Contents

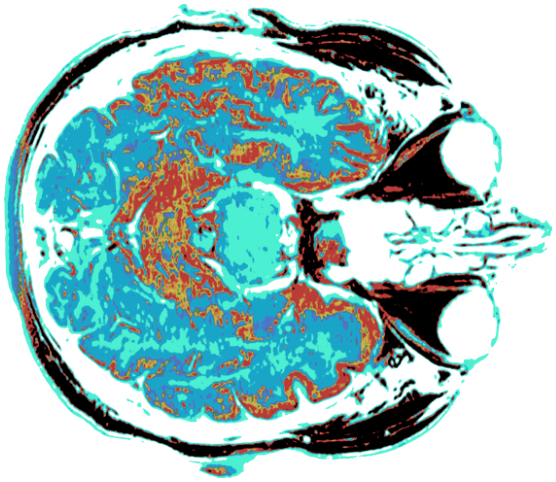
What If We Had Bigger Brains? Imagining Minds beyond Ours

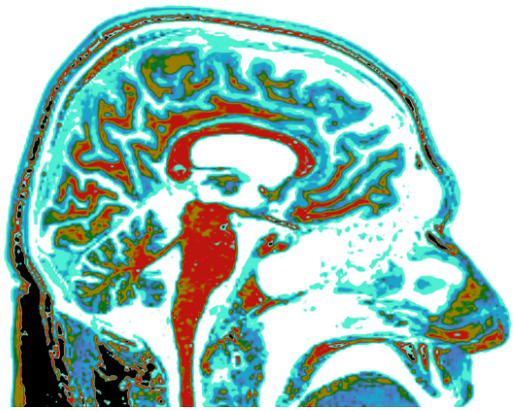
May 21, 2025

Cats Don't Talk

We humans have perhaps 100 billion neurons in our brains. But what if we had many more? Or what if the AIs we built effectively had many more? What kinds of things might then become possible? At 100 billion neurons, we know, for example, that compositional language of the kind we humans use is possible. At the 100 million or so neurons of a cat, it doesn't seem to be. But what would become possible with 100 trillion neurons? And is it even something we could imagine understanding?

My purpose here is to start exploring such questions, informed by what we've seen in recent years in neural nets and LLMs, as well as by what we now know about the fundamental nature of computation, and about neuroscience and the operation of actual brains (like the one that's writing this, imaged here):





One suggestive point is that as artificial neural nets have gotten bigger, they seem to have successively passed a sequence of thresholds in capability:

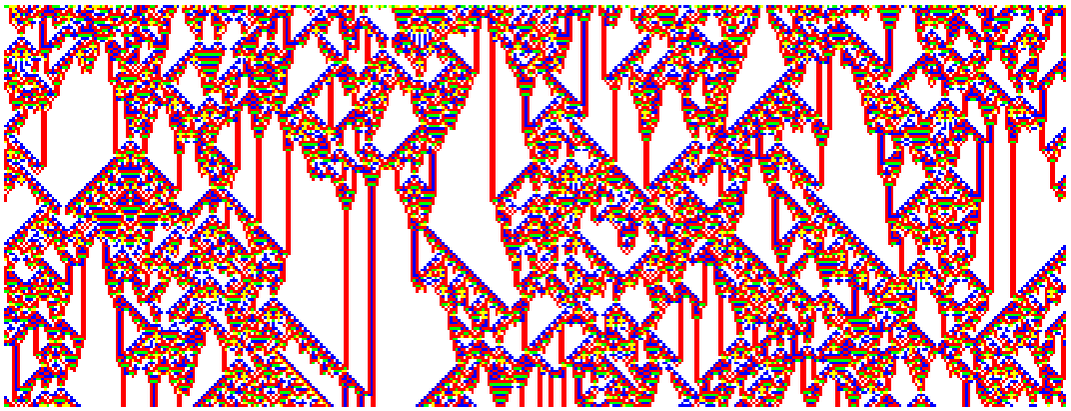
~10^5 connections	recognize images of 10 digits
~10^6 connections	recognize images of all letters
~10^7 connections	recognize images of 5000 picturable nouns
~10^8 connections	transcribe speech from audio
~10^9 connections	generate photorealistic images from text
~10^{11} connections	generate fluent natural language text

So what's next? No doubt there'll be things like humanoid robotic control that have close analogs in what we humans already do. But what if we go far beyond the ~10^{14} connections that our human brains have? What qualitatively new kinds of capabilities might there then be?

If this was about "computation in general" then there wouldn't really be much to talk about. The Principle of Computational Equivalence implies that beyond some low threshold computational systems can generically produce behavior that corresponds to computation that's as sophisticated as it can ever be. And indeed that's the kind of thing we see both in lots of abstract settings, and in the natural world.

But the point here is that we're not dealing with "computation in general". We're dealing with the kinds of computations that brains fundamentally do. And the essence of these seems to have to do with taking in large amounts of sensory data and then coming up with what amount to decisions about what to do next.

It's not obvious that there'd be any reasonable way to do this. The world at large is full of computational irreducibility—where the only general way to work out what will happen in a system is just to run the underlying rules for that system step by step and see what comes out:



And, yes, there are plenty of questions and issues for which there's essentially no choice but to do this irreducible computation—just as there are plenty of cases where LLMs need to call on our Wolfram Language computation system to get computations done. But brains, for the things most important to them, somehow seem to routinely manage to "jump ahead" without in effect simulating every detail. And what makes this possible is the fundamental fact that within any system that shows overall computational irreducibility there must inevitably be an infinite number of "pockets of computational reducibility", in effect associated with "simplifying features" of the behavior of the system.

It's these "pockets of reducibility" that brains exploit to be able to successfully "navigate" the world for their purposes in spite of its "background" of computational irreducibility. And in these terms things like the progress of science (and technology) can basically be thought of as the identification of progressively more pockets of computational reducibility. And we can then imagine that the capabilities of bigger brains could revolve around being able to "hold in mind" more of these pockets of computational reducibility.

We can think of brains as fundamentally serving to "compress" the complexity of the world, and extract from it just certain features—associated with pockets of reducibility—that we care about. And for us a key manifestation of this is the idea of concepts, and of language that uses them. At the level of raw sensory input we

might see many detailed images of some category of thing—but language lets us describe them all just in terms of one particular symbolic concept (say "rock").

In a rough first approximation, we can imagine that there's a direct correspondence between concepts and words in our language. And it's then notable that human languages all tend to have perhaps 30,000 common words (or word-like constructs). So is that scale the result of the size of our brains? And could bigger brains perhaps deal with many more words, say millions or more?

"What could all those words be about?" we might ask. After all, our everyday experience makes it seem like our current 30,000 words are quite sufficient to describe the world as it is. But in some sense this is circular: we've invented the words we have because they're what we need to describe the aspects of the world we care about, and want to talk about. There will always be more features of, say, the natural world that we could talk about. It's just that we haven't chosen to engage with them. (For example, we could perfectly well invent words for all the detailed patterns of clouds in the sky, but those patterns are not something we currently feel the need to talk in detail about.)

But given our current set of words or concepts, is there "closure" to it? Can we successfully operate in a "self-consistent slice of concept space" or will we always find ourselves needing new concepts? We might think of new concepts as being associated with intellectual progress that we choose to pursue or not. But insofar as the "operation of the world" is computationally irreducible it's basically inevitable that we'll eventually be confronted with things that cannot be described by our current concepts.

So why is it that the number of concepts (or words) isn't just always increasing? A fundamental reason is abstraction. Abstraction takes collections of potentially large numbers of specific things ("tiger", "lion", …) and allows them to be described "abstractly" in terms of a more general thing (say, "big cats"). And abstraction is useful if it's possible to make collective statements about those general things ("all big cats have…"), in effect providing a consistent "higher-level" way of thinking about things.

If we imagine concepts as being associated with particular pockets of reducibility, the phenomenon of abstraction is then a reflection of the existence of networks of

these pockets. And, yes, such networks can themselves show computational irreducibility, which can then have its own pockets of reducibility, etc.

