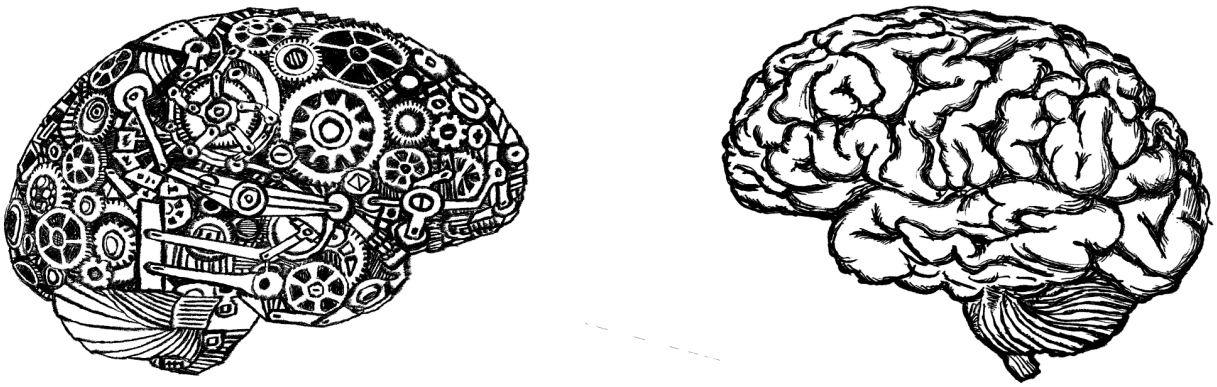
The Cyborg Myth

Text: Johannes Jaeger

Artwork: Marcus Neustetter



This argument was presented as an invited talk for the "*Hybrid Minds*" symposium — organized by the Telos Circle at the Medical University of Vienna on February 26, 2026. A re-recording with cleaned-up slides and sound is available here. I thank the organizer, Constantin Convalexius, for inviting me.

Many of the presentations at the symposium were concerned with the technology of neural implants and how they interface with the human brain. But in the background loomed the idea of using this kind of technology for *longevity* — the science of extending human life span, possibly forever. Unfortunately, investigations in this area are motivated by rather dubious philosophical assumptions and arguments.

First, there is the obvious objection that no one in their right mind should want to live forever, once they have properly thought the idea through. I refer the reader to Bernard Williams' brilliant and timeless essay "*The Makropulos case: reflections on the tedium of immortality*" [PDF]. Or maybe just watch Guillermo del Toro's recent adaptation of *"Frankenstein*" — the most faithful in spirit to Mary Shelley's original so far. You could also ponder why Q in *Star Trek* is such a cynical bastard. He's bored to death because he cannot die. Isn't it ironic? Immortal life, it turns out, is utterly devoid of meaning.

And, besides: isn't it a bit arrogant to presume that your particular life, your individual identity, is worth extending forever? Shouldn't we make space, when our time comes, for those who come after?

These may be valid doubts. But they are not really what concerns me here. Instead, I'd like to highlight another philosophical dispute that surfaced at the symposium, which revolves around the possibility of

"uploading" the human mind into a non-biological substrate — let's say the persistent memory of some sort of computer.

The problem is that any *squishy*, *fleshy* *embodiment of the mind* (no matter how much it has been enhanced and extended) remains prone to accidental harm and untimely demise. What shame if you could live forever, but get run over by a drunk driver at the age of 25! Life is full of nasty surprises. And, despite Hollywood stubbornly pretending otherwise, we all know that a biological clone is just an identical twin — a separate individual — not an actual replica of you as the person you are. Read Edward Ashton's "*Mickey7*" and "*Antimatter Blues*" — as entertaining as they are smart — if you do not believe me[1]. Then you will also understand why only *mind uploading* yields true immortality through redundant storage of your very own consciousness on interchangeable hardware. This classical science-fiction trope gives the idea of a "personal backup" a radically new meaning: minds can now be copied and propagated forever — across perishable physical instantiations.

[1] However, do *not* watch Bong Joon Ho's dreadful movie adaptation "*Mickey17*." It misses literally *every* philosophical point raised by the books — not to mention that it totally butchers the story as well as the characters, features the worst acting and introduces the most cringy, in-your-face attempts at political satire I've seen in a very long while …

But let's not beat about the bush: *mind uploading is utterly implausible*. In fact, I dare say, the probability of it ever happening and helping us achieve human immortality is basically zero.

For one, it involves tons of really daunting technological and logistic challenges without any practicable solutions anywhere in sight. Yet, tech-optimists believe those may be surmountable — if not tomorrow then eventually. And maybe they are right. I wouldn't know.

Still, there is a more serious problem underneath it all: the whole chain of reasoning "supporting" the idea of mind uploading is based on an elementary logical error. We can call it the *abstraction fallacy* or — with philosopher Alfred North Whitehead, who first stated it explicitly in 1925 — the *fallacy of misplaced concreteness*. Basically, it means *mistaking the map for the territory*: confusing the abstract and the concrete in our experience. This has been causing whole shiploads of seemingly irresolvable *paradoxes* concerning self-reference, personal identity, and subjective experience.

Evidently, to be uploadable, the mind must be somehow separable from its physical substrate. If we accept this premise, then we subscribe to an ideology called *computational functionalism*, which I abbreviate to *computationalism* in what follows. It treats the mind as something symbolic, constituted of *rule-based* (and hence *computational*) *patterns* — discretized and abstracted from the continuous tangle of underlying organ-level, tissue-mechanical, cellular, biochemical, and electrophysiological events, processes, and interactions. Only if computationalism applies can we artificially emulate or extract these higher-level patterns and project them onto unrelated kinds of physical substrates (the process of "uploading").

Somewhat ironically, computationalism springs from a hardcore reductionist and mechanistic approach to the study of mind, yet puts us on a slippery slope towards an anti-materialist *mind-body dualism*, as it invites the dissociation of mental and material phenomena from each other. Until recently, any whiff of such dualism would have been dismissed as heretic mysticism in well-respected scientific circles — its proponents driven from the community, intellectually tarred and feathered. But, as we shall see, computationalism has given it an unexpected kind of renaissance lately. And this is *not* a good thing.

Since its beginnings in the 1970s, computationalism has become established to an extent that we rarely see it questioned anymore, at least not in public. Beyond cognitive neuroscience and the philosophy of mind, its popularity is tethered to the current *AI hype cycle*, fueling wild-eyed speculations about artificial general intelligence (AGI), superintelligence (ASI), and sentient algorithmic systems. It has even led to the bizarre idea that we need some kind of "AI welfare:" public care not for humans who actually need it, but for lifeless machines. Many researchers and laypeople believe such talk to be supported by empirical evidence. But I assure you, it is not.

By its very nature, computationalism is a *metaphysical framework*. It is based on philosophical assumptions and arguments that are *not* experimentally testable, even in principle. Yet, many of us simply take them for granted, even if they lead to absurd conclusions. So, yes, it is okay to call this a dogmatic belief system — a cult, really. It certainly is *not* science.

