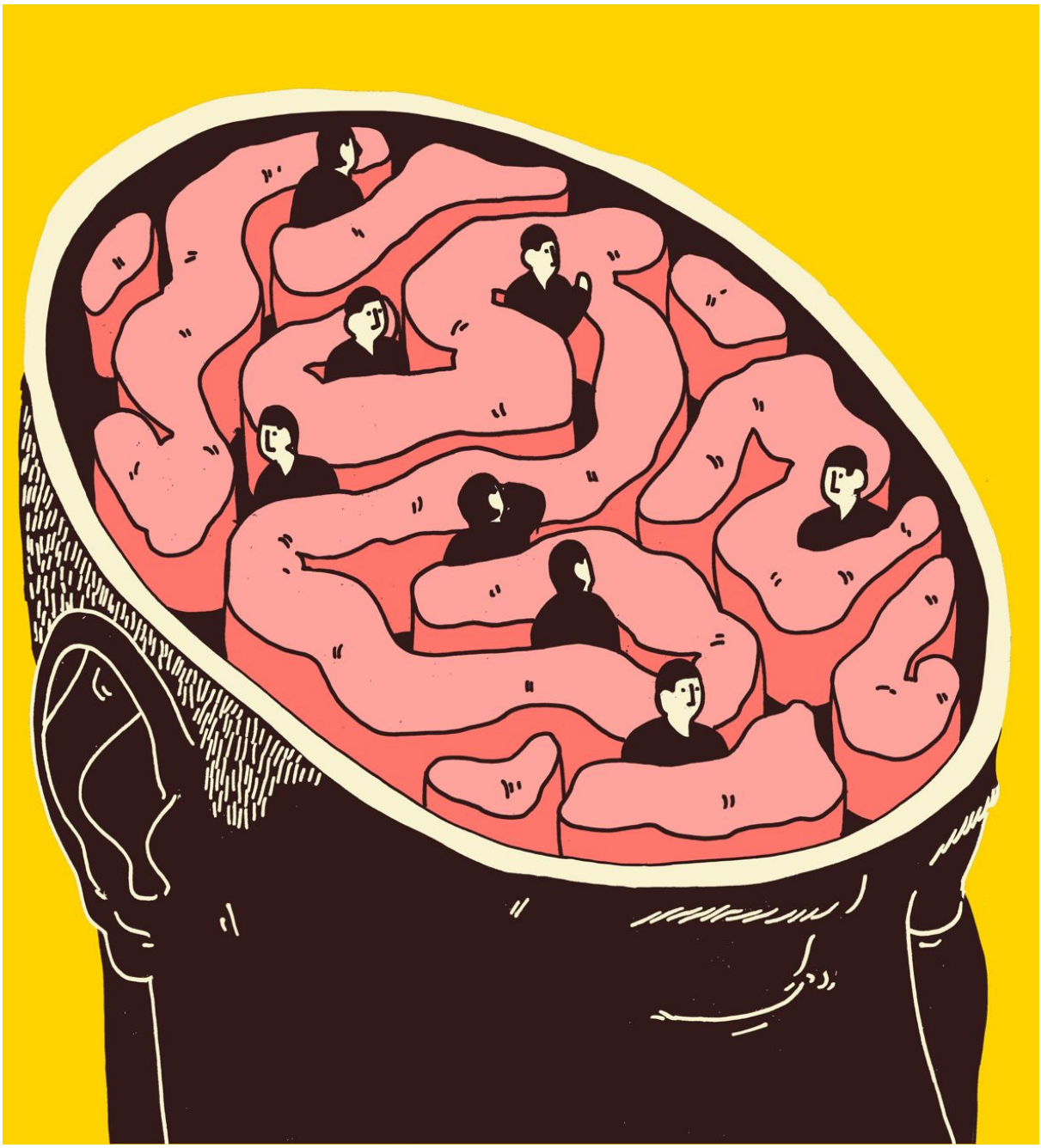
Neuroscience is pre-paradigmatic. Consciousness is why

Nothing in the brain makes sense except in the light of consciousness

Erik Hoel

Date: Jan 09, 2024

URL: https://www.theintrinsicperspective.com/p/neuroscience-is-pre-paradigmatic



Art for *The Intrinsic Perspective* is by Alexander Naughton

If academia is under question, it should be for things that are fundamental. So let's ask: Has neuroscience, my own discipline, *fundamentally* progressed since the turn of the millennium?

