims of interpretation is the difference between **crystallized** and **fluid** intelligence, as proposed originally by *Raymond Cattell*. Crystallized intelligence refers to the accumulated knowledge and skills learned over experience, whereas fluid intelligence refers to the ability to actively juggle information in your mind (e.g., in the service of mental arithmetic). Thus, fluid intelligence corresponds to the contributions of the prefrontal cortex and basal ganglia, supporting working memory and the ability to shuffle information rapidly around within your mental workspace. By contrast, crystallized intelligence refers to the accumulated synaptic changes from learning. These two forms of intelligence are often considered from an aging perspective, where crystallized intelligence generally increases over the lifespan (i.e., wisdom accumulates over time), while fluid intelligence unfortunately declines. This decline is thought to mirror the declines in dopamine levels with aging, which

may suggest that there may be an important motivational component to the story as well. Could a strongly-motivated senior still muster the computational horsepower of a young adult?

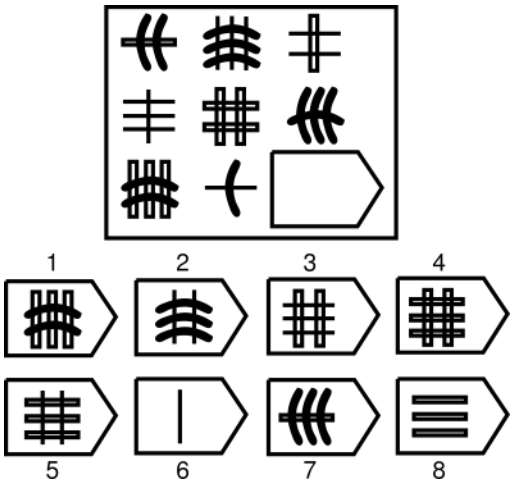


Figure 50: Fig 7-10: Raven's progressive matricies task, which resembles the matrix reasoning component of the WAIS, and is often used as the single best measure of general fluid intelligence. Your task is to figure out which of the options at the bottom best completes the pattern shown in the *matrix* at the top – row-wise, column-wise, and the diagonal progressions of the pattern are all relevant! This need to juggle multiple different sequences makes this a particularly challenging task.

One of the best single measures of general fluid intelligence is the Raven's progressive matricies task, shown in Figure 7-10 (J. C. Raven, Court, and Raven 1977; Conway et al. 2002). This task is also similar to the matrix tasks used in the WISC, to assess non-verbal intelligence. The non-verbal nature of the task likely helps to keep it from being "contaminated" by crystallized knowledge factors, and thus contributes to its ability to more directly measure the construct of fluid intelligence. It clearly requires considerable juggling of information in working memory, to figure out which of the possible patterns best completes the missing cell in the matrix. Furthermore, this Raven's task and the fluid intelligence construct have been closely associated with working memory function (Conway et al. 2002).

As discussed earlier, many people interpret this notion of fluid intelligence and working memory function in computer-like terms – some people just have better "RAM" than other people. However, as we concluded earlier, the available evidence more strongly implicates motivational factors as paramount. The revised interpretation then becomes: general fluid intelligence measures reflect the extent to which an individual is willing to exert significant cognitive effort on arbitrary lab tasks. People who are more willing to do this score higher on these

measures, and, perhaps not coincidently, also tend to do a bit better in school overall – school similarly requires you to allocate cognitive effort on things that you might not otherwise want to spend time and effort on.

Control

This brings us back to the recurring theme of this chapter: the importance of motivational factors in understanding intelligence, and the functions of the prefrontal cortex and basal ganglia more generally. Ultimately, it comes down to a question of *control*. These brain structures are critical for cognitive control, but also for overall control of your entire self, at all the relevant levels. The ventral and medial areas of prefrontal cortex receive extensive inputs from brainstem emotional and body-state areas, and integrate this "hot" state information with "cold" planning and sensory-motor control strategies, all in the service of keeping *you* in control of *you*.

Ultimately, control is about achieving your own goals, satisfying your own needs, etc. Thus, when scientists bring you into the lab, you really are a *participant* in the experiment, and the degree to which different people actually participate varies according to their motivational goals. If someone is really interested in what is going on inside their brain, and motivated to demonstrate their own sense of mental superiority, they may devote considerable effort to a given lab task. But others may have "more important" things on their minds, and spend less overall effort.