The cultish nature of computationalism is further exposed by the fact that its foundational assumptions and arguments turn out to be incredibly shaky — logically inconsistent, even. As an example, I will pick out a particularly popular argument for computationalism and dismantle it in front of your eyes.

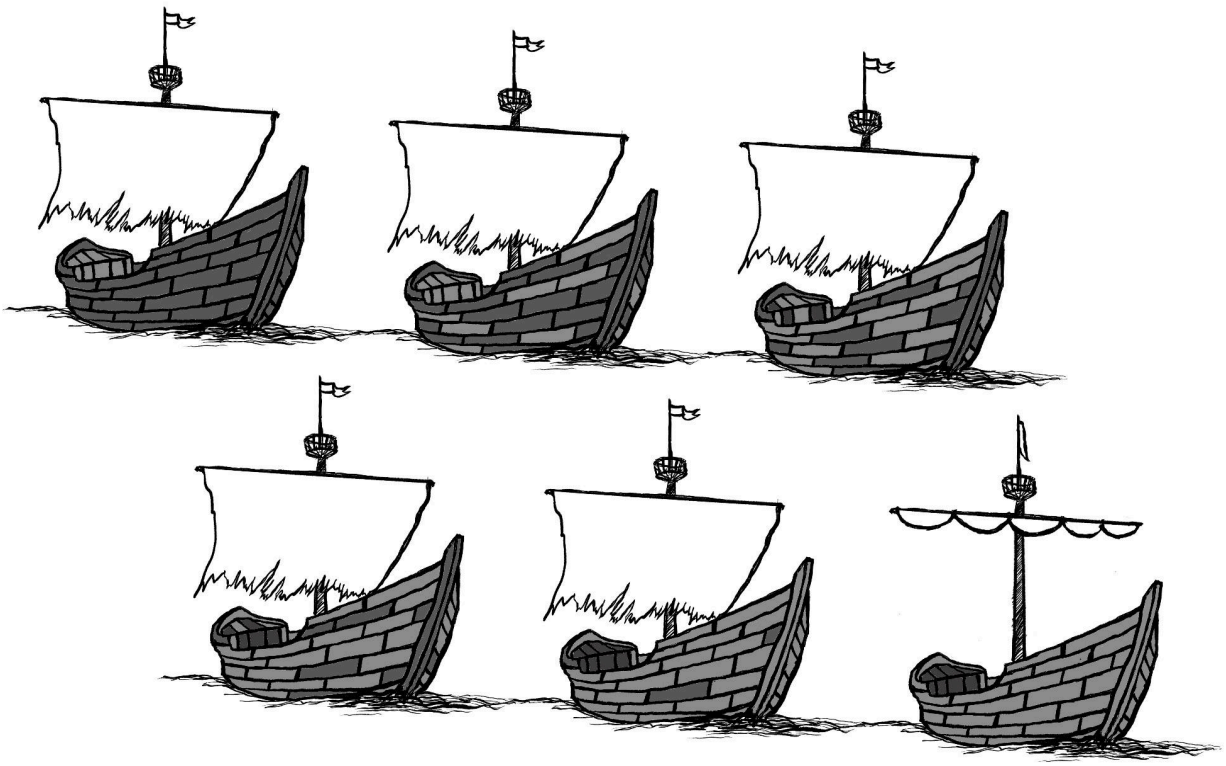
It is the famous *neural substitution* (or *neural replacement*) *thought experiment*. Austrian computer scientist Hans Moravec was the first to formulate it in his 1988 book "*Mind Children: The Future of Robot and Human Intelligence*." He was followed, several years later, by Australian philosopher David Chalmers — never one to miss a bandwagon to jump onto — who presents an identical line of reasoning in a 1995 paper with the quirky title "*Absent Qualia, Fading Qualia, Dancing Qualia*." Not only is this argument fallacious, but it is riddled with logical holes and basic misconceptions — a veritable Emmental cheese of a thought experiment[2]. So let's take a closer look.

The Ship of Theseus

The *substitution argument* has roots in a venerable philosophical paradox, which was first recorded by Plutarch in his classic "*Parallel Lives*" in the 2nd century BCE. It's called the *Ship of Theseus*, a riddle that concerns the identity of an object over time — in this case the ship that the mythical King Theseus used to rescue the children of Athens after he had slain the Minotaur on Crete. Upon his return, the Athenians decided to keep the ship. Every year, they would sail it to Delos to worship Apollo and to commemorate Theseus' deed. As time went by, its wooden planks were slowly rotting away and had to be replaced — one by one — until, at some point, there was no original plank left.

This caused the Athenian philosophers to question whether the ship was really still *the* Ship of Theseus. And if it was not, at what time had it ceased to be the original? Much later, in the 17th century, Thomas Hobbes added a further complication to the riddle: with some effort, you could restore the old rotting planks and build a "new" ship from them, identical in design to the original one. Which ship, then, should be considered the *real* Ship of Theseus? These questions seem impossible to answer.

[2] It is probably no coincidence that David Chalmers is the champion of many other atrocious thought experiments — most famously, the abominable philosophical zombie, an argument so grotesque that it fully deserves the label "not even wrong." Every time this man publishes an argument, it seems, we understand the mind (and reality in general) a bit *less*.



Make no mistake: the argument behind Theseus' Paradox is logically sound. It points to a valid philosophical problem and there are several proposed resolutions that address it. But none of the details really matter here.

The only thing that *does* matter is that the ship has a *precisely defined function* — it is used, once a year, to sail the Aegean Sea in commemoration of Theseus' deed. Importantly, it performs this function as a classical *mechanism*. The ship is a machine: its parts (the wooden planks) are finite in number, the shape of each one of them well-defined. Planks are cleanly separable and arranged in a particular way with regard to each other. This gives the ship its characteristic shape and capacities. Therefore, we can argue that new planks function *exactly* the same way the old ones did: they are equivalent in composition, form, and arrangement. Nothing else matters. As long as the ship remains structurally the same, it serves the same function. Honestly: who cares whether it still *is* the same ship — metaphysically

speaking — or not? Only philosophers do! Ships of Theseus are *fungible*, which means that one will work just as well as any other for the purposes we want to put them to.

Neural Substitution

The *substitution* or *replacement argument* is nothing but a modern version of the Ship of Theseus applied to a human context. It imagines what would happen if we replace each *neuron* in your brain, one by one, with a mechanical or electronic device that performs exactly the same function as the neuron it replaces. At some point, we'll have substituted *all* your neurons with such devices. You are now a classic type of *cyborg*: a human body with a machine brain. Moravec goes one step further: what if we replace every *cell* in your body with equivalent artificial devices? You'd end up a *robot* replica: a machine with the same functional capacities as the human body it replaces. Surely, this machine would be alive?