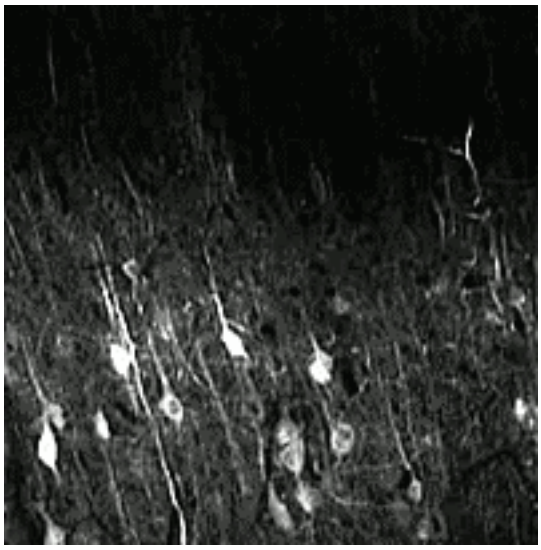
The term "fundamental" is doing a lot of work in this question, no doubt. But if you take stock of the progress in neuroscience since 2000, it's noticeable that the most popular papers are more aptly described as "cool" rather than fundamental. If you look at, say, *Scientific American*'s 2022 list of "This year's most thought-provoking brain discoveries," the results are firmly in the cool

category, not the fundamental. What about 2023? Let's look at a Chair of Computational Neuroscience's 2023 review of the field's year: again, merely cool results, except for how the question of what dopamine's role in the brain is, *fundamentally*, has suddenly come up for debate again, and it turns out no one knows the answer. What if we get more objective and examine the most-cited articles published since 2021 in the journal *Neuroscience*? The pattern continues. Cool, you can use micro-RNA to promote spinal cord regeneration in rats! Cool, you can use deep learning to help diagnose cognitive impairment using fMRI! Fundamental? No.

Getting even more objective, in 2017 researchers used a bibliometric analysis to examine the 100-most-cited neuroscience papers; almost none were from past 2000. Have recent results just not had time to accumulate citations? Nope, since the majority of the most-cited neuroscience papers are from the 90s, just a decade earlier. Perhaps this is because, since the turn of the millennium, a lot of neuroscience has focused on refining methodologies—things like optogenetics and advances in calcium imaging and growing cerebral organoids—which are then used to generate cool papers.

But coolness can disguise all sorts of problems. Neuroscience has not gone through the public fall from grace that its sister field, psychology, has endured. Since 2000 reproducibility has become a popular watch-word in psychology due to classic much-touted psychological effects failing to replicate. Stanford Prison Experiment? Out. Dunning-Kruger Effect? Out. Ego depletion? Out. Yet neuroscience, which is essentially psychology on hard mode (asking similar questions but with the added difficulty of accessing the brain), has been spared public skepticism. The machines are so fancy, the data so beautiful, the cool factor undeniable.

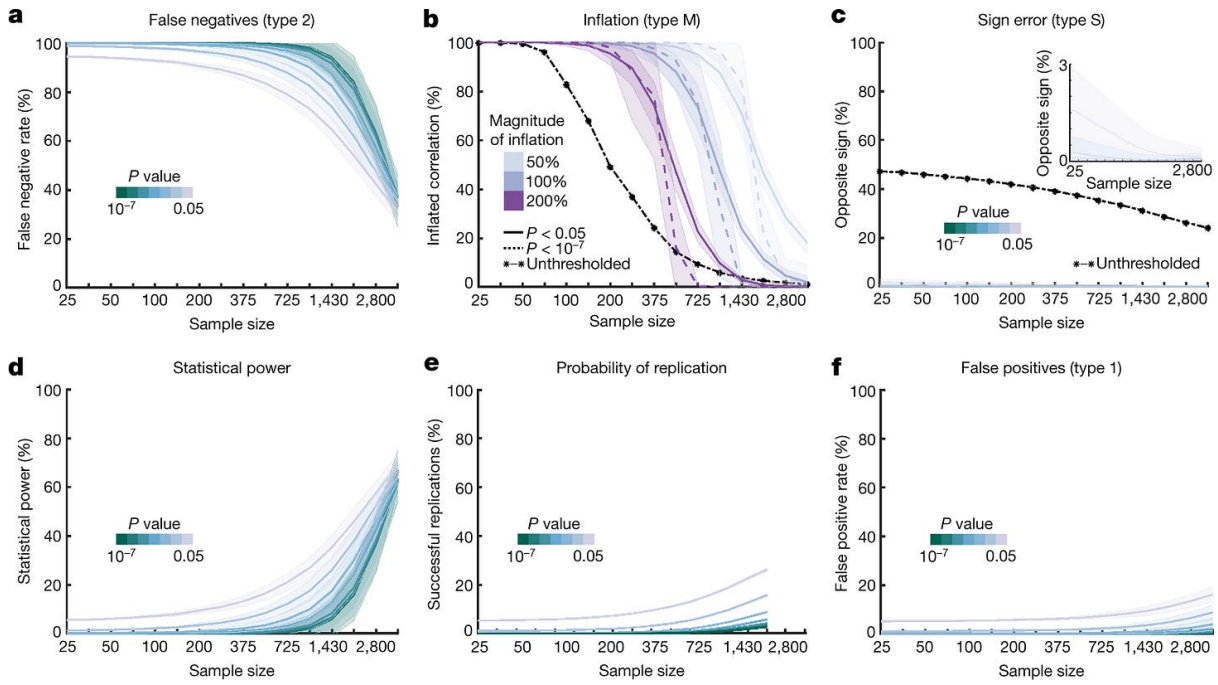
And there is no doubt that we have made progress on our understanding of individual neurons, like finding that human neurons can perform complicated logical functions inside themselves. Additionally, there have been undeniable practical advances. We can tell, for instance, if a patient in a supposed vegetative state, one who cannot even move the muscles of their eyes, is actually conscious, simply by asking them to imagine something like playing tennis and then seeing if the activation looks the same as in normal brains. We can record more and more neurons at finer spatiotemporal scales. I've stood in university labs and watched in real time neurons inside a rat's brain spark white with calcium waves as they fire. But even then, the activity looked to me as random as Christmas tree lights, because honestly that's how it always looks. I got no more information than from staring into the affixed rat's inscrutable red eyes.