So what about (artificial) neural nets? It's routine to "look inside" these, and for example see the possible patterns of activation at a given layer based on a range of possible ("real-world") inputs. We can then think of these patterns of activation as forming points in a "feature space". And typically we'll be able to see clusters of these points, which we can potentially identify as "emergent concepts" that we can view as having been "discovered" by the neural net (or rather, its training). Normally there won't be existing words in human languages that correspond to most of these concepts. They represent pockets of reducibility, but not ones that we've identified, and that are captured by our typical 30,000 or so words. And, yes, even in today's neural nets, there can easily be millions of "emergent concepts".

But will these be useful abstractions or concepts, or merely "incidental examples of compression" not connected to anything else? The construction of neural nets implies that a pattern of "emergent concepts" at one layer will necessarily feed into the next layer. But the question is really whether the concept can somehow be useful "independently"—not just at this particular place in the neural net.

And indeed the most obvious everyday use for words and concepts—and language in general—is for communication: for "transferring thoughts" from one mind to another. Within a brain (or a neural net) there are all kinds of complicated patterns of activity, different in each brain (or each neural net). But a fundamental role that concepts, words and language play is to define a way to "package up" certain features of that activity in a form that can be robustly transported between minds, somehow inducing "comparable thoughts" in all of them.

The transfer from one mind to another can never be precise: in going from the pattern of activity in one brain (or neural net) to the pattern of activity in another, there'll always be translation involved. But—at least up to a point—one can expect that the "more that's said" the more faithful a translation can be.

But what if there's a bigger brain, with more "emergent concepts" inside? Then to communicate about them at a certain level of precision we might need to use

more words—if not a fundamentally richer form of language. And, yes, while dogs seem to understand isolated words ("sit", "fetch", ...), we, with our larger brains, can deal with compositional language in which we can in effect construct an infinite range of meanings by combining words into phrases, sentences, etc.

At least as we currently imagine it, language defines a certain model of the world, based on some finite collection of primitives (words, concepts, etc.). The existence of computational irreducibility tells us that such a model can never be complete. Instead, the model has to "approximate things" based on the "network of pockets of reducibility" that the primitives in the language effectively define. And insofar as a bigger brain might in essence be able to make use of a larger network of pockets of reducibility, it can then potentially support a more precise model of the world.

And it could then be that if we look at such a brain and what it does, it will inevitably seem closer to the kind of "incomprehensible and irreducible computation" that's characteristic of so many abstract systems, and systems in nature. But it could also be that in being a "brain-like construct" it'd necessarily tap into computational reducibility in such a way that—with the formalism and abstraction we've built—we'd still meaningfully be able to talk about what it can do.

At the outset we might have thought any attempt for us to "understand minds beyond ours" would be like asking a cat to understand algebra. But somehow the universality of the concepts of computation that we now know—with their ability to address the deepest foundations of physics and other fields—makes it seem more plausible we might now be in a position to meaningfully discuss minds beyond ours. Or at least to discuss the rather more concrete question of what brains like ours, but bigger than ours, might be able to do.

How Brains Seem to Work

As we've mentioned, at least in a rough approximation, the role of brains is to turn large amounts of sensory input into small numbers of decisions about what to do. But how does this happen?

Human brains continually receive input from a few million "sensors", mostly associated with photoreceptors in our eyes and touch receptors in our skin. This input is processed by a total of about 100 billion neurons, each responding in a few milliseconds, and mostly organized into a handful of layers. There are altogether perhaps 100 trillion connections between neurons, many quite long range. At any given moment, a few percent of neurons (i.e. perhaps a billion) are firing. But in the end, all that activity seems to feed into particular structures in the lower part of the brain that in effect "take a majority vote" a few times a second to determine what to do next—in particular with the few hundred "actuators" our bodies have.

This basic picture seems to be more or less the same in all higher animals. The total number of neurons scales roughly with the number of "input sensors" (or, in a first approximation, the surface area of the animal—i.e. volume$^{2/3}$—which determines the number of touch sensors). The fraction of brain volume that consists of connections ("white matter") as opposed to main parts of neurons ("gray matter") increases as a power of the number of neurons. The largest brains—like ours—have a roughly nested pattern of folds that presumably reduce average connection lengths. Different parts of our brains have characteristic functions (e.g. motor control, handling input from our eyes, generation of language, etc.), although there seems to be enough universality that other parts can usually learn to take over if necessary. And in terms of overall performance, animals with smaller brains generally seem to react more quickly to stimuli.

So what was it that made brains originally arise in biological evolution? Perhaps it had to do with giving animals a way to decide where to go next as they moved around. (Plants, which don't move around, don't have brains.) And perhaps it's because animals can't "go in more than one direction at once" that brains seem to have the fundamental feature of generating a single stream of decisions. And, yes, this is probably why we have a single thread of "conscious experience", rather than a whole collection of experiences associated with the activities of all our neurons. And no doubt it's also what we leverage in the construction of language—and in communicating through a one-dimensional sequence of tokens.

It's notable how similar our description of brains is to the basic operation of large language models: an LLM processes input from its "context window" by feeding it

through large numbers of artificial neurons organized in layers—ultimately taking something like a majority vote to decide what token to generate next. There are differences, however, most notably that whereas brains routinely intersperse learning and thinking, current LLMs separate training from operation, in effect "learning first" and "thinking later".

But almost certainly the core capabilities of both brains and neural nets don't depend much on the details of their biological or architectural structure. It matters that there are many inputs and few outputs. It matters that there's irreducible computation inside. It matters that the systems are trained on the world as it is. And, finally, it matters how "big" they are, in effect relative to the "number of relevant features of the world".

In artificial neural nets, and presumably also in brains, memory is encoded in the strengths (or "weights") of connections between neurons. And at least in neural nets it seems that the number of tokens (of textual data) that can reasonably be "remembered" is a few times the number of weights. (With current methods, the number of computational operations of training needed to achieve this is roughly the product of the total number of weights and the total number of tokens.) If there are too few weights, what happens is that the "memory" gets fuzzy, with details of the fuzziness reflecting details of the structure of the network.

But what's crucial—for both neural nets and brains—is not so much to remember specifics of training data, but rather to just "do something reasonable" for a wide range of inputs, regardless of whether they're in the training data. Or, in other words, to generalize appropriately from training data.

But what is "appropriate generalization"? As a practical matter, it tends to be "generalization that aligns with what we humans would do". And it's then a remarkable fact that artificial neural nets with fairly simple architectures can successfully do generalizations in a way that's roughly aligned with human brains. So why does this work? Presumably it's because there are universal features of "brain-like systems" that are close enough between human brains and neural nets. And once again it's important to emphasize that what's happening in both cases seems distinctly weaker than "general computation".

A feature of "general computation" is that it can potentially involve unbounded amounts of time and storage space. But both brains and typical neural nets have just a fixed number of neurons. And although both brains and LLMs in effect have an "outer loop" that can "recycle" output to input, it's limited.

And at least when it comes to brains, a key feature associated with this is the limit on "working memory", i.e. memory that can readily be both read and written "in the course of a computation". Bigger and more developed brains typically seem to support larger amounts of working memory. Adult humans can remember perhaps 5 or 7 "chunks" of data in working memory; for young children, and other animals, it's less. Size of working memory (as we'll discuss later) seems to be important in things like language capabilities. And the fact that it's limited is no doubt one reason we can't generally "run code in our brains".

As we try to reflect on what our brains do, we're most aware of our stream of conscious thought. But that represents just a tiny fraction of all our neural activity. Most of the activity is much less like "thought" and much more like typical processes in nature, with lots of elements seemingly "doing their own thing". We might think of this as an "ocean of unconscious neural activity", from which a "thread of consensus thought" is derived. Usually—much like in an artificial neural net—it's difficult to find much regularity in that "unconscious activity". Though when one trains oneself enough to get to the point of being able to "do something without thinking about it", that presumably happens by organizing some part of that activity.