But is that still *you* in there? Is it still *your* mind after you have become a cyborg? Is it still *your* body now that you are a robot? Well, in analogy to Theseus' Ship, if function is maintained continuously during each swap, your identity — and thus your life and mind — should also persist. And if this is indeed so, then it follows that your consciousness must depend on the *function* of components only, not on the biological *material substrate* that we have been swapping out.

If the story of the Ship of Theseus makes sense (and it *does*, as we have seen), shouldn't the substitution argument hold up as well? After all, both appear to be analogous. This is what Moravec and Chalmers claim. And if the substitution argument is valid, then it provides us with a solid logical grounding for computationalism: mind and life are nothing but abstract patterns, which can be imposed on any material substrate if you only get the functional encoding right. Uploading, here we come!

Superficially, it all seems to pan out. There is no obvious point during the swapping procedure at which we could point and say: "this is where you cease to be you and become an unconscious machine without a self." Imagine replacing a missing limb with a prosthetic. Nobody in their right mind would claim that the "cyborg" with an artificial leg is not the person they used to be before the replacement. Clearly, there *is* something strange about the whole situation. But don't be fooled: it is not as straightforward as Theseus' Ship. There are several hidden issues that sneak up on us when we transfer the argument from ship to human — or any other kind of living being, for that matter.

The first and foremost issue concerns how we determine the *function* of a component. We should not forget that in a natural (rather than an engineered) system functions are not pregiven, nor are they designed into the system from outside. Instead, we need to second-guess them: researchers *assign* function through the deliberate act of *analytical abstraction*, a central pillar of our scientific method. But *how* our analysis decomposes a complex natural system into parts depends on many things — not least what we can actually measure, and what specific phenomena or behavior we would like to explain.

Our task is further complicated by the fact that functions in *living* systems are rarely fixed, precisely localizable, or intrinsic to a specific part. Instead, *organismic functions* hinge on history and the dynamic relations between components and their systemic context as a whole. The same neuron behaves very differently at different times, and its behavior depends on where it is located in the organism and which other neurons and non-neuronal cells it interacts with (and has interacted with in the past). A cell is also an *autonomous agent*: it not only passively responds but actively *anticipates*.

This poses a challenge: for the substitution argument to work, we must make sure that the functions of the parts are defined such that original and replacement parts are equivalent in *every* relevant way. But it is really hard, if not impossible, to exactly pin down all the possible functional behaviors of a living cell.

Compare this to Theseus' Ship — an engineered artifact — where assigning function is easy. Even if we didn't know the shipbuilder's original intentions, the problem is precisely delimited: planks are static objects of a specified material and shape; they are manufactured independently of the rest of the ship, and only later inserted into its overall design according to a stable and well-defined blueprint. This determines their contribution to the ship's function unambiguously and completely. No problem here.

In contrast, cells in a living organism grow, divide, and together form tissues without an external builder or blueprint. Each is an active, dynamic, and motile *self-manufacturing process* with an agenda of its own. In any multicellular tissue, cellular behavior must be *coordinated* and *constrained* (more so than induced) to make cells cooperate. This happens in diverse and ever-changing ways. Cells sustain themselves, grow and reproduce, while bathed in a complex soup of nutrients, hormones, growth factors, electric and chemical signals. How they generate, relay, and interpret these signals depends not only on their environment, but also on internal state. And as they collectively co-construct internal and external states, each cell's repertoire of possible behaviors keeps changing — their capabilities never well-defined but continuously and spontaneously emerging from context, evolving over time in a radically open-ended manner.

This brings us to *the crux of the matter*: a question-begging assumption at the heart of the substitution argument. In contrast to what I just said about cells, it presumes that the functional contribution of a neuron to the mind is fixed, rule-based, and clearly defined. More precisely, it presupposes that neurons communicate with each other exclusively via the propagation and exchange of standardized *action potentials* (*nerve impulses* or *spikes*) through a limited number of anatomical structures such as axons, dendrites, and synapses. If this is indeed their entire contribution to the functioning of the mind, then we get an equivalence of *natural* and *artificial neural networks*, and the substitution argument applies.

However, this begs the question: it does not show, but simply *assumes* that these abstract patterns of spike trains *can* be abstracted away from the underlying substrate, and that they are the only abstract pattern that matters for our understanding of the mind and our identity.

Let's be frank: there is no evidence to support these oversimplified assumptions. In fact, there is plenty of evidence that the situation is much more complicated than this. It bears repeating that neuronal cells are complex living beings that grow and maintain themselves. They are not only in touch with other

neurons, but also non-neuronal cells (such as microglia, for example). They are bathed in a rich medium of chemicals and other signals that influence their behavior, growth, and viability. They can actively modulate the transduction and idiosyncratically interpret the content of these signals, and they can alter their own base rates of signaling — all of this in a radically context- and history-dependent manner.

To cut a long story short: it is simply not possible to summarize — in advance, and as a neat and finite list — all the possible functions of a cell such as a neuron. Cells are not planks. As theoretical biologist Stuart Kauffman and his colleagues remind us: *no laws entail the behavior and evolution of living beings.*

We'll come back to this in the next section, to show that it is a *fundamental* limitation. But even just practically considered, you'll have to agree that there is no way of knowing whether our replacement device really captures *all* potentially relevant functions of a neuron. We're just guessing here. And, obviously, this problem becomes even more acute when replacing all the cells of your body. We cannot assign a specific function to each in any unambiguous manner. Cells … these little buggers keep eluding us, always poised to change and behave in ways that we did not expect them to.

And, if this were not bad enough, we humans are also notoriously prone to miss *unconceived alternatives* to our preconceived notions. Our imagination is severely stymied in this way. Even worse: we sometimes even fail to see what is right in front of our eyes.

All of this means the substitution argument fails right away, at its outset. It succumbs to a breakdown of imagination *and* analysis: in a living system, we can *never* be sure that the functional capacities we have assigned to a cell are truly equivalent before and after any of the swaps. And that is that.

But we're in for even more trouble than that: another gaping hole in the substitution argument is the idea that function is maintained continuously during swaps. It should be obvious that this is an *idealization* at best. In practice, it will take a non-zero amount of time to execute a swap. And while it may be true that we are made of an astronomical number of cells — around 86 billion neurons in the adult brain, and more than 30 trillion cells in an average human body, according to the latest counts — a cell is and remains a discrete unit. Therefore, the assumption that function remains *continuous* while we swap out *discrete* components over time is just that: an unproven assumption.

No two cells in your body are the same. And, as we have seen, context radically matters for their behavior and function. Therefore, we cannot logically rule out that life, identity, and mind may very well be lost at a discrete and specific point during the swapping procedure. We'd have to do the experiment to find out. But as long as we haven't, the substitution argument is *not* logically sound.

And it must be said: the burden of proof really rests on you, if you disagree with my criticism. Go on. Show me that I'm wrong. Do the swaps! I'll be patiently waiting. It'll be a while before we can attempt anything like it. But, as long as you haven't provided the empirical evidence, I don't buy the substitution argument, or computationalism more generally, as a suitable framework to understand life and mind.