Neuroimaging studies have provided a vivid window onto this battle between your own internal desires and goals, and those of the psychology experimenter. In any given experiment, your brain can be seen switching between a "task engaged" mode and the "default mode" (Fox et al. 2005). The default mode (see Figure 7-5) involves all the motivational / emotional areas of the frontal cortex, interconnected with the hippocampus and other brain areas that are relevant for thinking over important recent events and planning upcoming activities. Basically, the stuff you think about when left to your own devices. By contrast, the task-engaged brain areas involve the working-memory and problem-solving, cognitive control areas that are needed to keep you focused and solving complicated tasks.

As we saw in the working memory capacity results discussed earlier (Adam and Vogel 2016), the major determiner of measured overall capacity is how often people were willing to actually engage in the working memory task. Presumably, the rest of the time they were thinking about more self-relevant things, and exerting their own personal control over what they allocate their attention to. Likewise, you may have drifted off into your own mindwandering space while nominally reading this chapter – we can throw the words at you, but ultimately you are in control of whether you want to read them!

Summary of Key Terms

This is a checklist of key terms / concepts that you should know about from this chapter.

- Fixed vs. growth mindset and the malleability of intelligence
- Algorithm = program for solving a given problem (e.g., mental arithmetic)
- Automatic vs. controlled processing
- Stroop task: word reading is automatic compared to color naming that requires controlled processing
- Neural CPU:
  - Working memory supported by prefrontal cortex = active memory / RAM in a computer
  - Basal ganglia = production system = sequentializing cognitive steps
  - Natural language (e.g., English) = program
- Cognitive biases / heuristics: shortcuts that leverage strengths, avoid hard-to-compute optimal, rational solutions.
  - Representative heuristic: use similarity to prototypes / stereotypes instead of actual statistics. Compression!
  - Availability heuristic: use familiarity instead of actual statistics.
  - Base-rate neglect: focus on percents instead of overall probability. Contrast!
  - Better at concrete vs. abstract reasoning: Wason card selection task
  - Transfer of learning: not much = situated learning: learning is specific to situations
- Problem solving:
  - Strategies: trial-and-error, hill-climbing / gradient ascent, means-ends
  - Puzzles designed to frustrate obvious / hill-climbing solution
  - Insight problems: mental-set and functional fixedness
- Intelligence tests:
  - Stanford-Binet
  - Wechsler Adult Intelligence Scale (WAIS) – standardized scale (100 mean, 15 standard deviation)
  - reliability, construct validity, predictive validity – importance of predictive validity for
- Multiple intelligences
  - Generalized intelligence factor g
  - Crystallized vs. fluid intelligence
  - Raven's progressive matricies as non-verbal test of fluid intelligence
- Control:
  - Cognitive control in service of overall control

Chapter 8: Language

Chapter 9: Evolution, Genetics, and Development

- Miyake & Friedman and the genetic basis of IQ: gets stronger over time, just like a learning system.
- important considerations in interpreting genetic results..
- A-not-B, DCCS same as functional fixedness in adults – just the baby forms of them..

Acknowledgments

Thanks to the current beta-testers for reading!

Glossary

About the Authors

Randall C. O'Reilly is Professor of Psychology and Neuroscience at the University of Colorado Boulder.

References

Adam, Kirsten C. S., and Edward K. Vogel. 2016. "Reducing Failures of Working Memory with Performance Feedback." *Psychonomic Bulletin & Review* 23 (5): 1520–7. doi:10.3758/s13423-016-1019-4.

Alarcon, Juan M., Angel Barco, and Eric R. Kandel. 2006. "Capture of the Late Phase of Long-Term Potentiation Within and Across the Apical and Basilar Dendritic Compartments of Ca1 Pyramidal Neurons: Synaptic Tagging Is Compartment Restricted." *Journal of Neuroscience* 26 (1): 256–64. doi:10.1523/JNEUROSCI.3196-05.2006.

Anagnostaras, S. G., S. Maren, and M. S. Fanselow. 1999. "Temporally Graded Retrograde Amnesia of Contextual Fear After Hippocampal Damage in Rats: Within-Subjects Examination." *The Journal of Neuroscience : The Official Journal of the Society for Neuroscience* 19 (February): 1106.

Anderson, John R., Lynne M. Reder, and Herbert A. Simon. 1996. "Situated Learning and Education." *Educational Researcher* 25 (4): 5–11. doi:10.3102/0013189X025004005.

Aston-Jones, Gary, and Jonathan D. Cohen. 2005. "An Integrative Theory of Locus Coeruleus-Norepinephrine Function: Adaptive Gain and Optimal Performance." *Annual Review of Neuroscience* 28 (July): 403–50.

Atkinson, R. C., and R. M. Shiffrin. 1968. "Human Memory: A Proposed System and Its Control Processes." In *The Psychology of Learning and Motivation: Advances in Research and Theory*, edited by K. W. Spence, 89–195. New York: Academic Press.