There's always a question of what kinds of things we can learn. We can't overcome computational irreducibility. But how broadly can we handle what's computationally reducible? Artificial neural nets show a certain genericity in their operation: although some specific architectures are more efficient than others, it doesn't seem to matter much whether the input they're fed is images or text or numbers, or whatever. And for our brains it's probably the same—though what we've normally experienced, and learned from, are the specific kinds of input that come from our eyes, ears, etc. And from these, we've ended up recognizing certain types of regularities—that we've then used to guide our actions, set up our environment, etc.

And, yes, this plugs into certain pockets of computational reducibility in the world. But there's always further one could go. And how that might work with brains bigger than ours is at the core of what we're trying to discuss here.

Language and Beyond

At some level we can view our brains as serving to take the complexity of the world and extract from it a compressed representation that our finite minds can handle. But what is the structure of that representation? A central aspect of it is that it ignores many details of the original input (like particular configurations of pixels). Or, in other words, it effectively equivalences many different inputs together.

But how then do we describe that equivalence class? Implementationally, say in a neural net, the equivalence class might correspond to an attractor to which many different initial conditions all evolve. In terms of the detailed pattern of activity in the neural net the attractor will typically be very hard to describe. But on a larger scale we can potentially just think of it as some kind of robust construct that represents a class of things—or what in terms of our process of thought we might describe as a "concept".

At the lowest level there's all sorts of complicated neural activity in our brains—most of it mired in computational irreducibility. But the "thin thread of conscious experience" that we extract from this we can for many purposes treat as being made up of higher-level "units of thought", or essentially "discrete concepts".

And, yes, it's certainly our typical human experience that robust constructs—and particularly ones from which other constructs can be built—will be discrete. In principle one can imagine that there could be things like "robust continuous spaces of concepts" ("cat and dog and everything in between"). But we don't have anything like the computational paradigm that shows us a consistent universal way that such things could fit together (there's no robust analog of computation theory for real numbers, for example). And somehow the success of the computational paradigm—potentially all the way down to the foundations of the physical universe—doesn't seem to leave much room for anything else.

So, OK, let's imagine that we can represent our thread of conscious experience in terms of concepts. Well, that's close to saying that we're using language. We're "packaging up" the details of our neural activity into "robust elements" which we can think of as concepts—and which are represented in language essentially by words. And not only does this "packaging" into language give a robust way for different brains to communicate; it also gives a single brain a robust way to "remember" and "redeploy" thoughts.

Within one brain one could imagine that one might be able to remember and "think" directly in terms of detailed low-level neural patterns. But no doubt the "neural environment" inside a brain is continually changing (not least because of its stream of sensory input). And so the only way to successfully "preserve a thought" across time is presumably to "package it up" in terms of robust elements, or essentially in terms of language. In other words, if we're going to be able to consistently "think a particular thought" we probably have to formulate it in terms of something robust—like concepts.

But, OK, individual concepts are one thing. But language—or at least human language—is based on putting together concepts in structured ways. One might take a noun ("cat") and qualify it with an adjective ("black") to form a phrase that's in effect a finer-grained version of the concept represented by the noun. And in a rough approximation one can think of language as formed from trees of nested phrases like this. And insofar as the phrases are independent in their structure (i.e. "context free"), we can parse such language by recursively understanding each phrase in turn—with the constraint that we can't do it if the nesting goes too deep for us to hold the necessary stack of intermediate steps in our working memory.

An important feature of ordinary human language is that it's ultimately presented in a sequential way. Even though it may consist of a nested tree of phrases, the words that are the leaves of that tree are spoken or written in a one-dimensional sequence. And, yes, the fact that this is how it works is surely closely connected to the fact that our brains construct a single thread of conscious experience.

In the actuality of the few thousand human languages currently in use, there is considerable superficial diversity, but also considerable fundamental

commonality. For example, the same parts of speech (noun, verb, etc.) typically show up, as do concepts like "subject" and "object". But the details of how words are put together, and how things are indicated, can be fairly different. Sometimes nouns have case endings; sometimes there are separate prepositions. Sometimes verb tenses are indicated by annotating the verb; sometimes with extra words. And sometimes, for example, what would usually be whole phrases can be smooshed together into single words.

It's not clear to what extent commonalities between languages are the result of shared history, and to what extent they're consequences either of the particulars of our human sensory experience of the world, or the particular construction of our brains. It's not too hard to get something like concepts to emerge in experiments on training neural nets to pass data through a "bottleneck" that simulates a "mind-to-mind communication channel". But how compositionality or grammatical structure might emerge is not clear.

OK, but so what might change if we had bigger brains? If neural nets are a guide, one obvious thing is that we should be able to deal directly with a larger number of "distinct concepts", or words. So what consequences would this have? Presumably one's language would get "grammatically shallower", in the sense that what would otherwise have had to be said with nested phrases could now be said with individual words. And presumably this would tend to lead to "faster communication", requiring fewer words. But it would likely also lead to more rigid communication, with less ability to tweak shades of meaning, say by changing just a few words in a phrase. (And it would presumably also require longer training, to learn what all the words mean.)

In a sense we have a preview of what it's like to have more words whenever we deal with specialized versions of existing language, aimed say at particular technical fields. There are additional words of "jargon" available, that make certain things "faster to say" (but require longer to learn). And with that jargon comes a certain rigidity, in saying easily only what the jargon says, and not something slightly different.

So how else could language be different with a bigger brain? With larger working memory, one could presumably have more deeply nested phrases. But what about

more sophisticated grammatical structures, say ones that aren't "context free", in the sense that different nested phrases can't be parsed separately? My guess is that this quickly devolves into requiring arbitrary computation—and runs into computational irreducibility. In principle it's perfectly possible to have any program as the "message" one communicates. But if one has to run the program to "determine its meaning", that's in general going to involve computational irreducibility.

And the point is that with our assumptions about what "brain-like systems" do, that's something that's out of scope. Yes, one can construct a system (even with neurons) that can do it. But not with the "single thread of decisions from sensory input" workflow that seems characteristic of brains. (There are finer gradations one could consider—like languages that are context sensitive but don't require general computation. But the Principle of Computational Equivalence strongly suggests that the separation between nested context-free systems and ones associated with arbitrary computation is very thin, and there doesn't seem to be any particular reason to expect that the capabilities of a bigger brain would land right there.)

Said another way: the Principle of Computational Equivalence says it's easy to have a system that can deal with arbitrary computation. It's just that such a system is not "brain like" in its behavior; it's more like a typical system we see in nature.

OK, but what other "additional features" can one imagine, for even roughly "brain-like" systems? One possibility is to go beyond the idea of a single thread of experience, and to consider a multiway system in which threads of experience can branch and merge. And, yes, this is what we imagine happens at a low level in the physical universe, particularly in connection with quantum mechanics. And indeed it's perfectly possible to imagine, for example, a "quantum-like" LLM system in which one generates a graph of different textual sequences. But just "scaling up the number of neurons" in a brain, without changing the overall architecture, won't get to this. We have to have a different, multiway architecture. Where we have a "graph of consciousness" rather than a "stream of

consciousness", and where, in effect, we're "thinking a graph of thoughts", notably with thoughts themselves being able to branch and merge.

In our practical use of language, it's most often communicated in spoken or written form—effectively as a one-dimensional sequence of tokens. But in math, for example, it's common to have a certain amount of 2D structure, and in general there are also all sorts of specialized (usually technical) diagrammatic representations in use, often based on using graphs and networks—as we'll discuss in more detail below.

But what about general pictures? Normally it's difficult for us to produce these. But in generative AI systems it's basically easy. So could we then imagine directly "communicating mental images" from one mind to another? Maybe as a practical matter some neural implant in our brain could aggregate neural signals from which a displayed image could be generated. But is there in fact something coherent that could be extracted from our brains in this way? Perhaps that can only happen after "consensus is formed", and we've reduced things to a much thinner "thread of experience". Or, in other words, perhaps the only robust way for us to "think about images" is in effect to reduce them to discrete concepts and language-like representations.

But perhaps if we "had the hardware" to display images directly from our minds it'd be a different story. And it's sobering to imagine that perhaps the reason cats and dogs don't appear to have compositional language is just that they don't "have the hardware" to talk like we do (and it's too laborious for them to "type with their paws", etc.). And, by analogy, that if we "had the hardware" for displaying images, we'd discover we could also "think very differently".