The cyborg — and the living robot that is supposed to be you — both remain a *myth*: not rational science supported by empirical evidence, but wishful figments of computationalist imagination.

Machine Architecture

In sum: what invalidates the substitution argument is that we can never tell whether we are actually replacing all the relevant functions of a biological component (an open-ended, self-manufacturing process) with those we have built into the supposedly equivalent machine part (a pre-manufactured thing). Neither can we justifiably claim that functional continuity is maintained during each swap.

None of this is a problem for Theseus' Ship, where planks get swapped for functionally equivalent planks while we're *not* sailing. In fact, we'd better be in dry dock when this happens, applying a new layer of paint while we're at it. And once we set sail again, the ship functions just like it did before.

In stark contrast, components *need* to be *hot-swapped* in a living organism[3]. We cannot simply shut down system functions, do the component swaps, and then boot everything back up. There is no dry dock. Any interruption would be lethal — and death is the ultimate *discontinuity* of function.

Thus, we can neither be sure that function continues, nor that biological and machine functions are equivalent before and after a swap. In other words, we are comparing apples with oranges when we draw an analogy between the Ship of Theseus and the substitution argument. In the latter, we are not swapping out like for like: machine function does *not* equal organismic function. On the contrary, they are not even remotely similar. Moravec and Chalmers walked into a logical trap without noticing when they came up with this comparison. Philosophers call this kind of mistake a *category error*. It's pretty elementary: cells are *not* planks — and this distinction matters.

The difference, by the way, does *not* lie in the *composition* of machine parts and biological components. Both are made of the same kind of chemical elements and follow the same laws of physics. Instead, the fundamental difference between living and non-living matter lies in the way the former is *organized* — how living component processes, each exhibiting perfectly respectable physico-chemical dynamics, *relate*

[3] So do the planks in Neurath's Boat, also famous in philosophy — but arguably a less realistic nautical metaphor than Theseus' Ship.

and *interact* with one another. These relations are contingent: they may be *constrained* but are not *determined* by the underlying laws of physics and chemistry, because they radically depend on history and context, and do not necessarily emerge and evolve in any law-like manner. And it is this "lawless" contingency that explains why functions in machines and organisms are not the same — not only in practice, but also in principle. Machines are well-behaved, law-like; but organisms are *not* like that.

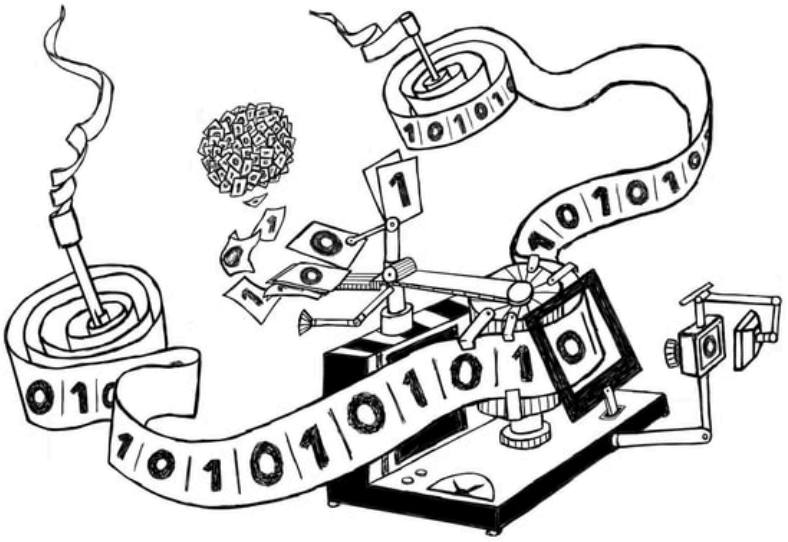
Now, you may ask: how is this even possible? Doesn't physics *prescribe* a mechanistic worldview? Isn't scientific explanation *necessarily* mechanistic? And aren't we ascribing *mystical properties* to living systems?

No. As a matter of fact, it does not, it is not, and we do not! To better understand why this is, we must compare the *architecture* that makes it easy to determine machine function with the biological *organization* that renders organismic function (and behavior) so difficult to pin down with formalistic reasoning.

Let's do the easy part first: *machine architecture*. Instead of Theseus' Ship, I'll consider a more interesting machine — the modern *computer*. Apart from a few exotic and experimental exceptions (*e.g.*, *neuromorphic computing devices*), computers are built on von Neumann's architecture — the familiar stored-program computer, basically. Your laptop is one, so is your smartphone, and the largest supercomputers in the world share this architecture too. In fact, pretty much any device that contains electronics these days is some kind of variant of von Neumann. This predominant design, in turn, is a real-world approximation of an abstract model of computation called a *Turing Machine*.

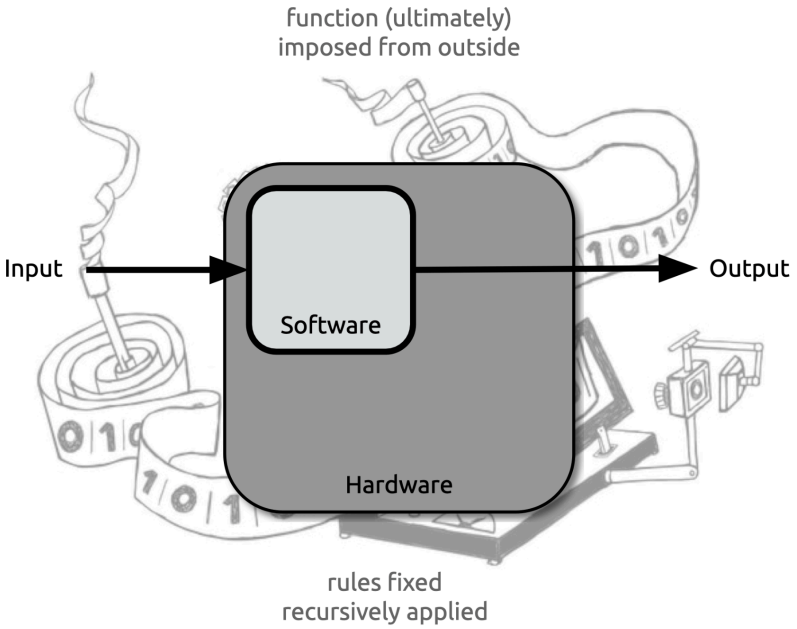
Mind you, Turing's machine is not a real machine. This abstract model cannot be realized as a physical system, because it requires an *infinite tape*, which is subdivided into cells that contain discrete (digital) chunks of information that serve as input to the machine. The input is processed, in sequence, by a *head* that reads the information and then transforms it according to a finite set of instructions (called a *transition table*), which resides within the internal *memory* of the machine. The transition table also determines whether the head then moves the tape to the left or to the right, or whether it stays exactly where it is. This results in a change of information content of the tape — the output of the machine.