Baddeley, A. D., and G. J. Hitch. 1974. "Working Memory." In *The Psychology of Learning and Motivation*, edited by G. Bower, VIII:47–89. New York: Academic Press.

Baddeley, A., S. Gathercole, and C. Papagno. 1998. "The Phonological Loop as a Language Learning Device." *Psychological Review* 105 (March): 158.

Balleine, B. W., and A. Dickinson. 1998. "Goal-Directed Instrumental Action: Contingency and Incentive Learning and Their Cortical Substrates." *Neuropharmacology* 37 (May): 407–19.

Bartlett, F. C. 1932. *Remembering: A Study in Experimental and Social Psychology.* Cambridge: Cambridge University Press.

Bjork, Robert A. 1994. "Memory and Metamemory Considerations in the Training of Human Beings." In *Metacognition: Knowing About Knowing*, 185–205. Cambridge, MA, US: The MIT Press.

Bliss, T. V., and T. Lomo. 1973. "Long-Lasting Potentiation of Synaptic Transmission in the Dentate Area of the Anaesthetized Rabbit Following Stimulation of the Perforant Path." *The Journal of Physiology* 232 (October): 331–56.

Buschman, Timothy J., Markus Siegel, Jefferson E. Roy, and Earl K. Miller. 2011. "Neural Substrates of Cognitive Capacity Limitations." *Proceedings of the*

National Academy of Sciences 108 (27): 11252–5.

Buzsáki, G. 1989. "Two-Stage Model of Memory Trace Formation: A Role for 'Noisy' Brain States." *Neuroscience* 31 (3): 551–70. doi:10.1016/0306-4522(89)90423-5.

Cameron, Judy, Katherine M. Banko, and W. David Pierce. 2001. "Pervasive Negative Effects of Rewards on Intrinsic Motivation: The Myth Continues." *The Behavior Analyst* 24 (1): 1–44. doi:10.1007/BF03392017.

Cardinal, Rudolf N., John A. Parkinson, Jeremy Hall, and Barry J. Everitt. 2002. "Emotion and Motivation: The Role of the Amygdala, Ventral Striatum, and Prefrontal Cortex." *Neuroscience and Biobehavioral Reviews* 26 (May): 321–52.

Carver, C S, and M F Scheier. 1990. "Origins and Functions of Positive and Negative Affect: A Control-Process View." *Psychological Review* 97 (December): 19–35.

Carver, C S, and T White. 1994. "Behavioral Inhibition, Behavioral Activation, and Affective Responses to Impending Reward and Punishment: The BIS/BAS Scales." *Journal of Personality and Social Psychology* 67 (December): 319–33.

Ceci, Stephen J., Mary Lyndia Crotteau Huffman, Elliott Smith, and Elizabeth F. Loftus. 1994. "Repeatedly Thinking About a Non-Event: Source Misattributions Among Preschoolers." *Consciousness and Cognition* 3 (3): 388–407. doi:10.1006/ccog.1994.1022.

Cohen, J. D., and D. Servan-Schreiber. 1992. "Context, Cortex, and Dopamine: A Connectionist Approach to Behavior and Biology in Schizophrenia." *Psychological Review* 99 (April): 45–77.

Conway, A. R. A., N. Cowan, M. F. Bunting, D. J. Therriault, and S. R. B. Minkoff. 2002. "A Latent Variable Analysis of Working Memory Capacity, Short Term Memory Capacity, Processing Speed, and General Fluid Intelligence." *Intelligence* 30 (January): 163–83.

Cowan, N. 2001. "The Magical Number 4 in Short-Term Memory: A Reconsideration of Mental Storage Capacity." *Behavioral and Brain Sciences* 24 (August): 87–185.

Craik, F. I. M., and R. S. Lockhart. 1972. "Levels of Processing: A Framework for Memory Research." *Journal of Verbal Learning and Verbal Behavior* 11 (January): 671–84.

Crick, F. 1989. "The Recent Excitement About Neural Networks." *Nature* 337 (February): 129–32.

Deci, E. L., R. Koestner, and R. M. Ryan. 2000. "A Meta-Analytic Review of Experiments Examining the Effects of Extrinsic Rewards on Intrinsic Motivation." *Psychological Bulletin* 125 (February): 627.

Deese, James. 1959. "On the Prediction of Occurrence of Particular Verbal Intrusions in Immediate Recall." *Journal of Experimental Psychology* 58 (1):

17–22. doi:10.1037/h0046671.

Diamond, A. 1990. "The Development and Neural Bases of Memory Functions as Indexed by the A-Not-B Task: Evidence for Dependence on Dorsolateral Prefrontal Cortex." In *The Development and Neural Bases of Higher Cognitive Functions*, edited by A. Diamond, 267–317. New York: New York Academy of Science Press.