Of course, in some small ways we do have the ability to "directly communicate with images", for example in our use of gestures and body language. Right now, these seem like largely ancillary forms of communication. But, yes, it's conceivable that with bigger brains, they could be more.

And when it comes to other animals the story can be different. Cuttlefish are notable for dynamically producing elaborate patterns on their skin—giving them in a sense the hardware to "communicate in pictures". But so far as one can tell, they produce just a small number of distinct patterns—and certainly nothing like

a "pictorial generalization of compositional language". (In principle one could imagine that "generalized cuttlefish" could do things like "dynamically run cellular automata on their skin", just like all sorts of animals "statically" do in the process of growth or development. But to decode such patterns—and thereby in a sense enable "communicating in programs"—would typically require irreducible amounts of computation that are beyond the capabilities of any standard brain-like system.)

Sensors and Actuators

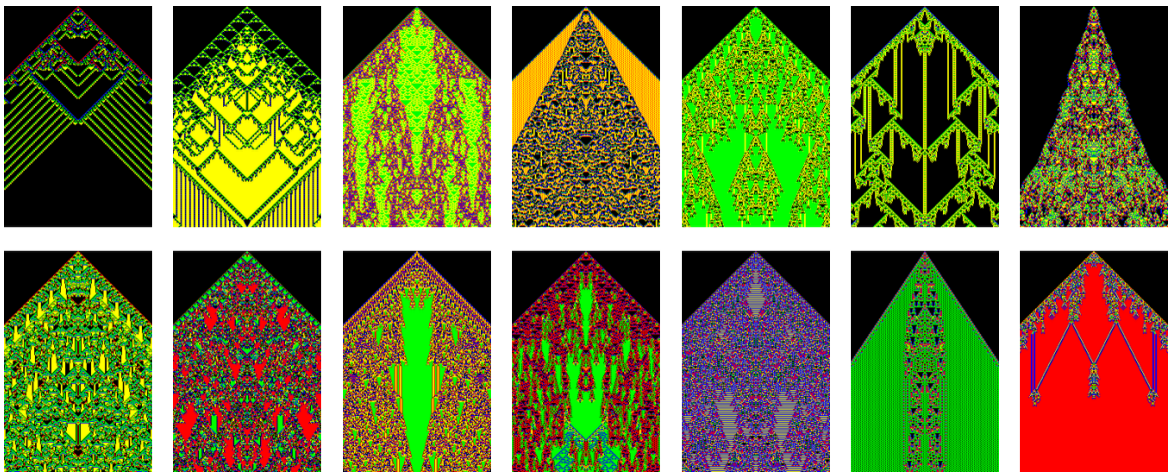
We humans have raw inputs coming into our brains from a few million sensors distributed across our usual senses of touch, sight, hearing, taste and smell (together with balance, temperature, hunger, etc.). In most cases the detailed sensor inputs are not independent; in a typical visual scene, for example, neighboring pixels are highly correlated. And it doesn't seem to take many layers of neurons in our brains to distill our typical sensory experience from pure pieces of "raw data" to what we might view as "more independent features".

Of course there'll usually be much more in the raw data than just those features. But the "features" typically correspond to aspects of the data that we've "learned are useful to us"—normally connected to pockets of computational reducibility that exist in the environment in which we operate. Are the features we pick out all we'll ever need? In the end, we typically want to derive a small stream of decisions or actions from all the data that comes in. But how many "intermediate features" do we need to get "good" decisions or actions?

That really depends on two things. First, what our decisions and actions are like. And second, what our raw data is like. Early in the history of our species, everything was just about "indigenous human experience": what the natural world is like, and what we can do with our bodies. But as soon as we were dealing with technology, that changed. And in today's world we're constantly exposed, for example, to visual input that comes not from the natural world, but, say, from digital displays.

And, yes, we often try to arrange our "user experience" to align with what's familiar from the natural world (say by having objects that stay unchanged when

they're moved across the screen). But it doesn't have to be that way. And indeed it's easy—even with simple programs—to generate for example visual images very different from what we're used to. And in many such cases, it's very hard for us to "tell what's going on" in the image. Sometimes it'll just "look too complicated". Sometimes it'll seem like it has pieces we should recognize, but we don't:



When it's "just too complicated", that's often a reflection of computational irreducibility. But when there are pieces we might "think we should recognize", that can be a reflection of pockets of reducibility we're just not familiar with. If we imagine a space of possible images—as we can readily produce with generative AI—there will be some that correspond to concepts (and words) we're familiar with. But the vast majority will effectively lie in "interconcept space": places where we could have concepts, but don't, at least yet:

So what could bigger brains do with all this? Potentially they could handle more features, and more concepts. Full computational irreducibility will always in effect ultimately overpower them. But when it comes to handling pockets of reducibility, they'll presumably be able to deal with more of them. So in the end,

it's very much as one might expect: a bigger brain should be able to track more things going on, "see more details", etc.

Brains of our size seem like they are in effect sufficient for "indigenous human experience". But with technology in the picture, it's perfectly possible to "overload" them. (Needless to say, technology—in the form of filtering, data analysis, etc.—can also reduce that overload, in effect taking raw input and bringing our actual experience of it closer to something "indigenous".)

It's worth pointing out that while two brains of a given size might be able to "deal with the same number of features or concepts", those features or concepts might be different. One brain might have learned to talk about the world in terms of one set of primitives (such as certain basic colors); another in terms of a different set of primitives. But if both brains are sampling "indigenous human experience" in similar environments one can expect that it should be possible to translate between these descriptions—just as it is generally possible to translate between things said in different human languages.

But what if the brains are effectively sampling "different slices of reality"? What if one's using technology to convert different physical phenomena to forms (like images) that we can "indigenously" handle? Perhaps we're sensing different electromagnetic frequencies; perhaps we're sensing molecular or chemical properties; perhaps we're sensing something like fluid motion. The kinds of features that will be "useful" may be quite different in these different modalities. Indeed, even something as seemingly basic as the notion of an "object" may not be so relevant if our sensory experience is effectively of continuous fluid motion.

But in the end, what's "useful" will depend on what we can do. And once again, it depends on whether we're dealing with "pure humans" (who can't, for example, move like octopuses) or with humans "augmented by technology". And here we start to see an issue that relates to the basic capabilities of our brains.

As "pure humans", we have certain "actuators" (basically in the form of muscles) that we can "indigenously" operate. But with technology it's perfectly possible for us to use quite different actuators in quite different configurations. And as a practical matter, with brains like ours, we may not be able to make them work.

For example, while humans can control helicopters, they never managed to control quadcopters—at least not until digital flight controllers could do most of the work. In a sense there were just too many degrees of freedom for brains like ours to deal with. Should bigger brains be able to do more? One would think so. And indeed one could imagine testing this with artificial neural nets. In millipedes, for example, their actual brains seem to support only a couple of patterns of motion of their legs (roughly, same phase vs. opposite phase). But one could imagine that with a bigger brain, all sorts of other patterns would become possible.

Ultimately, there are two issues at stake here. The first is having a brain be able to "independently address" enough actuators, or in effect enough degrees of freedom. The second is having a brain be able to control those degrees of freedom. And for example with mechanical degrees of freedom there are again essentially issues of computational irreducibility. Looking at the space of possible configurations—say of millipede legs—does one effectively just have to trace the path to find out if, and how, one can get from one configuration to another? Or are there instead pockets of reducibility, associated with regularities in the space of configurations, that let one "jump ahead" and figure this out without tracing all the steps? It's those pockets of reducibility that brains can potentially make use of.

When it comes to our everyday "indigenous" experience of the world, we are used to certain kinds of computational reducibility, associated for example with familiar natural laws, say about motion of objects. But what if we were dealing with different experiences, associated with different senses?