We can further assume that the transition table itself can be read into memory from the tape, so that the stored information can be interpreted as either data or code. This gets us a *Universal Turing Machine* (UTM), able to implement *any* possible rule-based sequence of processing steps (*any* algorithm) you could possibly imagine. Turing (building on work by Kleene and Church) stated this mathematically in the 1930s, thereby laying the groundwork for what we now call the *theory of computation*.



For the sake of argument, let us then define a *machine* in general as "any physical system whose functions can be simulated perfectly by a UTM." According to this definition, von-Neumann computers are machines that come really close to being universal simulators, able to mimic the behavior of almost any other (more specialized) machine. This is why computational methods and models are so good at reproducing and predicting mechanistic behavior. They are powerful scientific tools indeed!

We can summarize and visualize this kind of *computational architecture* as follows:



A computer receives input data (from the left) and processes it into output (to the right). What makes this computational architecture so powerful is, at its very heart, the strict *separation of hardware and software* (as shown by the nested boxes). After all, the whole point of the UTM is to be a *universal* model of computation, as flexible as possible when it comes to executing the rule-based instructions of an algorithm. For this to work, the symbols constituting the software must be completely independent of how the machine is built: we may inscribe *any* possible sequence of information on the tape.

It may not be immediately obvious, but the drawback entailed by this design is that function *must* be imposed from outside — at the hardware level through the design of the physical computing machinery which gets pre-assembled in a factory, and at the level of software through the rules that determine the programming (pre-scheduled execution) of the instructions. Neither hardware nor software construct or run themselves entirely. Not even *rewrite systems* — computational formalisms that treat their own instructions as data — transcend this limitation. They too, must be based on fixed rules imposed from outside the machine, even if they no longer determine the sequence of instructions directly, but only how they get transformed from some initial formal expression into one that is eventually executed.

Whether specified directly or through rewriting, algorithmic rules are implemented in Turing's framework through *recursion*: the self-referential application of symbols to symbols — all ultimately read from the tape. For this to work, the hardware (tape, memory, and reading head) must be designed in a way that conforms to the architecture prescribed by the UTM model. Importantly, it cannot itself be altered by software: recursion may be self-referential within the *symbolic* realm, but it must not affect the *physical* hardware, as this would violate the fundamental hardware/software separation, which is what yields the universality that gives the system its power in the first place.

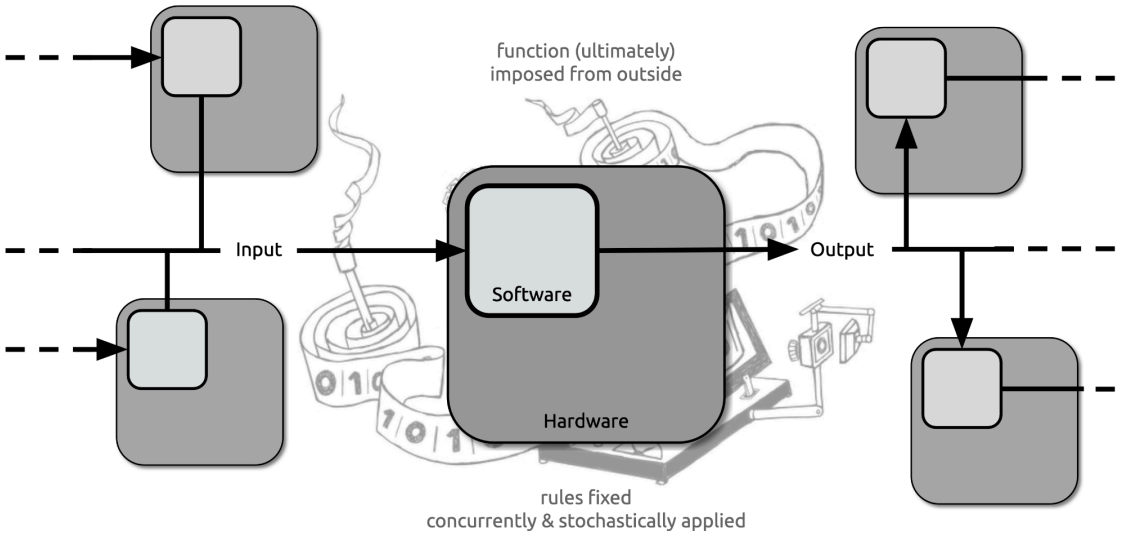
One common objection is that you simply need to put the factory which produces the hardware under the control of the software. Et voilà: a computer system that builds itself! But not so fast: how do you produce the hardware (factory machines and their parts) that produces the hardware? You quickly descend into an *infinite regress* that way, which is only interrupted, once again, by imposing functions from outside. The buck, in such a system, has to stop somewhere if it is to be physically realizable.

All of this means that a computational or algorithmic system may be *complicated* (think *large language models* with their gazillions of connection weights), but it is not *complex* in the same sense that living systems are: no von Neumann computer — or any algorithm running on it, no matter how intricate — will ever achieve the ability to self-manufacture within the constraints of its own design. That would

require what theoretical biologist Howard Pattee calls *semiotic closure*: the existential *integration* of "hardware" (physics) and "software" (symbols) we see in living organisms, which breaks the fundamental design principles of the computational system. We'll come back to this limitation shortly.

In the meantime, you may object that we are relying on a narrow and outdated model of computation: since the internet was invented, computers have become networked, and we run software on hardware devices such as robots that move around and interact with the outside world. This changes the situation fundamentally: algorithms become deeply embedded in an uncertain and unpredictable world! And you may have a point.

To account for this, we further relax our assumptions about what it means to be a computer. As a first step, we should let UTMs *interact*:



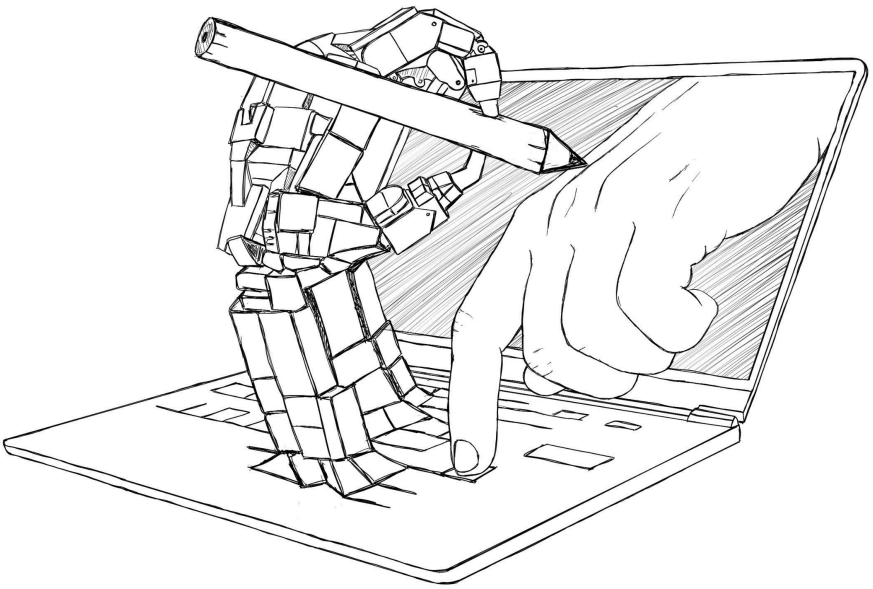
Here, the output of one UTM becomes the input to another one. This has important consequences. It means we no longer expect a UTM to calculate solutions to specific problems and then halt, as traditional computer science did, because each node in the network keeps on receiving new input (imagine the information on its tape being constantly rewritten). Furthermore, interaction can give us *concurrency*, as algorithmic instructions become processed simultaneously and in parallel. This, in turn, introduces an amount of indeterminacy, because parallel processes may be executed at variable rates, which makes input/output patterns unpredictable. Another kind of indeterminacy can be added through *stochastic* execution of instructions: at each processing step, multiple instructions may get