calcium imaging from live rat cortex

Of course, ever since researchers in the mid-aughts threw a dead salmon into an fMRI machine and got statistically significant results, it's been well-known that neuroimaging, the working core of modern neuroscience, has all sorts of problems. It's impossible to walk through every paper, and many are, in their choices of methods and analysis, individually defensible. But when considered collectively, we have very good reasons to be skeptical.

E.g., we now know that reproducible effects in neuroimaging likely have a far higher bar than the one practiced by working neuroscientists. As the title of a 2022 *Nature* paper bluntly puts it: "Reproducible brain-wide association studies require thousands of individuals." That's in comparison to the mere dozens used in the overwhelming majority of papers. Correlating brain states to psychological states (like searching for the neural correlates of depression, or something equivalent) requires sample sizes in the thousands, far to the right on the x-axis below. Yet most neuroscience takes place to the left, in the shadowland of false negatives, false positives, and poor reproducibility.



Expected error rates for neuroimaging correlations

This dirty secret is why the only neuroimaging I ever tried to directly replicate showed opposite results (the original paper still has hundreds of citations), and over my career I have personally seen many puzzling failures of well-known effects that left postdocs and graduate students privately stumped.

The problem goes beyond the paucity of data; it's the analyses and techniques themselves. Data from any single trial of a neuroimaging experiment is almost always a complete mess; looking at them is a "marginal" activity for neuroscientists. Therefore, the constant background assumption is that the noisiness of the brain can be smoothed out by averaging—that if you average enough, the "real" signal will pop out. Yet, does the brain even care about the statistical averages neuroscientists publish about? After all, *you* don't think in averages. Shown visual stimuli in a neuroscience experiment, you'd experience one thing, and then another, and they'd each be crystal clear. The neuroscientist's data wouldn't reflect that. Does the neuroscientist's average even have anything to do with what you saw?

Be clever enough with experimental design and you can test this. In 2021 researchers proposed two criteria that, if the brain did care about the averages neuroscientists put in their papers, would have to be true:

(a) *that neural responses repeat consistently enough that their averaged response should be theoretically recognizable to downstream regions.*

(b) *that single-trial responses more like the averaged response should be correlated with the animal doing better on the task (like identifying a visual stimulus).*

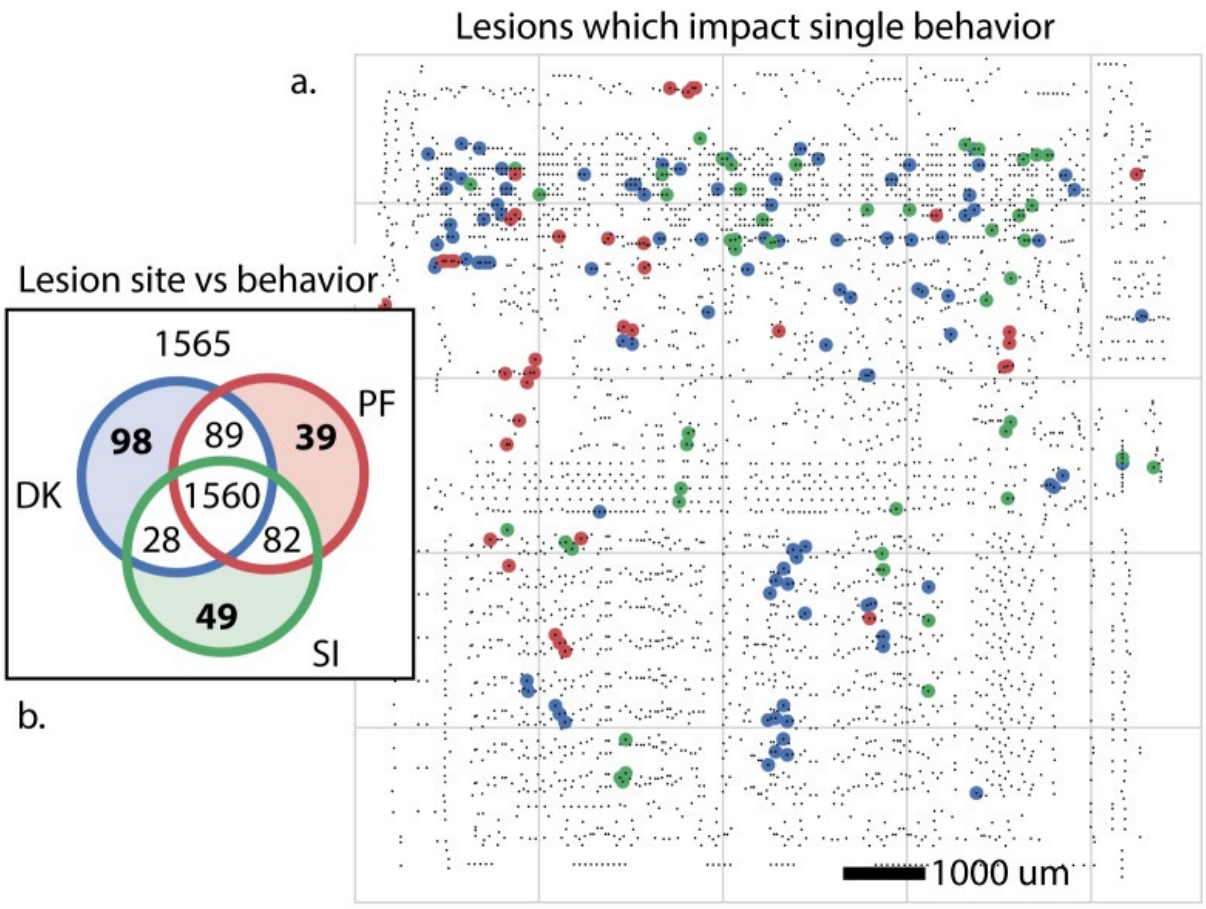
Lo and behold, they found that neither **(a)** nor **(b)** are true. When the researchers started correlating electrophysiological recordings in mice brains to behavior, they found that identifying the average from the single-trial response was extremely difficult even for ideal observers, and that how well the single-trial response matched the "platonic" average had almost no bearing on behavior. Despite a plea by the authors to perform the same analysis on other experiments, that clever paper currently has one citation, and still has not been let through peer review (despite its methods looking sound). No one wants to hear its message.