For example, imagine (as with dogs) that our sense of smell was better developed than our sense of sight—as reflected by more nerves coming into our brains from our noses than our eyes. Our description of the world would then be quite different, based for example not on geometry revealed by the line-of-sight arrival of light, but instead by the delivery of odors through fluid motion and diffusion—not to mention the probably-several-hundred-dimensional space of odors, compared to the red, green, blue space of colors. Once again there would be features that could be identified, and "concepts" that could be defined. But those

might only be useful in an environment "built for smell" rather than one "built for sight".

And in the end, how many concepts would be useful? I don't think we have any way to know. But it certainly seems as if one can be a successful "smell-based animal" with a smaller brain (presumably supporting fewer concepts) than one needs as a successful "sight-based animal".

One feature of "natural senses" is that they tend to be spatially localized: an animal basically senses things only where it is. (We'll discuss the case of social organisms later.) But what if we had access to a distributed array of sensors—say associated with IoT devices? The "effective laws of nature" that one could perceive would then be different. Maybe there would be regularities that could be captured by a small number of concepts, but it seems more likely that the story would be more complicated, and that in effect one would "need a bigger brain" to be able to keep track of what's going on, and make use of whatever pockets of reducibility might exist.

There are somewhat similar issues if one imagines changing the timescales for sensory input. Our perception of space, for example, depends on the fact that light travels fast enough that in the milliseconds it takes our brain to register the input, we've already received light from everything that's around us. But if our brains operated a million times faster (as digital electronics does) we'd instead be registering individual photons. And while our brains might aggregate these to something like what we ordinarily perceive, there may be all sorts of other (e.g. quantum optics) effects that would be more obvious.

Abstraction

The more abstractly we try to think, the harder it seems to get. But would it get easier if we had bigger brains? And might there perhaps be fundamentally higher levels of abstraction that we could reach—but only if we had bigger brains.

As a way to approach such questions, let's begin by talking a bit about the history of the phenomenon of abstraction. We might already say that basic perception involves some abstraction, capturing as it does a filtered version of the world as it actually is. But perhaps we reach a different level when we start to ask "what if?"

questions, and to imagine how things in the world could be different than they are.

But somehow when it comes to us humans, it seems as if the greatest early leap in abstraction was the invention of language, and the explicit delineation of concepts that could be quite far from our direct experience. The earliest written records tend to be rather matter of fact, mostly recording as they do events and transactions. But already there are plenty of signs of abstraction. Numbers independent of what they count. Things that should happen in the future. The concept of money.

There seems to be a certain pattern to the development of abstraction. One notices that some category of things one sees many times can be considered similar, then one "packages these up" into a concept, often described by a word. And in many cases, there's a certain kind of self amplification: once one has a word for something (as a modern example, say "blog"), it becomes easier for us to think about the thing, and we tend to see it or make it more often in the world around us. But what really makes abstraction take off is when we start building a whole tower of it, with one abstract concept recursively being based on others.

Historically this began quite slowly. And perhaps it was seen first in theology. There were glimmerings of it in things like early (syllogistic) logic, in which one started to be able to talk about the form of arguments, independent of their particulars. And then there was mathematics, where computations could be done just in terms of numbers, independent of where those numbers came from. And, yes, while there were tables of "raw computational results", numbers were usually discussed in terms of what they were numbers of. And indeed when it came to things like measures of weight, it took until surprisingly modern times for there to be an absolute, abstract notion of weight, independent of whether it was a weight of figs or of wool.

The development of algebra in the early modern period can be considered an important step forward in abstraction. Now there were formulas that could be manipulated abstractly, without even knowing what particular numbers x stood for. But it would probably be fair to say that there was a major acceleration in abstraction in the 19th century—with the development of formal systems that

could be discussed in "purely symbolic form" independent of what they might (or might not) "actually represent".

And it was from this tradition that modern notions of computation emerged (and indeed particularly ones associated with symbolic computation that I personally have extensively used). But the most obvious area in which towers of abstraction have been built is mathematics. One might start with numbers (that could count things). But soon one's on to variables, functions, spaces of functions, category theory—and a zillion other constructs that abstractly build on each other.

The great value of abstraction is that it allows one to think about large classes of things all at once, instead of each separately. But how do those abstract concepts fit together? The issue is that often it's in a way that's very remote from anything about which we have direct experience from our raw perception of the world. Yes, we can define concepts about transfinite numbers or higher categories. But they don't immediately relate to anything we're familiar with from our everyday experience.

As a practical matter one can often get a sense of how high something is on the tower of abstraction by seeing how much one has to explain to build up to it from "raw experiential concepts". Just sometimes it turns out that actually, once one hears about a certain seemingly "highly abstract" concept, one can actually explain it surprisingly simply, without going through the whole historical chain that led to it. (A notable example of this is the concept of universal computation—which arose remarkably late in human intellectual history, but is now quite easy to explain, albeit particularly given its actual widespread embodiment in technology.) But the more common case is that there's no choice but to explain a whole tower of concepts.

At least in my experience, however, when one actually thinks about "highly abstract" things, one does it by making analogies to more familiar, more concrete things. The analogies may not be perfect, but they provide scaffolding which allows our brains to take what would otherwise be quite inaccessible steps.

At some level any abstraction is a reflection of a pocket of computational reducibility. Because if a useful abstraction can be defined, what it means is that it's possible to say something in a "summarized" or reduced way, in effect

"jumping ahead", without going through all the computational steps or engaging with all the details. And one can then think of towers of abstraction as being like networks of pockets of computational reducibility. But, yes, it can be hard to navigate these.

Underneath, there's lots of computational irreducibility. And if one is prepared to "go through all the steps" one can often "get to an answer" without all the "conceptual difficulty" of complex abstractions. But while computers can often readily "go through all the steps", brains can't. And that's in a sense why we have to use abstraction. But inevitably, even if we're using abstraction, and the pockets of computational reducibility associated with it, there'll be shadows of the computational irreducibility underneath. And in particular, if we try to "explore everything", our network of pockets of reducibility will inevitably "get complicated", and ultimately also be mired in computational irreducibility, albeit with "higher-level" constructs than in the computational irreducibility underneath.

No finite brain will ever be able to "go all the way", but it starts to seem likely that a bigger brain will be able to "reach further" in the network of abstraction. But what will it find there? How does the character of abstraction change when we take it further? We'll be able to discuss this a bit more concretely when we talk about computational language below. But perhaps the main thing to say now is that—at least in my experience—most higher abstractions don't feel as if they're "structurally different" once one understands them. In other words, most of the time, it seems as if the same patterns of thought and reasoning that one's applied in many other places can be applied there too, just to different kinds of constructs.

Sometimes, though, there seem to be exceptions. Shocks to intuition that seem to separate what one's now thinking about from anything one's thought before. And, for example, for me this happened when I started looking broadly at the computational universe. I had always assumed that simple rules would lead to simple behavior. But many years ago I discovered that in the computational universe this isn't true (hence computational irreducibility). And this led to a whole different paradigm for thinking about things.

It feels a bit like in metamathematics. Where one can imagine one type of abstraction associated with different constructs out of which to form theorems. But where somehow there's another level associated with different ways to build new theorems, or indeed whole spaces of theorems. Or to build proofs from proofs, or proofs from proofs of proofs, etc. But the remarkable thing is that there seems to be an ultimate construct that encompasses it all: the ruliad.

We can describe the ruliad as the entangled limit of all possible computations. But we can also describe it as the limit of all possible abstractions. And it seems to lie underneath all physical reality, as well as all possible mathematics, etc. But, we might ask, how do brains relate to it?

Inevitably, it's full of computational irreducibility. And looked at as a whole, brains can't get far with it. But the key idea is to think about how brains as they are—with all their various features and limitations—will "parse" it. And what I've argued is that what "brains as they are" will perceive about the ruliad are the core laws of physics (and mathematics) as we know them. In other words, it's because brains are the way they are that we perceive the laws of physics that we perceive.

Would it be different for bigger brains? Not if they're the "same kind of brains". Because what seems to matter for the core laws of physics are really just two properties of observers. First, that they're computationally bounded. And second, that they believe they are persistent in time, and have a single thread of experience through time. And both of these seem to be core features of what makes brains "brain-like", rather than just arbitrary computational systems.