Dweck, Carol S. 2008. *Mindset: The New Psychology of Success*. Ballantine Books.

Ebbinghaus (1885), Hermann. 2013. "Memory: A Contribution to Experimental Psychology." *Annals of Neurosciences* 20 (4): 155–56. doi:10.5214/ans.0972.7531.200408.

Ekman, P., and W. V. Friesen. 1976. "Measuring Facial Movement." *Environmental Psychology and Nonverbal Behavior* 1 (1): 56–75.

Engle, Randall W. 2002. "Working Memory Capacity as Executive Attention." *Current Directions in Psychological Science* 11 (January): 19–23.

———. 2018. "Working Memory and Executive Attention: A Revisit." *Perspectives on Psychological Science* 13 (2): 190–93. doi:10.1177/1745691617720478.

Ericcson, K A, W G Chase, and S Faloon. 1980. "Acquisition of a Memory Skill." *Science* 208 (June).

Ericsson, K. A., and A. C. Lehmann. 1996. "Expert and Exceptional Performance: Evidence of Maximal Adaptation to Task Constraints." *Annual Review of Psychology* 47 (1): 273–305. doi:10.1146/annurev.psych.47.1.273.

Eslinger, Paul J., Claire V. Flaherty-Craig, and Arthur L. Benton. 2004. "Developmental Outcomes After Early Prefrontal Cortex Damage." *Brain and Cognition*, Development of orbitofrontal function, 55 (1): 84–103. doi:10.1016/S0278-2626(03)00281-1.

Fox, Michael D., Abraham Z. Snyder, Justin L. Vincent, Maurizio Corbetta, David C. Van Essen, and Marcus E. Raichle. 2005. "The Human Brain Is Intrinsically Organized into Dynamic, Anticorrelated Functional Networks." *Proceedings of the National Academy of Sciences of the United States of America* 102 (27): 9673–8. doi:10.1073/pnas.0504136102.

Frank, M. J. 2005. "When and When Not to Use Your Subthalamic Nucleus: Lessons from a Computational Model of the Basal Ganglia." *Modelling Natural Action Selection: Proceedings of an International Workshop*, January, 53–60.

Frey, U., and R. G. M. Morris. 1998. "Weak Before Strong: Dissociating Synaptic Tagging and Plasticity-Factor Accounts of Late-LTP." *Neuropharmacology* 37 (May): 545–52.

Fuster, J. M., and G. E. Alexander. 1971. "Neuron Activity Related to Short-Term Memory." *Science* 173 (January): 652–54.

Gallese, Vittorio, Christian Keysers, and Giacomo Rizzolatti. 2004. "A Unifying View of the Basis of Social Cognition." *Trends in Cognitive Sciences* 8

(9): 396–403.

Gerfen, Charles R., and D. James Surmeier. 2011. "Modulation of Striatal Projection Systems by Dopamine." *Annual Review of Neuroscience* 34: 441–66.

Gershman, Samuel J., David M. Blei, and Yael Niv. 2010. "Context, Learning, and Extinction." *Psychological Review* 117 (1): 197–209.

Gigerenzer, Gerd. 2006. "Out of the Frying Pan into the Fire: Behavioral Reactions to Terrorist Attacks." *Risk Analysis* 26 (2): 347–51. doi:10.1111/j.1539-6924.2006.00753.x.

Gobet, Fernand, and Herbert A. Simon. 1996. "The Roles of Recognition Processes and Look-Ahead Search in Time-Constrained Expert Problems Solving: Evidence from Grand-Master-Level Chess." *Psychological Science* 7 (January): 52.

Godden, D. R., and A. D. Baddeley. 1975. "Context-Dependent Memory in Two Natural Environments: On Land and Under Water." *British Journal of Psychology* 66 (January): 325–31.

Goldman-Rakic,.P S. 1995. "Architecture of the Prefrontal Cortex and the Central Executive." *Annals of the New York Academy of Sciences* 769 (December): 71–84.

Gollwitzer, P. M. 1993. "Goal Achievement: The Role of Intentions." *European Review of Social Psychology* 4: 141–85.

Goodwin, Donald W., Barbara Powell, David Bremer, Haskel Hoine, and John Stern. 1969. "Alcohol and Recall: State-Dependent Effects in Man." *Science* 163 (3873): 1358–60. doi:10.1126/science.163.3873.1358.

Greeno, James G., Joyce L. Moore, and David R. Smith. 1993. "Transfer of Situated Learning." In *Transfer on Trial: Intelligence, Cognition, and Instruction.*, 99–167. Westport, CT, US: Ablex Publishing.