Neuroscientists' focus on epiphenomenal statistical constructs might explain why the story of modern neuroscience has so often been: "Whoops, function *x* is more complex than we thought" (like how the classic motor homunculus is actually much more variating than in textbooks). More and more, scientists are realizing that the entire brain is at work for any given phenomenon (predicted, by the way, in a novel of mine several years ago), such as the recent revelation that "'Language regions' are artefacts of averaging."

The constant revision toward more shoulder-shrugging complexity, and the concomitant drop of clarity in explanations, is elided by the field's jargon. In neuroscience, the terms "computation" or "representation" or "information processing" or even "storage" or "retrieval" are used interchangeably, and vary between subfields. Only rarely does it signify a real difference in the neural activity being observed. What most terms really mean is "the neurons fired differently between my conditions." But what does the difference *mean*?

For example, in the 2017 paper "Can a neuroscientist understand a microprocessor?" the authors apply a suite of popular techniques from neuroscience to a system far less complicated than the brain itself: the MOS 6502 microchip (used in the 80s to power Nintendos). Despite having only 3,510 transistors it can run a few games like Space Invaders, Donkey Kong, and Pitfall.

Pretending to not understand the function of the chip, the researchers attempted the common techniques of neuroscience: looking at the connectomics (the wiring diagram of the chip); performing "lesions" on the transistors to mimic studies wherein a part of the brain is damaged; analyzing the individual transistor behavior using techniques usually thrown at neurons (like looking at firing rates or tuning curves); correlating the transistors to each other in pairs; averaging the chip's activity into fMRI-like voxels; along with other more esoteric analyses like dimension reduction and Granger causality. Many pretty graphs could be made, yet, despite having the complete microscale information we wish we knew for the brain, the conclusions were utterly obscure. E.g., lesions to ~50% of the transistors did nothing, while the remaining ~50% shut down the chip completely.