It's a remarkable thing that just these features are sufficient to make core laws of physics inevitable. But if we want to understand more about the physics we've constructed—and the laws we've deduced—we probably have to understand more about what we're like as observers. And indeed, as I've argued elsewhere, even our physical scale (much bigger than molecules, much smaller than the whole universe) is for example important in giving us the particular experience (and laws) of physics that we have.

Would this be different with bigger brains? Perhaps a little. But anything that something brain-like can do pales in comparison to the computational irreducibility that exists in the ruliad and in the natural world. Nevertheless, with

every new pocket of computational reducibility that's reached we get some new abstraction about the world, or in effect, some new law about how the world works.

And as a practical matter, each such abstraction can allow us to build a whole collection of new ways of thinking about the world, and making things in the world. It's challenging to trace this arc. Because in a sense it'll all be about "things we never thought to think about before". Goals we might define for ourselves that are built on a tower of abstraction, far away from what we might think of as "indigenous human goals".

It's important to realize that there won't just be one tower of abstraction that can be built. There'll inevitably be an infinite network of pockets of computational reducibility, with each path leading to a different specific tower of abstraction. And indeed the abstractions we have pursued reflect the particular arc of human intellectual history. Bigger brains—or AIs—have many possible directions they can go, each one defining a different path of history.

One question to ask is to what extent reaching higher levels of abstraction is a matter of education, and to what extent it requires additional intrinsic capabilities of a brain. It is, I suspect, a mixture. Sometimes it's really just a question of knowing "where that pocket of reducibility is", which is something we can learn from education. But sometimes it's a question of navigating a network of pockets, which may only be possible when brains reach a certain level of "computational ability".

There's another thing to discuss, related to education. And that's the fact that over time, more and more "distinct pieces of knowledge" get built up in our civilization. There was perhaps a time in history when a brain of our size could realistically commit to memory at least the basics of much of that knowledge. But today that time has long passed. Yes, abstraction in effect compresses what one needs to know. But the continual addition of new and seemingly important knowledge, across countless specialties, makes it impossible for brains of our size to keep up.

Plenty of that knowledge is, though, quite siloed in different areas. But sometimes there are "grand analogies" to make—say pulling an idea from

relativity theory and applying it to biological evolution. In a sense such analogies reveal new abstractions—but to make them requires knowledge that spans many different areas. And that's a place where bigger brains—or AIs—can potentially do something that's in a fundamental way "beyond us".

Will there always be such "grand analogies" to make? The general growth of knowledge is inevitably a computationally irreducible process. And within it there will inevitably be pockets of reducibility. But how often in practice will one actually encounter "long-range connections" across "knowledge space"? As a specific example one can look at metamathematics, where such connections are manifest in theorems that link seemingly different areas of mathematics. And this example leads one to realize that at some deep level grand analogies are in a sense inevitable. In the context of the ruliad, one can think of different domains of knowledge as corresponding to different parts. But the nature of the ruliad—encompassing as it does everything that is computationally possible—inevitably imbues it with a certain homogeneity, which implies that (as the Principle of Computational Equivalence might suggest) there must ultimately be a correspondence between different areas. In practice, though, this correspondence may be at a very "atomic" (or "formal") level, far below the kinds of descriptions (based on pockets of reducibility) that we imagine brains normally use.

But, OK, will it always take an "expanding brain" to keep up with the "expanding knowledge" we have? Computational irreducibility guarantees that there'll always in principle be "new knowledge" to be had—separated from what's come before by irreducible amounts of computation. But then there's the question of whether in the end we'll care about it. After all, it could be that the knowledge we can add is so abstruse that it will never affect any practical decisions we have to make. And, yes, to some extent that's true (which is why only some tiny fraction of the Earth's population will care about what I'm writing here). But another consequence of computational irreducibility is that there will always be "surprises"—and those can eventually "push into focus" even what at first seems like arbitrarily obscure knowledge.

Computational Language

Language in general—and compositional language in particular—is arguably the greatest invention of our species. But is it somehow "the top"—the highest possible representation of things? Or if, for example, we had bigger brains, is there something beyond it that we could reach?

Well, in some very formal sense, yes, compositional language (at least in idealized form) is "the top". Because—at least if it's allowed to include utterances of any length—then in some sense it can in principle encode arbitrary, universal computations. But this really isn't true in any useful sense—and indeed to apply ordinary compositional language in this way would require doing computationally irreducible computations.

So we return to the question of what might in practice lie beyond ordinary human language. I wondered about this for a long time. But in the end I realized that the most important clue is in a sense right in front of me: the concept of computational language, that I've spent much of my life exploring.

It's worth saying at the outset that the way computational language plays out for computers and for brains is somewhat different, and in some respects complementary. In computers you might specify something as a Wolfram Language symbolic expression, and then the "main action" is to evaluate this expression, potentially running a long computation to find out what the expression evaluates to.

Brains aren't set up to do long computations like this. For them a Wolfram Language expression is something to use in effect as a "representation of a thought". (And, yes, that's an important distinction between the computational language concept of Wolfram Language, and standard "programming languages", which are intended purely as a way to tell a computer what to do, not a way to represent thoughts.)

So what kinds of thoughts can we readily represent in our computational language? There are ones involving explicit numbers, or mathematical expressions. There are ones involving cities and chemicals, and other real-world entities. But then there are higher-level ones, that in effect describe more abstract structures.

For example, there's **NestList**, which gives the result of nesting any operation, here named *f*:

```
In[•]:= NestList[f, x, 5]
Out[•]= {x, f[x], f[f[x]], f[f[f[x]]], f[f[f[f[x]]]], f[f[f[f[f[x]]]]]}
```

At the outset, it's not obvious that this would be a useful thing to do. But in fact it's a very successful abstraction: there are lots of functions *f* for which one wants to do this.

In the development of ordinary human language, words tend to get introduced when they're useful, or, in other words, when they express things one often wants to express. But somehow in human language the words one gets tend to be more concrete. Maybe they describe something that directly happens to objects in the world. Maybe they describe our impression of a human mental state. Yes, one can make rather vague statements like "I'm going to do something to someone". But human language doesn't normally "go meta", doing things like **NestList** where one's saying that one wants to take some "direct statement" and in effect "work with the statement". In some sense, human language tends to "work with data", applying a simple analog of code to it. Our computational language can "work with code" as "raw material".

One can think about this as a "higher-order function": a function that operates not on data, but on functions. And one can keep going, dealing with functions that operate on functions that operate on functions, and so on. And at every level one is increasing the generality—and abstraction—at which one is working. There may be many specific functions (a bit analogous to verbs) that operate on data (a bit analogous to nouns). But when we talk about operating on functions themselves we can potentially have just a single function (like **NestList**) that operates, quite generally, on many functions. In ordinary language, we might call such things "metaverbs", but they aren't something that commonly occurs.

But what makes them possible in computational language? Well, it's taking the computational paradigm seriously, and representing everything in computational terms: objects, actions, etc. In Wolfram Language, it's that we can represent everything as a symbolic expression. Arrays of numbers (or countries, or

whatever) are symbolic expressions. Graphics are symbolic expressions. Programs are symbolic expressions. And so on.

And given this uniformity of representation it becomes feasible—and natural—to do higher-order operations, that in effect manipulate symbolic structure without being concerned about what the structure might represent. At some level we can view this as leading to the ultimate abstraction embodied in the ruliad, where in a sense "everything is pure structure". But in practice in Wolfram Language we try to "anchor" what we're doing to known concepts from ordinary human language—so that we use names for things (like **NestList**) that are derived from common English words.

In some formal sense this isn't necessary. Everything can be "purely structural", as it is not only in the ruliad but also in constructs like combinators, where, say, the operation of addition can be represented by:

```
S•(K•S)•(S•(K•(S•(K•S)•K)))
```

Combinators have been around for more than a century. But they are almost impenetrably difficult for most humans to understand. Somehow they involve too much "pure abstraction", not anchored to concepts we "have a sense of" in our brains.