Hasselmo, Michael E., and Chantal E. Stern. 2006. "Mechanisms Underlying Working Memory for Novel Information." *Trends in Cognitive Sciences* 10 (November).

Hayne, Harlene. 2004. "Infant Memory Development: Implications for Childhood Amnesia." *Developmental Review*, The nature and consequences of very early memory development, 24 (1): 33–73. doi:10.1016/j.dr.2003.09.007.

Hazy, Thomas E., Michael J. Frank, and R. C. O'Reilly. 2010. "Neural Mechanisms of Acquired Phasic Dopamine Responses in Learning." *Neuroscience and Biobehavioral Reviews* 34 (5): 701–20.

Hearne, Luke J., Jason B. Mattingley, and Luca Cocchi. 2016. "Functional Brain Networks Related to Individual Differences in Human Intelligence at Rest." *Scientific Reports* 6 (August): 32328. doi:10.1038/srep32328.

Hebb, D. O. 1949. *The Organization of Behavior*. New York: Wiley.

Hodgkin, A. L., and A. F. Huxley. 1952. "A Quantitative Description of Membrane Current and Its Application to Conduction and Excitation in Nerve."

The Journal of Physiology 117 (4): 500–544. doi:10.1113/jphysiol.1952.sp004764.

Hopwood, Christopher J., Aidan G. C. Wright, Emily B. Ansell, and Aaron L. Pincus. 2013. "The Interpersonal Core of Personality Pathology." *Journal of Personality Disorders* 27 (3): 270–95. doi:10.1521/pedi.2013.27.3.270.

Howard, M. W., and M. J. Kahana. 1999. "Contextual Variability and Serial Position Effects in Free Recall." *Journal of Experimental Psychology. Learning, Memory, and Cognition* 25 (August): 923.

Hull, C. L. 1943. *Principles of Behavior*. Appleton.

Iacoboni, M., R. P. Woods, and G. Rizzolatti. 1999. "Cortical Mechanisms of Human Imitation." *Science* 286 (January): 2526.

Jacoby, L. L., J. P. Toth, and A. P. Yonelinas. 1993. "Separating Conscious and Unconscious Influences of Memory: Measuring Recollection." *Journal of Experimental Psychology: General* 122 (2): 139–54.

Jilk, David, Christian Lebiere, R. C. O'Reilly, and John Anderson. 2008. "SAL: An Explicitly Pluralistic Cognitive Architecture." *Journal of Experimental & Theoretical Artificial Intelligence* 20 (3): 197–218.

Klinger, E. 1975. "Consequences of Commitment to and Disengagement from Incentives." *Psychological Review* 82: 1–25.

Kotovsky, K., J. R. Hayes, and H. A. Simon. 1985. "Why Are Some Problems Hard? Evidence from Tower of Hanoi." *Cognitive Psychology* 17 (January): 248–94.

Kubota, K., and H. Niki. 1971. "Prefrontal Cortical Unit Activity and Delayed Alternation Performance in Monkeys." *Journal of Neurophysiology* 34 (3): 337–47.

Lambon-Ralph, Matthew A., Elizabeth Jefferies, Karalyn Patterson, and Timothy T. Rogers. 2017. "The Neural and Computational Bases of Semantic Cognition." *Nature Reviews Neuroscience* 18 (1): 42–55. doi:10.1038/nrn.2016.150.

Lave, Jean, and Etienne Wenger. 1991. *Situated Learning: Legitimate Peripheral Participation.* Situated Learning: Legitimate Peripheral Participation. New York, NY, US: Cambridge University Press. doi:10.1017/CBO9780511815355.

Loftus, Elizabeth F., and John C. Palmer. 1974. "Reconstruction of Automobile Destruction: An Example of the Interaction Between Language and Memory." *Journal of Verbal Learning and Verbal Behavior* 13 (5): 585–89. doi:10.1016/S0022-5371(74)80011-3.

Luck, S. J., and E. K. Vogel. 1997. "The Capacity of Visual Working Memory for Features and Conjunctions." *Nature* 390 (December): 279.

MacLeod, C. M. 1991. "Half a Century of Research on the Stroop Effect: An Integrative Review." *Psychological Bulletin* 109 (June): 163–203.

Maier, Steven F, and Linda R Watkins. 2010. "Role of the Medial Prefrontal

Cortex in Coping and Resilience." *Brain Research* 1355 (October): 52–60.

Maier, Steven F., and Martin E. P. Seligman. 1976. "Learned Helplesness: Theory and Evidence." *Journal of Experimental Psychology: General* 105: 3–46.