Knocking out parts of the chip vs. what games are affected

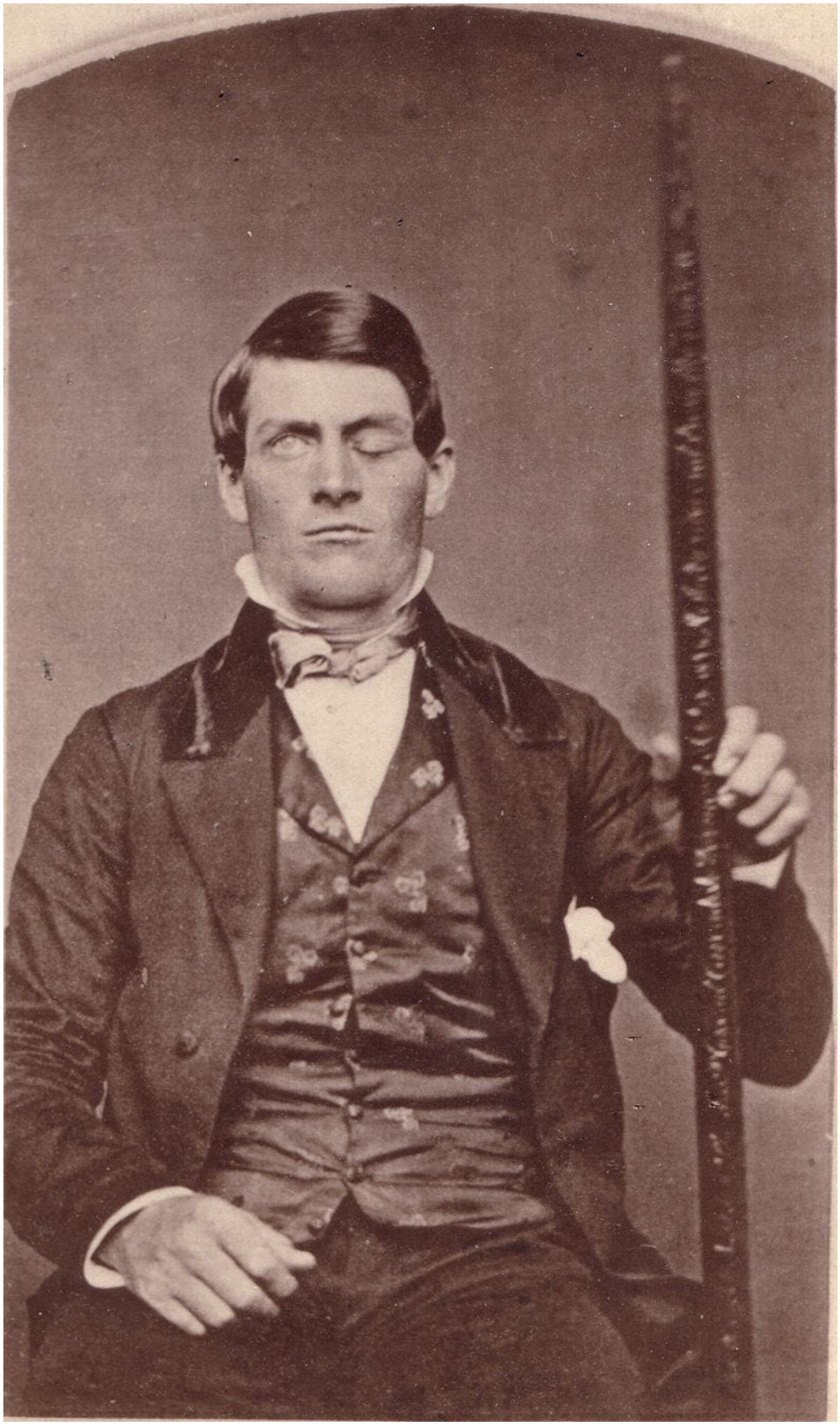
A small number were game-specific, like the 98 "lesions" that led to the unique failure of Donkey Kong! Aha! The Donkey Kong transistors have been found! An equivalent result in the brain would net a *Nature* paper for some hungry young graduate student. And yet, the reason the game failed under these lesions turned out to have nothing to do with Donkey Kong; it was just a fluke in the back-end code. You'd never be able to derive a single thing about barrel-throwing or bananas from the knowledge. Precisely these difficulties is why

neuroscience oversells low-information explanations.

Here's a cruel parlor trick: ask a neuroscientist to explain something about human cognition. Listen to their answer. Then ask them to explain again, but to say nothing in their explanation about location. The second ask is far harder, almost embarrassingly so (and I say this as a neuroscientist by training). "Speech is processed in Wernicke's… no wait."

Both popular science books about neuroscience, as well as most introductory textbooks, lead with spatial information. The prefrontal cortex. The temporal lobe. V1. The entorhinal cortex. The medial superior olive. You've heard plenty of this; e.g., college courses in neuroscience often start with Phineas Gage, the railroad worker in the 1800s who ended up with a metal rod blasting through his eye socket and out of his head. He miraculously survived, but his

personality supposedly changed; therefore, the story goes, the prefrontal cortex (what was damaged) surely houses that most high-level psychological construct, the self.



Gage with his infamous rod

Even if we put aside that his behavioral changes are likely an exaggerated scientific myth, if personality had instead been located to the side of the brain, rather than the front, what would that change? If you ask me how a car works, and I say "well right here is the engine, and there are the wheels, and the steering wheel, that's inside," and so on, you'd quickly come to the conclusion that I have no idea how a car actually works.

And unlike a car, the brain is extremely plastic, so spatial location is even more irrelevant. The actual physical architecture is always roiling, changing beneath our microscopes. In the macaque visual cortex, about 7% of synaptic boutons are turned over every week (boutons are the points of close contact between axons and dendrites, where neuronal communication occurs). There is now well-documented "representational drift" wherein parts of the brain shift what they respond to, not just at a timescale of days, but at a timescale of minutes. Representations and functions drift like clouds across the cortex. Just last year, it was shown representational drift can be seen on fMRI in humans; even in the visual cortex, which, we were all told, is supposed to have a stable architecture to perform the stable function of *seeing.* Such drift can be found in simpler artificial neural networks as well. What shenanigans are happening in the higher levels of the human brain?