It's been interesting for me to observe over the years what it's taken for people (including myself) to come to terms with the kind of higher-order constructs that exist in the Wolfram Language. The typical pattern is that over the course of months or years one gets used to lots of specific cases. And only after that is one able—often in the end rather quickly—to "get to the next level" and start to use some generalized, higher-order construct. But normally one can in effect only "go one level at a time". After one groks one level of abstraction, that seems to have to "settle" for a while before one can go on to the next one.

Somehow it seems as if one is gradually "feeling out" a certain amount of computational irreducibility, to learn about a new pocket of reducibility, that one can eventually use to "think in terms of".

Could "having a bigger brain" speed this up? Maybe it'd be useful to be able to remember more cases, and perhaps get more into "working memory". But I

rather suspect that combinators, for example, are in some sense fundamentally beyond all brain-like systems. It's much as the Principle of Computational Equivalence suggests: one quickly "ascends" to things that are as computationally sophisticated as anything—and therefore inevitably involve computational irreducibility. There are only certain specific setups that remain within the computationally bounded domain that brain-like systems can deal with.

Of course, even though they can't directly "run code in their brains", humans—and LLMs—can perfectly well use Wolfram Language as a tool, getting it to actually run computations. And this means they can readily "observe phenomena" that are computationally irreducible. And indeed in the end it's very much the same kind of thing observing such phenomena in the abstract computational universe, and in the "real" physical universe. And the point is that in both cases, brain-like systems will pull out only certain features, essentially corresponding to pockets of computational reducibility.

How do things like higher-order functions relate to this? At this point it's not completely clear. Presumably in at least some sense there are hierarchies of higher-order functions that capture certain kinds of regularities that can be thought of as associated with networks of computational reducibility. And it's conceivable that category theory and its higher-order generalizations are relevant here. In category theory one imagines applying sequences of functions ("morphisms") and it's a foundational assumption that the effect of any sequence of functions can also be represented by just a single function—which seems tantamount to saying that one can always "jump ahead", or in other words, that everything one's dealing with is computationally reducible. Higher-order category theory then effectively extends this to higher-order functions, but always with what seem like assumptions of computational reducibility.

And, yes, this all seems highly abstract, and difficult to understand. But does it really need to be, or is there some way to "bring it down" to a level that's close to everyday human thinking? It's not clear. But in a sense the core art of computational language design (that I've practiced so assiduously for nearly half a century) is precisely to take things that at first might seem abstruse, and somehow cast them into an accessible form. And, yes, this is something that's about as intellectually challenging as anything—because in a sense it involves

continually trying to "figure out what's really going on", and in effect "drilling down" to get to the foundations of everything.

But, OK, when one gets there, how simple will things be? Part of that depends on how much computational irreducibility is left when one reaches what one considers to be "the foundations". And part in a sense depends on the extent to which one can "find a bridge" between the foundations and something that's familiar. Of course, what's "familiar" can change. And indeed over the four decades that I've been developing the Wolfram Language quite a few things (particularly in areas like functional programming) that at first seemed abstruse and unfamiliar have begun to seem more familiar. And, yes, it's taken the collective development and dissemination of the relevant ideas to achieve that. But now it "just takes education"; it doesn't "take a bigger brain" to deal with these things.

One of the core features of the Wolfram Language is that it represents everything as a symbolic expression. And, yes, symbolic expressions are formally able to represent any kind of computational structure. But beyond that, the important point is that they're somehow set up to be a match for how brains work.

And in particular, symbolic expressions can be thought of "grammatically" as consisting of nested functions that form a tree-like structure; effectively a more precise version of the typical kind of grammar that we find in human language. And, yes, just as we manage to understand and generate human language with a limited working memory, so (at least at the grammatical level) we can do the same thing with computational language. In other words, in dealing with Wolfram Language we're leveraging our faculties with human language. And that's why Wolfram Language can serve as such an effective bridge between the way we think about things, and what's computationally possible.

But symbolic expressions represented as trees aren't the only conceivable structures. It's also possible to have symbolic expressions where the elements are nodes on a graph, and the graph can even have loops in it. Or one can go further, and start talking, for example, about the hypergraphs that appear in our Physics Project. But the point is that brain-like systems have a hard time processing such structures. Because to keep track of what's going on they in a sense have to keep

track of multiple "threads of thought". And that's not something individual brain-like systems as we current envision them can do.

Many Brains Together: The Formation of Society

As we've discussed several times here, it seems to be a key feature of brains that they create a single "thread of experience". But what would it be like to have multiple threads? Well, we actually have a very familiar example of that: what happens when we have a whole collection of people (or other animals).

One could imagine that biological evolution might have produced animals whose brains maintain multiple simultaneous threads of experience. But somehow it has ended up instead restricting each animal to just one thread of experience—and getting multiple threads by having multiple animals. (Conceivably creatures like octopuses may actually in some sense support multiple threads within one organism.)

Within a single brain it seems important to always "come to a single, definite conclusion"—say to determine where an animal will "move next". But what about in a collection of organisms? Well, there's still some kind of coordination that will be important to the fitness of the whole population—perhaps even something as direct as moving together as a herd or flock. And in a sense, just as all those different neuron firings in one brain get collected to determine a "final conclusion for what to do", so similarly the conclusions of many different brains have to be collected to determine a coordinated outcome.

But how can a coordinated outcome arise? Well, there has to be communication of some sort between organisms. Sometimes it's rather passive (just watch what your neighbor in a herd or flock does). Sometimes it's something more elaborate and active—like language. But is that the best one can do? One might imagine that there could be some kind of "telepathic coordination", in which the raw pattern of neuron firings is communicated from one brain to another. But as we've argued, such communication cannot be expected to be robust. To achieve robustness, one must "package up" all the internal details into some standardized form of communication (words, roars, calls, etc.) that one can expect can be "faithfully unpacked" and in effect "understood" by other, suitably similar brains.

But it's important to realize that the very possibility of such standardized communication in effect requires coordination. Because somehow what goes on in one brain has to be aligned with what goes on in another. And indeed the way that's maintained is precisely through continual communication.

So, OK, how might bigger brains affect this? One possibility is that they might enable more complex social structures. There are plenty of animals with fairly small brains that successfully form "all do the same thing" flocks, herds and the like. But the larger brains of primates seem to allow more complex "tribal" structures. Could having a bigger brain let one successfully maintain a larger social structure, in effect remembering and handling larger numbers of social connections? Or could the actual forms of these connections be more complex? While human social connections seem to be at least roughly captured by social networks represented as ordinary graphs, maybe bigger brains would for example routinely require hypergraphs.

But in general we can say that language—or standardized communication of some form—is deeply connected to the existence of a "coherent society". For without being able to exchange something like language there's no way to align the members of a potential society. And without coherence between members something like language won't be useful.

As in so many other situations, one can expect that the detailed interactions between members of a society will show all sorts of computational irreducibility. And insofar as one can identify "the will of society" (or, for that matter, the "tide of history"), it represents a pocket of computational reducibility in the system.

In human society there is a considerable tendency (though it's often not successful) to try to maintain a single "thread of society", in which, at some level, everyone is supposed to act more or less the same. And certainly that's an important simplifying feature in allowing brains like ours to "navigate the social world". Could bigger brains do something more sophisticated? As in other areas, one can imagine a whole network of regularities (or pockets of reducibility) in the structure of society, perhaps connected to a whole tower of "higher-order social abstractions", that only brains bigger than ours can comfortably deal with. ("Just being friends" might be a story for the "small brained". With bigger brains one

might instead have patterns of dependence and connectivity that can only be represented in complicated graph theoretic ways.)

Minds beyond Ours

We humans have a tremendous tendency to think—or at least hope—that our minds are somehow "at the top" of what's possible. But with what we know now about computation and how it operates in the natural world it's pretty clear this isn't true. And indeed it seems as if it's precisely a limitation in the "computational architecture" of our minds—and brains—that leads to that most cherished feature of our existence that we characterize as "conscious experience".

In the natural world at large, computation is in some sense happening quite uniformly, everywhere. But our brains seem to be set up to do computation in a more directed and more limited way—taking in large amounts of sensory data, but then filtering it down to a small stream of actions to take. And, yes, one can remove this "limitation". And while the result may lead to more computation getting done, it doesn't lead to something that's "a mind like ours".