Marr, D. 1969. "A Theory of Cerebellar Cortex." *Journal of Physiology (London)* 202 (January): 437–70.

———. 1971. "Simple Memory: A Theory for Archicortex." *Philosophical Transactions of the Royal Society of London B: Biological Sciences* 262 (841): 23–81. doi:10.1098/rstb.1971.0078.

Maslow, A. H. 1943. "A Theory of Human Motivation." *Psychological Review* 50: 370–96.

McClelland, J. L., and D. E. Rumelhart. 1986. "A Distributed Model of Human Learning and Memory." In *Parallel Distributed Processing. Volume 2: Psychological and Biological Models*, edited by J. L. McClelland, D. E. Rumelhart, and PDP Research Group, 170–215. Cambridge, MA: MIT Press.

McClelland, J. L., B. L. McNaughton, and R. C. O'Reilly. 1995. "Why There Are Complementary Learning Systems in the Hippocampus and Neocortex: Insights from the Successes and Failures of Connectionist Models of Learning and Memory." *Psychological Review* 102 (3): 419–57.

Meltzoff, Andrew N., and M. K. Moore. 1994. "Imitation, Memory, and the Representation of Persons." *Infant Behavior and Development* 17 (January): 83–99.

Miller, E. K., and R. Desimone. 1994. "Parallel Neuronal Mechanisms for Short-Term Memory." *Science (New York, N.Y.)* 263 (February): 520–22.

Miller, G. A., E. Galanter, and K. H. Pribram. 1960. *Plans and the Structure of Behavior*. New York: Holt.

Miller, George. 1956. *The Magical Number Seven, Plus or Minus Two: Some Limits on Our Capacity for Processing Information.* Vol. 101. Indiana : Bobbs-Merrill.

Miyake, Akira, Michael J. Emerson, Francisca Padilla, and Jeung-chan Ahn. 2004. "Inner Speech as a Retrieval Aid for Task Goals: The Effects of Cue Type and Articulatory Suppression in the Random Task Cuing Paradigm." *Acta Psychologica* 115 (February): 123–42.

Mnih, Volodymyr, Koray Kavukcuoglu, David Silver, Andrei A. Rusu, Joel Veness, Marc G. Bellemare, Alex Graves, et al. 2015. "Human-Level Control Through Deep Reinforcement Learning." *Nature* 518 (7540): 529–33.

Mollick, Jessica A., Thomas E. Hazy, Kai A. Krueger, Ananta Nair, Prescott Mackie, Seth A. Herd, and R. C. O'Reilly. 2018, submitted. "A Systems-Neuroscience Model of Phasic Dopamine."

Montague, P. Read, Peter Dayan, and Terrence J. Sejnowski. 1996. "A Framework for Mesencephalic Dopamine Systems Based on Predictive Hebbian Learning." *The Journal of Neuroscience* 16 (5): 1936–47.

Morris, Richard G. M. 2001. "Episodiclike Memory in Animals: Psycho-

logical Criteria, Neural Mechanisms and the Value of Episodiclike Tasks to Investigate Animal Models of Neurodegenerative Disease." *Philosophical Transactions of the Royal Society of London B: Biological Sciences* 356 (1413): 1453–65. doi:10.1098/rstb.2001.0945.

Neisser, Ulric, Gwyneth Boodoo, Thomas Bouchard, Nathan Brody, Stephen Ceci, Diane Halpern, John Loehlin, Robert Perloff, Robert Sternberg, and Susana Urbina. 1996. "Intelligence: Knowns and Unknowns." *American Psychologist* 51 (2): 77–101.

Newell, A., and H. A. Simon. 1972. *Human Problem Solving*. Englewood Cliffs, NJ: Prentice-Hall.

Newman, Ehren L, and Kenneth A Norman. 2010. "Moderate Excitation Leads to Weakening of Perceptual Representations." *Cerebral Cortex* 20 (11): 2760–70.

Norman, Kenneth A., and R. C. O'Reilly. 2003. "Modeling Hippocampal and Neocortical Contributions to Recognition Memory: A Complementary-Learning-Systems Approach." *Psychological Review* 110 (4): 611–46.

Ongür, D., and J. L. Price. 2000. "The Organization of Networks Within the Orbital and Medial Prefrontal Cortex of Rats, Monkeys and Humans." *Cerebral Cortex* 10 (3): 206–19.

O'Reilly, R. C. 1996. "Biologically Plausible Error-Driven Learning Using Local Activation Differences: The Generalized Recirculation Algorithm." *Neural Computation* 8 (5): 895–938.