And there are far harder questions for neuroscientists than just to explain phenomena without reference to spatial location. In fact, let's ask

the hardest question in neuroscience.

It's a very simple one, and devastating in its simplicity. If even small artificial neural networks are mathematical black boxes, such that the people who build them don't know how to control them, nor why they work, nor what their true capabilities are… why do you think that the brain, a higher-parameter and more messy biological neural network, is not similarly a black box?

And if the answer is that it is; indeed, that we should expect even greater opacity from the brain compared to artificial neural networks, then why does every neuroscience textbook not start with that? Because it means neuroscience is not just hard, it's closing in on *impossibly hard* for intelligible mathematical reasons. AI researchers struggle with these reasons every day, and that's despite having perfect access; meanwhile neuroscientists are always three steps removed, their choice of study locked behind blood and bone, with only partial views and poor access. Why is the assumption that you can just fumble at the brain, and that, if enough people are fumbling all together, the truth will reveal itself?

This skepticism is supported by a historical analysis of the relationship between AI research and neuroscience. If neuroscience had made a lot of *fundamental* progress since 2000, then AI researchers curious about interpretability (how their neural networks work) could simply look at neuroscience and steal the techniques already developed there, adapting them as necessary. And while such transfers do occasionally occur, they are piecemeal, rare, and unsatisfactory.

Most of the time, AI researchers are on their own. That's why the best recent paper on black box interpretability, published last year by Anthropic, cites only *one paper* that resembles a traditional neuroscience paper you might read in graduate school. One paper out of 75 (and it's possibly the least important citation). The slacking transferability of neuroscientific knowledge is one of the surest signs that

neuroscience really is pre-paradigmatic.

But what does this mean? When does a scientific field count as pre-paradigmatic? Because it's not the same as saying that the field is equivalent to astrology, or a complete waste of time. No, a pre-paradigmatic science, at least according to Thomas Kuhn's original conception of science proceeding via "paradigm shifts," is when a field is marked by insecurity over an anomaly. In *The Structure of Scientific Revolutions* Kuhn writes:

> Galileo's contributions to the study of motion depended closely upon difficulties discovered in Aristotle's theory by scholastic critics. Newton's new theory of light and color originated in the discovery that none of the existing pre-paradigms theories would account for the length of the spectrum, and the wave theory that replaced Newton's was announced in the midst of growing concern about **anomalies** in the relation of diffraction and polarization effects to Newton's theory. Thermodynamics was born from the collision of two existing nineteenth-century physical theories, and quantum mechanics from a variety of difficulties surrounding black-body radiation, specific heats, and the photoelectric effect. Furthermore, in all these cases except that of Newton the awareness of **anomaly** had lasted so long and penetrated so deep that one can appropriately describe the fields affected by it as in a state of growing crisis… the emergence of new theories is generally preceded by a period of pronounced professional insecurity.

In all those fields Kuhn mentioned, there were papers (or their historical equivalent) being published, experiments being done, all right before the paradigm revolution. The tell of being ripe for revolution is both deep troubles *and* that there is an unexplained anomaly that haunts the field. And there is an obvious anomaly in neuroscience: *consciousness*.

For to this day, neuroscience has no well-accepted theory of consciousness—what it is, how it works. While there is a small subfield of neuroscience engaged in what's called the "search for the neural correlates of consciousness," it took the efforts of many brave scholars, spear-headed by heavy-weights like two Nobel-Prize winners, Francis Crick and Gerald Edelman, to establish its marginal credibility. Previously, talk of consciousness had been effectively banished from science for decades (I've called this time the "consciousness winter" and shown that even the word "consciousness" declined in usage across culture as a whole during its reign).

Yet, if you go back and read the original psychologists and neuroscientists like William James or Wilhelm Wundt, while they were just as unsure about the ultimate metaphysical nature of consciousness as we are today—they too didn't know how nerve cells firing is accompanied by subjective experience—it's obvious that they believed what every lay person already knows,

which is that

consciousness is the primary function of the brain.

You wake up to a stream of consciousness, which is your perception of the world and its content, with you at the subjective center of the experience. Throughout the day this stream is the reason for everything you do. You get water because you're thirsty, take turns while driving because you have a destination in mind, talk because you're feeling gregarious. All of our cognitive functions, from memory to attention to cognition, take place within the domain of an ongoing stream of consciousness. It's the *raison d'être* of the brain as an organ. Indeed, our very survival depends on our consciousness being veridical and richly informative, and the pleasure or pain it provides, and all the valences in-between, guide our actions every day. Automatic or unconscious processes may underpin it, they may support it like marvelous butlers, but the stream of consciousness is the *point*. This is precisely the long-ignored main thesis of William James' field-inaugurating 1890 *The Principles of Psychology*:

> We talk, it is true, when we are darwinizing, as if the mere body that owns the brain had interests; we speak about the utilities of its various organs and how they help or hinder the body's survival; and we treat the survival as if it were an absolute end, existing as such in the physical world… The organs themselves, and all the rest of the physical world, will, however, all the time be quite indifferent to this consequence… But the moment you bring a consciousness into the midst, survival ceases to be a mere hypothesis…. Real ends appear for the first time now upon the world's stage. The conception of consciousness as a purely cognitive form of being… is thoroughly antipsychological, **as the remainder of this book will show**. Every actually existing consciousness seems to itself at any rate to be a fighter for ends, of which many, but for its presence, would not be ends at all. Its powers of cognition are mainly subservient to these ends, discerning which facts further them and which do not.