applied, each with a certain probability. All of this vastly expands the range of instruction sequences that can be run on a network like the one depicted above, compared to a single, isolated UTM.

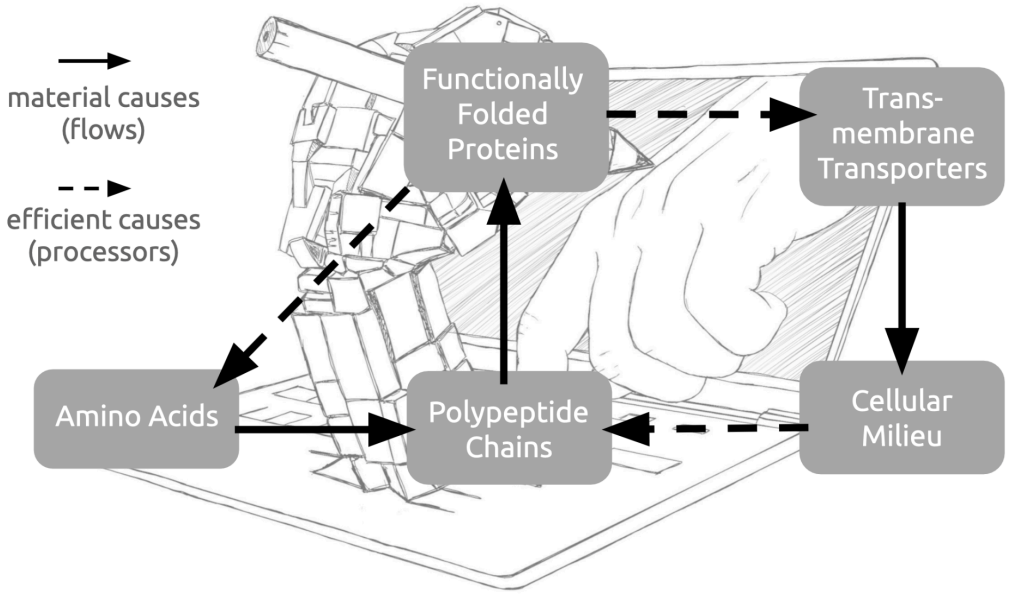
Still, some important design constraints remain in place in this extended computational framework. Most importantly: function must still be imposed externally, as each processing node is an independent UTM whose hardware cannot be altered by the software it is recursively running. And the rules that determine how program instructions are executed must also be codified, ultimately, from outside the network itself. The internet, as a real-world example, requires many technological standards and communication protocols for interaction to work successfully at both hardware and software levels.

In the end, this still precludes the kind of software/hardware *integration* (semiotic closure) you need for a true self-manufacturing system. And without it, we still get the infinite regress outlined for a single UTM above. The only way to change this is to design a system where code can directly and literally *become* hardware in some way, without losing universality. Last time I checked, such designs or technologies did not exist, and nobody was anywhere near inventing them. In biology, however, self-manufacturing organization has been around for billions of years, and we also have a plausible account of how it works, which is very useful for understanding why living beings (including their brains) are *not* computers — and not even *like* computers at all.

Living Organization



Here is where things may become a bit complicated and counterintuitive. Please bear with me. I'll do my best to explain everything using minimal jargon and technicalities. To begin, let's compare the following arrow diagram of a living cell — proposed by biochemist Jan-Hendrik Hofmeyr, based on earlier work by theoretical biologist Robert Rosen — with those of computer architecture above:



What you immediately notice is that arrows go around in circles here, when in the diagrams of computational architecture they consistently point straight from left to right. This is not an accident, but we need to understand the details of the diagram better before we can fully grasp what it means.

To be sure, this is only part of a much more complicated model. But since both have the same distilled shape when we leave out all unnecessary detail, it will suffice for my argument and be easier to parse than the full scheme. If you want to learn more about the latter, go read Hofmeyr's original papers.

For starters, imagine an input arrow pointing from a source of nutrients outside the organism to the node that says "*amino acids*." Together with the arrow that points further to the right, representing the concatenation of amino-acid building blocks into "*polypeptide chains*," this represents the part of metabolism responsible for *macromolecular synthesis*. The arrow is solid because it represents a *material transformation* (a *flow* — or, more accurately, a *flux*). This transformation is mediated through catalysis by functionally folded *proteins* called *enzymes*. This is marked by a dashed arrow, indicating an *efficient cause* (or *processor*), which enables an underlying material flow but does not itself participate in it.

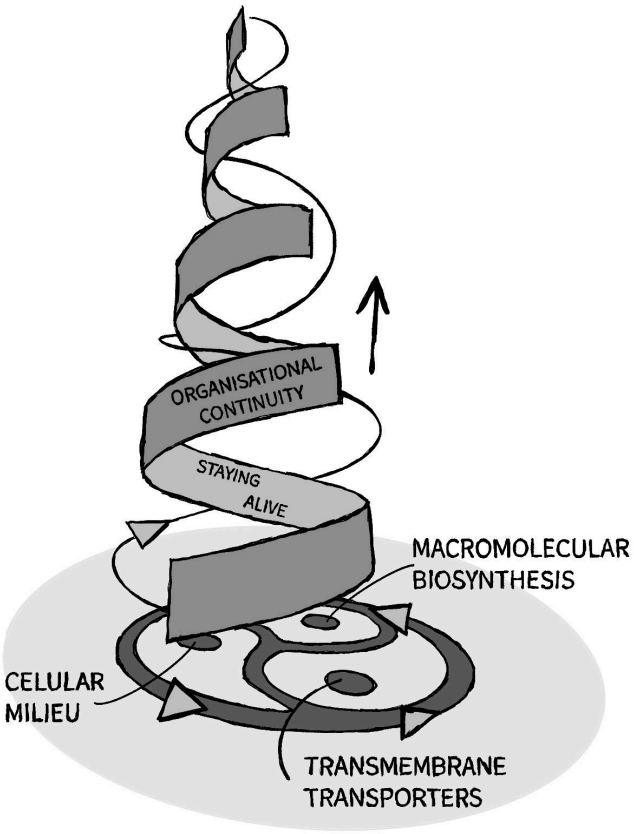
To become functional proteins, polypeptide chains must fold into the appropriate conformation. This happens spontaneously in most cases but requires very specific conditions to be present within a cell. These constitute the *cellular milieu*. pH, ionic and metabolite concentrations (including co-factors), and the spatio-temporal distribution of other proteins, macromolecular structures, and organelles must be precisely regulated for proteins to fold into their functional conformations. Here, the milieu plays the role of the processor: it enables folding but is not affected by the process itself.