O'Reilly, R. C., and Michael J. Frank. 2006. "Making Working Memory Work: A Computational Model of Learning in the Prefrontal Cortex and Basal Ganglia." *Neural Computation* 18 (2): 283–328.

O'Reilly, R. C., and J. L. McClelland. 1994. "Hippocampal Conjunctive Encoding, Storage, and Recall: Avoiding a Tradeoff." *Hippocampus* 4 (6): 661–82.

O'Reilly, R. C., Yuko Munakata, Michael J. Frank, Thomas E. Hazy, and Contributors. 2012. *Computational Cognitive Neuroscience*. Wiki Book, 1st Edition, URL: http://ccnbook.colorado.edu.

O'Reilly, R.C. 2006. "Biologically Based Computational Models of High-Level Cognition." *Science* 314 (5796): 91–94.

Powers, William T. 1973. *Behavior: The Control of Perception*. Hawthorne.

Quirk, Gregory J., and Devin Mueller. 2008. "Neural Mechanisms of Extinction Learning and Retrieval." *Neuropsychopharmacology* 33 (1): 56–72.

Quiroga, R. Quian, L. Reddy, G. Kreiman, C. Koch, and I. Fried. 2005. "Invariant Visual Representation by Single Neurons in the Human Brain." *Nature* 435 (7045): 1102–7. doi:10.1038/nature03687.

Raven, J. C., J. H. Court, and J. Raven. 1977. *Standard Progressive Matrices*. London: H. K. Lewis.

Read, Stephen J., Brian M. Monroe, Aaron L. Brownstein, Yu Yang, Gurveen Chopra, and Lynn C. Miller. 2010. "A Neural Network Model of the Structure

and Dynamics of Human Personality." *Psychological Review* 117 (1): 61–92.

Redish, A. D. 2004. "Neuroscience: Addiction as a Computational Process Gone Awry." *Science* 306 (5703): 1944–6.

Reike, F., D. Warland, R. van Steveninck, and W. Bialek. 1996. *Spikes: Exploring the Neural Code*. Cambridge, MA: MIT Press.

Rescorla, R. A., and A. R. Wagner. 1972. "A Theory of Pavlovian Conditioning: Variation in the Effectiveness of Reinforcement and Non-Reinforcement." In *Classical Conditioning II: Theory and Research*, edited by A. H. Black and W. F. Prokasy, 64–99. New York: Appleton-Century-Crofts.

Riva-Posse, Patricio, Ki Sueng Choi, Paul E Holtzheimer, Cameron C McIntyre, Robert E Gross, Ashutosh Chaturvedi, Andrea L Crowell, Steven J Garlow, Justin K Rajendra, and Helen S Mayberg. 2014. "Defining Critical White Matter Pathways Mediating Successful Subcallosal Cingulate Deep Brain Stimulation for Treatment-Resistant Depression." *Biological Psychiatry*, April.

Roediger, Henry L., and Kathleen B. McDermott. 1995. "Creating False Memories: Remembering Words Not Presented in Lists." *Journal of Experimental Psychology: Learning, Memory, and Cognition* 21 (4): 803–14. doi:10.1037/0278-7393.21.4.803.

Roumis, Demetris K, and Loren M Frank. 2015. "Hippocampal Sharp-Wave Ripples in Waking and Sleeping States." *Current Opinion in Neurobiology*, Circuit plasticity and memory, 35 (December): 6–12. doi:10.1016/j.conb.2015.05.001.

Rudebeck, Peter H., Mark E. Walton, Angharad N. Smyth, David M. Bannerman, and Matthew F. S. Rushworth. 2006. "Separate Neural Pathways Process Different Decision Costs." *Nature Neuroscience* 9 (9): 1161–8.

Rudy, Jerry. 2013. *The Neurobiology of Learning and Memory*. Second Edition. Oxford, New York: Oxford University Press.

Rumelhart, D. E., and J. L. McClelland. 1986. "PDP Models and General Issues in Cognitive Science." In *Parallel Distributed Processing. Volume 1: Foundations*, edited by D. E. Rumelhart, J. L. McClelland, and PDP Research Group, 110–46. Cambridge, MA: MIT Press.

Rumelhart, D. E., G. E. Hinton, and R. J. Williams. 1986. "Learning Representations by Back-Propagating Errors." *Nature* 323 (9): 533–36.

Schneider, W., and R. M. Shiffrin. 1977. "Controlled and Automatic Human Information Processing: I. Detection, Search, and Attention." *Psychological Review* 84 (January): 1–66.

Schultz, W. 1986. "Responses of Midbraine Dopamine Neurons to Behavioral Trigger Stimuli in the Monkey." *Journal of Neurophysiology* 56 (January): 1439–62.