Are there arguments against consciousness as the primary function of the brain, with powers of cognition its servants? Yes, but they must be weighted by the fact that, precisely because consciousness always seemed too ickily subjective for science, scholars have had strong non-scientific reasons to downplay its importance. Notably, I've long tracked how experimental findings used to argue for the insignificance of consciousness almost always end up being overturned, quietly, by later results.

E.g., for a long time it's been claimed that humans can perform complex actions without consciousness; the clearest case of this is a phenomenon given the oxymoronic name "blindsight." The canonical examples are patients with occipital lobe lesions. These patients can still perform tasks prompted from within their blindspot (the part of their visual field where they don't have any conscious perception)—they see without seeing. Or so the story goes.

The problem is that blindsight claims were always based on a tiny number of subjects—as in like three primary subjects known only as (for anonymity reasons) GY, DB, and TN. And yet these patients would report things like "feelings" or "visual pinpricks" or "dark shadows" or

“white halos” in their blindspots, which matters a lot when detecting the presence of the kind of stimuli used in experiments. Decades after blindsight rose to popularity in pop philosophy of mind books, precise experiments on the most-studied of the three, GY, revealed that, while his vision was indeed highly degraded, he still had visual experiences in his supposed blindspot. And in other cases, blindsight disappeared once the subject was allowed to answer outside binary confines and give gradations of perception. In fact, some of the most famous patients contradicted themselves, admitting to some researchers that they do indeed have degraded sight in their blindspot long after telling other researchers they didn’t! Besides, even if they were to be believed, it’s clear from the experiments that “seeing without seeing” requires extensive cueing and prompting in the experimental design, unlike normal vision.

As another example of how historical arguments against the primacy of consciousness end up overstated and oversold, consider the phenomenon of “change blindness.” In this experimental literature, viewers come across as disconnected from their own consciousness, like how most viewers supposedly don’t even notice a man in a gorilla suit moving through a crowd of people passing a basketball around.

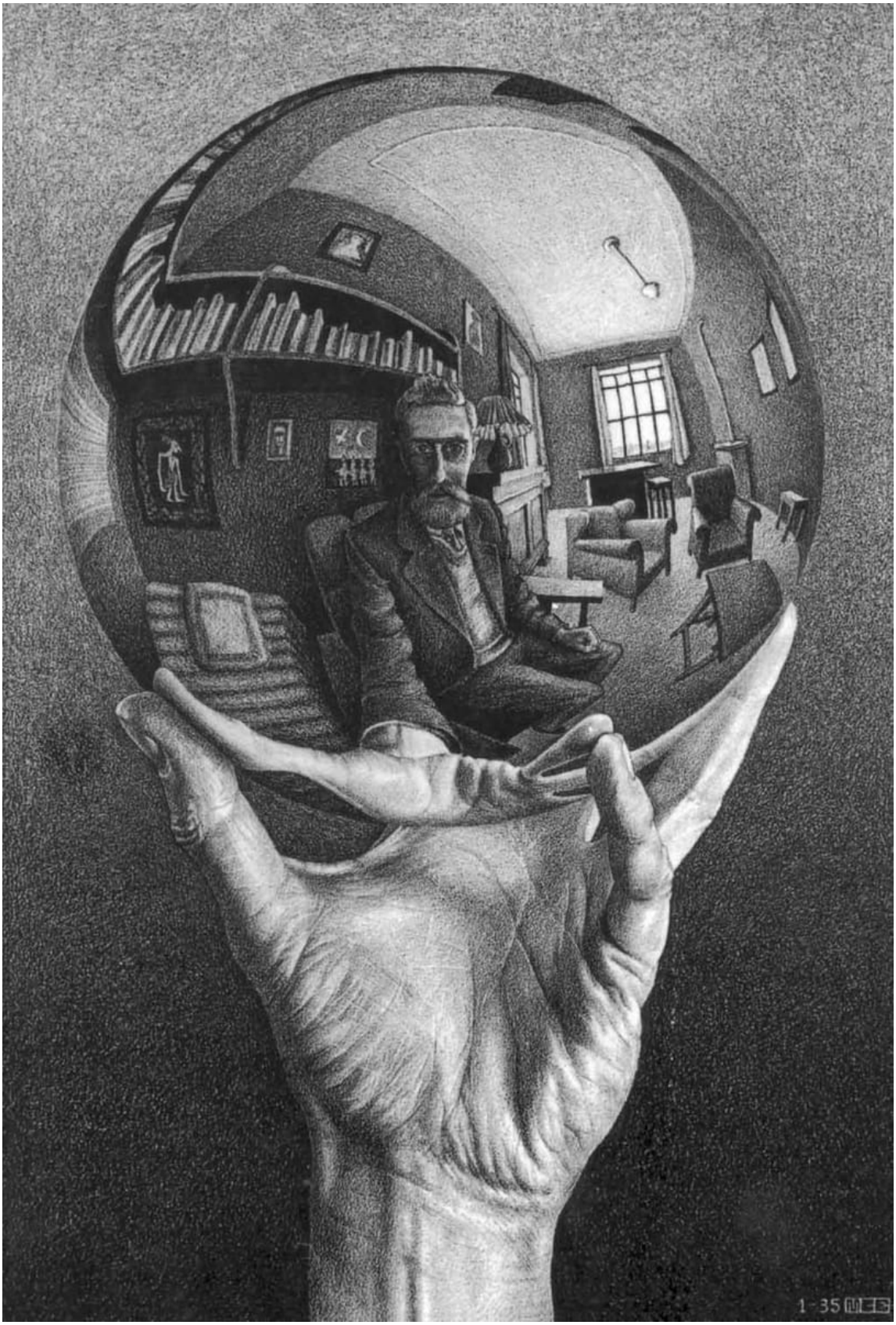
Frame from the original video used to demonstrate change blindness