And indeed in what we've done here, we've tended to be very conservative in how we imagine "extending our minds". We've mostly just considered what might happen if our brains were scaled up to have more neurons, while basically maintaining the same structure. (And, yes, animals physically bigger than us already have larger brains—as did Neanderthals—but what we really need to look at is size of brain relative to size of the animal, or, in effect "amount of brain for a given amount of sensory input".)

A certain amount about what happens with different scales of brains is already fairly clear from looking at different kinds of animals, and at things like their apparent lack of human-like language. But now that we have artificial neural nets that do remarkably human-like things we're in a position to get a more systematic sense of what different scales of "brains" can do. And indeed we've seen a sequence of "capability thresholds" passed as neural nets get larger.

So what will bigger brains be able to do? What's fairly straightforward is that they'll presumably be able to take larger amounts of sensory input, and generate larger amounts of output. (And, yes, the sensory input could come from existing

modalities, or new ones, and the outputs could go to existing "actuators", or new ones.) As a practical matter, the more "data" that has to be processed for a brain to "come to a decision" and generate an output, the slower it'll probably be. But as brains get bigger, so presumably will the size of their working memory—as well as the number of distinct "concepts" they can "distinguish" and "remember".

If the same overall architecture is maintained, there'll still be just a single "thread of experience", associated with a single "thread of communication", or a single "stream of tokens". At the size of brains we have, we can deal with compositional language in which "concepts" (represented, basically, as words) can have at least a certain depth of qualifiers (corresponding, say, to adjectival phrases). As brain size increases, we can expect there can both be more "raw concepts"—allowing fewer qualifiers—as well as more working memory to deal with more deeply nested qualifiers.

But is there something qualitatively different that can happen with bigger brains? Computational language (and particularly my experience with the Wolfram Language) gives some indications, the most notable of which is the idea of "going meta" and using "higher-order constructs". Instead of, say, operating directly on "raw concepts" with (say, "verb-like") "functions", we can imagine higher-order functions that operate on functions themselves. And, yes, this is something of which we see powerful examples in the Wolfram Language. But it feels as if we could somehow go further—and make this more routine—if our brains in a sense had "more capacity".

To "go meta" and "use higher-order constructs" is in effect a story of abstraction—and of taking many disparate things and abstracting to the point where one can "talk about them all together". The world at large is full of complexity—and computational irreducibility. But in essence what makes "minds like ours" possible is that there are pockets of computational reducibility to be found. And those pockets of reducibility are closely related to being able to successfully do abstraction. And as we build up towers of abstraction we are in effect navigating through networks of pockets of computational reducibility.

The progress of knowledge—and the fact that we're educated about it—lets us get to a certain level of abstraction. And, one suspects, the more capacity there is in a

brain, the further it will be able to go.

But where will it "want to go"? The world at large—full as it is with computational irreducibility, along with infinite numbers of pockets of reducibility—leaves infinite possibilities. And it is largely the coincidence of our particular history that defines the path we have taken.

We often identify our "sense of purpose" with the path we will take. And perhaps the definiteness of our belief in purpose is related to the particular feature of brains that leads us to concentrate "everything we're thinking" down into just a single stream of decisions and action.

And, yes, as we've discussed, one could in principle imagine "multiway minds" with multiple "threads of consciousness" operating at once. But we humans (and individual animals in general) don't seem to have those. Of course, in collections of humans (or other animals) there are still inevitably multiple "threads of consciousness" —and it's things like language that "knit together" those threads to, for example, make a coherent society.

Quite what that "knitting" looks like might change as we scale up the size of brains. And so, for example, with bigger brains we might be able to deal with "higher-order social structures" that would seem alien and incomprehensible to us today.

So what would it be like to interact with a "bigger brain"? Inside, that brain might effectively use many more words and concepts than we know. But presumably it could generate at least a rough ("explain-like-I'm-5") approximation that we'd be able to understand. There might well be all sorts of abstractions and "higher-order constructs" that we are basically blind to. And, yes, one is reminded of something like a dog listening to a human conversation about philosophy—and catching only the occasional "sit" or "fetch" word.

As we've discussed several times here, if we remove our restriction to "brain-like" operation (and in particular to deriving a small stream of decisions from large amounts of sensory input) we're thrown into the domain of general computation, where computational irreducibility is rampant, and we can't in general expect to say much about what's going on. But if we maintain "brain-like operation", we're

instead in effect navigating through "networks of computational reducibility", and we can expect to talk about things like concepts, language and towers of abstraction.

From a foundational point of view, we can imagine any mind as in effect being at a particular place in the ruliad. When minds communicate, they are effectively exchanging the rulial analog of particles—robust concepts that are somehow unchanged as they propagate within the ruliad. So what would happen if we had bigger brains? In a sense it's a surprisingly "mechanical" story: a bigger brain—encompassing more concepts, etc.—in effect just occupies a larger region of rulial space. And the presence of abstraction—perhaps learned from a whole arc of intellectual history—can lead to more expansion in rulial space.

And in the end it seems that "minds beyond ours" can be characterized by how large the regions of the ruliad they occupy are. (Such minds are, in some very literal rulial sense, more "broad minded".) So what is the limit of all this? Ultimately, it's a "mind" that spans the whole ruliad, and in effect incorporates all possible computations. But in some fundamental sense this is not a mind like ours, not least because by "being everything" it "becomes nothing"—and one can no longer identify it as having a coherent "thread of individual existence".

And, yes, the overall thrust of what we've been saying applies just as well to "AI minds" as to biological ones. If we remove restrictions like being set up to generate the next token, we'll be left with a neural net that's just "doing computation", with no obvious "mind-like purpose" in sight. But if we make neural nets do typical "brain-like" tasks, then we can expect that they too will find and navigate pockets of reducibility. We may well not recognize what they're doing. But insofar as we can, then inevitably we'll mostly be sampling the parts of "minds beyond ours" that are aligned with "minds like ours". And it'll take progress in our whole human intellectual edifice to be able to fully appreciate what it is that minds beyond ours can do.

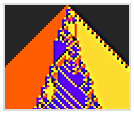
Thanks for recent discussions about topics covered here in particular to Richard Assar, Joscha Bach, Kovas Boguta, Thomas Dullien, Dugan Hammock, Christopher Lord, Fred Meinberg, Nora Popescu, Philip Rosedale, Terry Sejnowski, Hikari Sorensen, and James Wiles.

Cite this as	Stephen Wolfram (2025), "What If We Had Bigger Brains? Imagining Minds beyond Ours," Stephen Wolfram Writings. writings.stephenwolfram.com/2025/05/what-if-we-had-bigger-brains-imagining-minds-beyond-ours.

Posted in: Artificial Intelligence, Future Perspectives, Language & Communication, Life Science, Philosophy

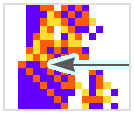
Join the discussion

Related Writings



What's Special about Life? Bulk Orchestration and the Rulial Ensemble in Biology and Beyond

November 11, 2025



Towards a Computational Formalization for Foundations of Medicine

February 3, 2025

Who Can Understand the Proof? A Window on Formalized Mathematics

January 9, 2025

Useful to the Point of Being Revolutionary: Introducing Wolfram Notebook Assistant

December 9, 2024

Popular Categories

Artificial Intelligence	Historical Perspectives	Personal Analytics
Big Picture	Language & Communication	Philosophy
Companies & Business	Life & Times	Physics
Computational Science	Life Science	Ruliology
Computational Thinking	Mathematica	Software Design
Data Science	Mathematics	Wolfram|Alpha
Education	New Kind of Science	Wolfram Language
Future Perspectives	New Technology	Other

Writings by Year

2025 | 2024 | 2023 | 2022 | 2021 | 2020 | 2019 | 2018 | 2017 | 2016 | 2015 |
2014 | 2013 | 2012 | 2011 | 2010 | 2009 | 2008 | 2007 | 2006 | 2004 | 2003 | All