Schultz, W., P. Dayan, and P. R. Montague. 1997. "A Neural Substrate of Prediction and Reward." *Science* 275 (5306): 1593–9.

Semendeferi, K., A. Lu, N. Schenker, and H. Damasio. 2002. "Humans and Great Apes Share a Large Frontal Cortex." *Nature Neuroscience* 5 (January):

272–76.

Shiffrin, R. M., and W. Schneider. 1977. "Controlled and Automatic Human Information Processing: II. Perceptual Learning, Automatic Attending, and a General Theory." *Psychological Review* 84 (January): 127–90.

Silver, David, Julian Schrittwieser, Karen Simonyan, Ioannis Antonoglou, Aja Huang, Arthur Guez, Thomas Hubert, et al. 2017. "Mastering the Game of Go Without Human Knowledge." *Nature* 550 (7676): 354–59. doi:10.1038/nature24270.

Simons, Daniel J., Walter R. Boot, Neil Charness, Susan E. Gathercole, Christopher F. Chabris, David Z. Hambrick, and Elizabeth A. L. Stine-Morrow. 2016. "Do 'Brain-Training' Programs Work?" *Psychological Science in the Public Interest* 17 (3): 103–86. doi:10.1177/1529100616661983.

Sperling, George. 1960. "The Information Available in Brief Visual Presentations." *Psychological Monographs: General and Applied* 74 (11): 1–29. doi:10.1037/h0093759.

Squire, L. R. 1992. "Declarative and Nondeclarative Memory: Multiple Brain Systems Supporting Learning and Memory." *Journal of Cognitive Neuroscience* 4 (3): 232–43.

Stocco, A., C. Lebiere, and J.R. Anderson. 2010. "Conditional Routing of Information to the Cortex: A Model of the Basal Ganglia's Role in Cognitive Coordination." *Psychological Review* 117: 541–74.

Stroop, J. R. 1935. "Studies of Interference in Serial Verbal Reactions." *Journal of Experimental Psychology* 18 (January): 643–62.

Stuss, D. T., D. Floden, M. P. Alexander, B. Levine, and D. Katz. 2001. "Stroop Performance in Focal Lesion Patients: Dissociation of Processes and Frontal Lobe Lesion Location." *Neuropsychologia* 39 (May): 771–86.

Sutherland, Robert J, Jamus O'Brien, and Hugo Lehmann. xx 2008. "Absence of Systems Consolidation of Fear Memories After Dorsal, Ventral, or Complete Hippocampal Damage." *Hippocampus* 18 (7): 710–18.

Sutton, R. S., and A.G. Barto. 1981. "Toward a Modern Theory of Adaptive Networks: Expectation and Prediction." *Psychological Review* 88 (2): 135–70.

Thorndike, E. L. 1911. *Animal Intelligence: Experimental Studies*. New York: The MacMillan Company.

Tolman, E.C. 1948. "Cognitive Maps in Rats and Men." *Psychological Review* 55 (4): 189–208.

Tulving, E. 1972. "Episodic and Semantic Memory." In *Organization of Memory*, edited by E. Tulving and W. Donaldson, 381–403. San Diego, CA: Academic Press.

———. 1983. *Elements of Episodic Memory*. Oxford, England: Clarendon Press.

Turing, A. M. 1936. "On Computable Numbers, with an Application to the Entscheidungsproblem." *Proceedings of the London Mathematical Society* s2-42

(1): 230–65. doi:10.1112/plms/s2-42.1.230.

von Neumann, John. 1945. “First Draft of a Report on the EDVAC.”

Wallis, J D, R Dias, T W Robbins, and A C Roberts. 2001. “Dissociable Contributions of the Orbitofrontal and Lateral Prefrontal Cortex of the Marmoset to Performance on a Detour Reaching Task.” *The European Journal of Neuroscience* 13 (May).

Wallis, Jonathan D., and Steven W. Kennerley. 2011. “Contrasting Reward Signals in the Orbitofrontal Cortex and Anterior Cingulate Cortex.” *Annals of the New York Academy of Sciences* 1239 (December): 33–42.

Wason, P. 1968. “Reasoning About a Rule.” *Quarterly Journal of Experimental Psychology* 20 (January): 273–81.

Wilson, M. A., and B. L. McNaughton. 1994. “Reactivation of Hippocampal Ensemble Memories During Sleep.” *Science (New York, N.Y.)* 265 (August): 676–78.

Yerkes, R. M., and J. D. Dodson. 1908. “The Relation of Strength of Stimulus to Rapidity of Habit Formation.” *Journal of Comparative Neurology and Psychology* 18 (January): 459–82.