Finally, maintaining the milieu requires a semipermeable *cell membrane* (whose lipid building blocks are another product of metabolism). Embedded in this membrane are *transmembrane transporters* (also functionally folded proteins), which act as processors for maintaining the cellular milieu, by selectively letting, or actively pumping, ions and other chemicals in and out of the cell.

We've literally come full circle: you will notice that all three essential processes involved in cellular organization — macromolecular metabolism, protein folding, and maintenance of the milieu — are processed (efficiently caused) by components *within* the organization of the system. In contrast to machine function, there is nothing imposed from outside when it comes to cellular and organismic functions! But note: this does *not* mean that a cell is *thermodynamically closed*. Quite the contrary, the membrane must allow the entry of nutrients and ions, after all. Otherwise, macromolecular synthesis and regulation of the milieu could not occur. Yet, while being open thermodynamically, it also exhibits *organizational closure*: all the functions (processors) required to keep cellular processes going are provided from *within* the cell.

This is how we avoid the infinite regress we kept bumping into when we were trying to construct self-manufacturing computers before. Basically, a living cell is organized like an *origami* folded in upon itself. And this crucially involves at least one factor — the cellular milieu — that is a feature of the cell as a whole, not of any localized, individual subcomponent.

This kind of hierarchical circularity, by the way, is not at all the same as *cybernetic feedback*, which only regulates flow among *preexisting* processes that occur at the same level of organization. Nor is it mere *recursion*, which is purely *symbolic* self-referentiality. Instead, it means that *every* process essential to the continued existence of the cell is physically *co-constructed* by all the other processes which constitute the cell. If this does not make your head spin, then I don't know what does. Each cellular process depends on all the other ones *for its very existence*. This is what *self-manufacture* actually means: not only the physical fabrication of all required parts, but also their constitutive and continuous self-assembly. This can be visualized as the following upward spiral:



Because everything is constantly converted into everything else in such a system (figuratively, but still accurately speaking) there can be no strict separation or segregation of processes. Take, for example, the *genetic encoding* of the amino-acid sequence that constitutes the primary structure of a protein. The code that we could call the "software" of the cell is stored in genomic DNA. This genetic sequence first gets transcribed into mRNA, and then translated into a polypeptide chain before folding into a functional protein component (which we can consider the basic "hardware" of the living cell). Finally, a subset of these proteins (transcription factors) then bind to DNA to regulate gene expression.

Again, we get a closed, continuous cycle — "software" to "hardware" to "software" — and no separation of the symbolic (coded sequence) and the physical (functional folding). Pattee's semiotic closure is achieved! Hardware and software are combined into one inseparable constructive process. And, as theoretical biologist C. H. Waddington pointed out, we can no longer tell if the genome encodes the cell or if the cell interprets its genome. They are complementary perspectives on the same hierarchical and circular *self-referential organization*.

Now, and this is crucial to understand: the *rules* governing the dynamics of such a system *can no longer be predefined*. This is what Kauffman means when he says "no laws entail organismic behavior." Instead, the rules are (re)generated, continuously and on the spot, from *within* the organization of the system.

Mathematically, we can explain this via the closed loop to the right in our arrow diagram of the cell (see above). Its abstract form is what mathematicians call *collectively impredicative*, as it defines the processes of protein folding and regulation of the milieu in terms of each other. A perfectly harmless example of an impredicative definition is "the biggest fish in the pond." Here, the fish to be defined is included as part of the sample ("all the fish in the pond") that serves to define it in the first place. Clearly, this is circular reasoning, but evidently not vicious in this case. Similarly, collective impredicativity among cellular processes is also *not* vicious. It only indicates collective co-construction, which means these processes can only exist — and only proceed up the spiral — if they are *all* active and present at the same time.

As you can imagine, this is a real headache for explaining the origin of life: instead of the *halting problem* of computer science we get the mother of a *starting problem* in biology. All of a living cell's processes need to get going *together* to kick-start life. How astronomically improbable is that? Surely, we must look for more plausible, step-by-step pathways towards living organization. But the starting problem also introduces a new kind of indeterminacy into cells and organisms that exist right now: it makes it impossible for us to formulate a model able to predict all the possible future states of the system. In more technical terms (for those who care about such things): the circularity introduced by collective impredicativity makes it impossible to precisely define the system's *state space*. This is the mathematical reason why we cannot pin down biological behavior, and hence function, precisely in advance. It is not just a practical limitation, but impossible *in principle*, because of the impredicative nature of the system.

In fact, Robert Rosen proved mathematically — using the *theory of categories* — that the category of living systems (without well-defined dynamic rules or state spaces) is fundamentally distinct from the category of machines (whose models are based on fixed rules and state spaces). This is called *Rosen's conjecture*. It states that organisms are not *computable* (or *simulable*, as Rosen calls it), and has led to much controversy over the years, because the way it is formulated tends to confuse people. The problem is as follows: Rosen does *not* say that it is impossible to come up with simulations of a cell's dynamic behavior; all he says is that these simulations won't capture *all* the possible future dynamics. It is an *incompleteness argument* for the life sciences, analogous to Gödel's *incompleteness theorems* for formal systems. Life cannot be *completely* formalized: it will always surprise us — its behavior and evolution truly open-ended indeed.

Biological Naturalism

Let's recap: Moravec and Chalmers' substitution argument is frequently used to prop up the ideology of *computational functionalism* (or *computationalism* for short). However, the substitution argument fails. It does so because it is not clear that function is maintained while hot-swapping discrete components such as neurons and other cells, and because we cannot replace organismic with machine function, as they are not at all the same thing. In contrast to a machine part, we cannot pin down the function of a living cell in any definite way. This is where the analogy between the Ship of Theseus and the substitution argument breaks down. We are comparing apples and oranges.

Beneath all of this lies the contrast between *machine architecture* and the *dynamic self-manufacturing organization* of cells and organisms with *organizational closure*. This kind of closure is based on processors and material flows that *collectively co-construct* each other in a self-referential and collectively impredicative way. Mathematical analysis of this kind of closure reveals why we cannot capture the full range of possible organismic behaviors, and thus functions, in any well-defined way. The starting problem (all processes having to get going at the same time) makes the formal construction of a definite state space impossible. This implies that the behavioral potential of a living system is fundamentally open-ended, and thus not completely simulable on our current computational architectures.