Again, however, replications reveal the truth to be more complicated. In the most difficult condition, the original result reported only 8% of the subjects noticed the gorilla; in a recent replication attempt, it jumped to 60% (how many of the remaining, in turn, were bored undergraduates not paying attention at all?). At this rate, the change blindness effect will have

disappeared entirely in a decade, like so much of the rest of psychology, wherein the size of effects notoriously decline over time (likely due to a bias against common-sense for publication reasons).

In comparison, modern research shows your consciousness is probably about as richly detailed as you think. Consider just your visual consciousness: recent research has shown that humans can identify colors even in their furthest parts of peripheral vision, and indeed have a rich pointillist fusion of perceptions of the objects there (albeit coarse-grained), and that even when visually searching for a single face in a crowd humans form a huge number of instantaneous short-term memories of the non-target faces they can then recall. It's all pretty much what you'd expect from introspecting about your own phenomenology.

More on the theory side, as my scientific mentor in graduate school, Giulio Tononi, has pointed out, the vast numbers of possible conscious experiences (every frame of every possible movie) mathematically imply consciousness is highly informative, the same way a trillion-sided die would be highly informative. And not only that, the information in your consciousness is integrated, bound together in ways that information stored on a disk drive isn't.



"Hand with Reflecting Sphere," a self-portrait by M. C. Escher

Trusting our own introspection about the primacy of consciousness, and trusting how we all use consciousness to understand the behavioral repertoire of not just humans, but also other animals, the argument for neuroscience being pre-paradigmatic becomes very simple. So let's call it the Very Simple Proof of Pre-Paradigm-ism (VSPP). The VSPP goes:

P1: Consciousness is the primary function of the brain.

P2: There is no well-accepted theory of consciousness.

QED, neuroscience is pre-paradigmatic.

P2 no one can quibble with. P1 people could quibble with, but they usually base their arguments either on *a priori* metaphysical commitments (like eliminativism about consciousness) or oversold early psychology experiments about the minimal importance of consciousness, results that don't hold up anymore. Meanwhile, everything about the neuroscience's struggles is explained by the proof, and there's also a clear anomaly on offer that fits Kuhn's definition of a pre-paradigmatic field.

Just as the father of the modern synthesis of evolution, biologist Theodosius Dobzhansky, declared that "Nothing in biology makes sense except in the light of evolution," to move to a post-paradigmatic science neuroscientists must declare that "Nothing in the brain makes sense except in the light of consciousness."

I don't expect most neuroscientists to accept my arguments (although I've found that if you raise the same points privately while getting beers after a conference, the majority will agree with 80% of it). I'm sure many can cite dozens of studies showing what they deem as fundamental progress in their chosen fields; in fact, to many this will come across as insulting or derogatory.

While I don't mean it that way, fortunately or unfortunately, science doesn't care. And one must be forceful about traditional neuroscience's failures, because studying consciousness directly remains so difficult to get funding for, especially from government institutions. Even within the subfield of neuroscience where talk of consciousness is accepted, the number of scientists actually working with an open mind on new scientific theories of consciousness is extremely small. The number of those that are young and talented is even smaller still. If some rich philanthropist (or an ambitious startup) wants to have a major impact on scientific progress for minimal investment, this is the place, and consciousness is the question. Otherwise, neuroscience must await its *annus mirabilis*.

TIP occasionally gets outside sponsors who allow me to write and research while keeping the results open to all. This post is sponsored by Prophetic, a company researching consciousness and neurotechnology. It is co-posted on their blog. The content reflects only my own views. It contains small excerpts from my recent book *The World Behind the World: Consciousness, Free Will, and the Limits of Science*.