I need to stress again that there is no magic involved here, whatsoever — no vitalistic life-giving force or essence, or anything like it: the organizational account of life presented here is perfectly compatible with all the known laws of physics — especially thermodynamics and chemical kinetics. In addition, it is a perfectly *scientific* and *naturalistic explanation*, formulated entirely in terms of cause-and-effect between material processes. But, crucially, it is *not* a *mechanistic explanation*! The notion of "mechanism" implies a chain of causation from some starting point to a specific outcome, and that the parts of a mechanistic system interact with each other according to fixed and well-defined rules. Neither of these assumptions holds in a living system, as I have outlined in the previous section.

This has important consequences. First of all, it means that *organisms are not machines*. Of course, organisms contain mechanisms as subcomponents (the metabolic part of the diagram above being one particularly obvious example). And yet, an organism belongs to an entirely separate category of natural systems when we consider it as an organized whole. This is what Rosen demonstrated mathematically.

Second, this further implies that there are organismic functions, which *cannot* be implemented in machine architecture. Organisms and their subsystems — your brain, for example — can do things a

machine or algorithm will *never* be able to do, because of the constraints imposed on them by their design.

Such uniquely organismic functions matter: it is *not* an open empirical question whether our current computational architectures can support *selfhood* or *sentience* as we define these concepts in the living world. *They cannot, by design.* All this talk about AGI, ASI, or sentient algorithms is moot speculation without any deeper meaning or scientific merit. My argument shows that computationalism, with its claim that mental or living patterns can be abstracted from their material substrate, is neither plausible given our current evidence, nor is it based on logically consistent thinking. It stands on very shaky ground, and there is no reason to take renewed claims of dualism about mind and matter seriously.

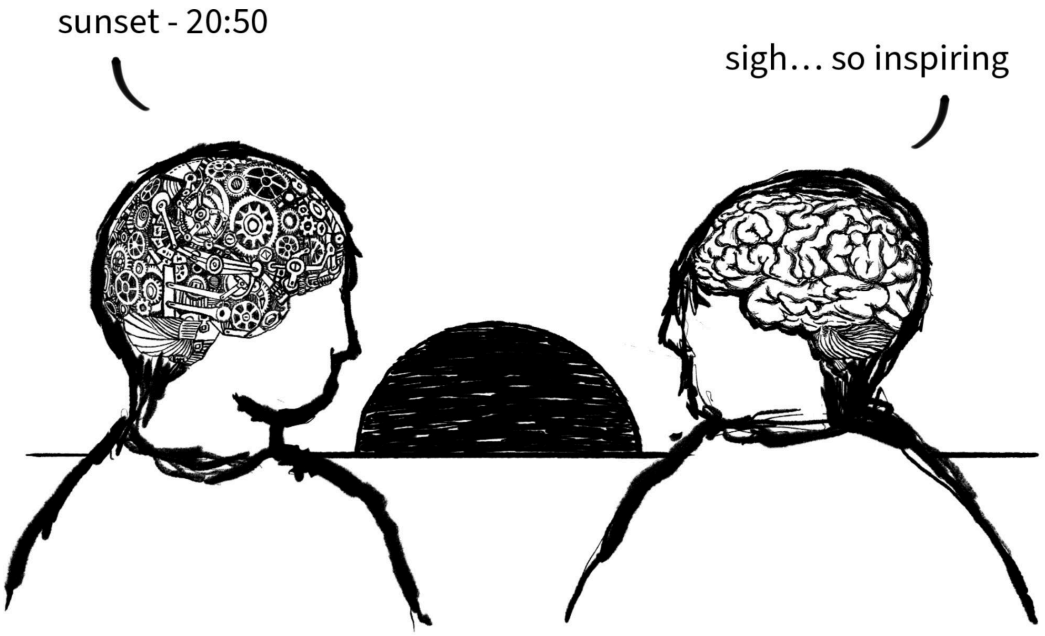
As an alternative, our discussion suggests that the material substrate *is* indeed important. Specifically, it needs to be the kind of substrate that allows self-manufacturing organizational closure to emerge and sustain itself. This view is called *biological naturalism*. It is gaining some traction lately — and I've just given you a naturalistic explanation that makes it a lot more plausible still. Right now, it is by far our best explanation for the observation that organisms and machines have different behavioral capacities. Additional empirical support is needed, but at the very least, biological naturalism is a possibility that we should take seriously — more seriously, anyway, than its far-fetched computationalist antagonists.

An important caveat remains: none of this means that *only* our specific meat-based configuration is suitable for supporting life. There may well be many exotic selves and minds out there that are based on utterly alien chemistries. Philosophers call this *multiple realizability*. It is different, however, from the *separability* of living and mental phenomena from substrate that is required for computationalism: the metal- and silicate-based technology we have today, for instance, seems too rigid for supporting living organization; it is very probably *not* a suitable substrate. And, in any case, as long as we maintain the strict software-hardware separation in the design of our machines — including our most powerful computers and algorithms today — we are *not* going to get living or sentient machines, not even if we embed our algorithms in motile robotic frames or network them across the globe. And, of course, that copy of your consciousness in the computational cloud? Well, that's not you! It's just a simulation.

To conclude: it is time to say bye-bye to brain uploading, to sentient machines, to synthetic minds, and to Scotty beaming you up any time soon. Computationalism — the separability of mind and life from material substrate — is absolutely required for all of these. Sadly, it is not a plausible, empirically supported, or logically consistent ideology. Rather, it is wishful thinking by people who view themselves and the entire world they live in as a machine to be manipulated and optimized. It is an irrational belief

that is deeply rooted in our desire for power and control, as well as our fear of unpredictability and surprise. It is utterly human this way but also profoundly confused, and highly dangerous, if you ask me, because it leads us to the hubristic *accelerationism* and *transhumanism* that are frightfully popular today.

A central pillar of this ideological movement — of this *cult of technology*, really — is the yearning for longevity. And a cult it is: we are not accelerating or transforming ourselves towards any real, achievable or even desirable goals. The immortal mind in the cloud, backed up and on its way to the stars, is cheap science fiction at best. The technologies and treatments we pretend to be creating are mirages, based on bad science and flawed thinking. And the unintended consequences we are generating in the process are severe, and sometimes existentially threatening to our human future on this planet.

My ultimate goal here is to make you stop for a second and think — especially if you are an accelerationist. What is life really about? What makes it so special? What makes you unique? What makes you (and no-one else)? What drives you? What enables you to feel, desire, judge, and prioritize? What gives you dignity and creativity? Well, it is your living organization, your ability to literally construct yourself the way you decide to. At least to some degree, you are autonomous — in charge of your life! You couldn't do this if you were a mere machine. The downside of this power is that you have to constantly invest work to stay alive. And when this work is done, it is your time to make room for other living beings, with their own way of constructing themselves, to take your place in this world of limited resources. To partake in this larger process, to me that's what it truly means to be alive.

Can you feel it?