New Orleans, LA
April 3, 2021

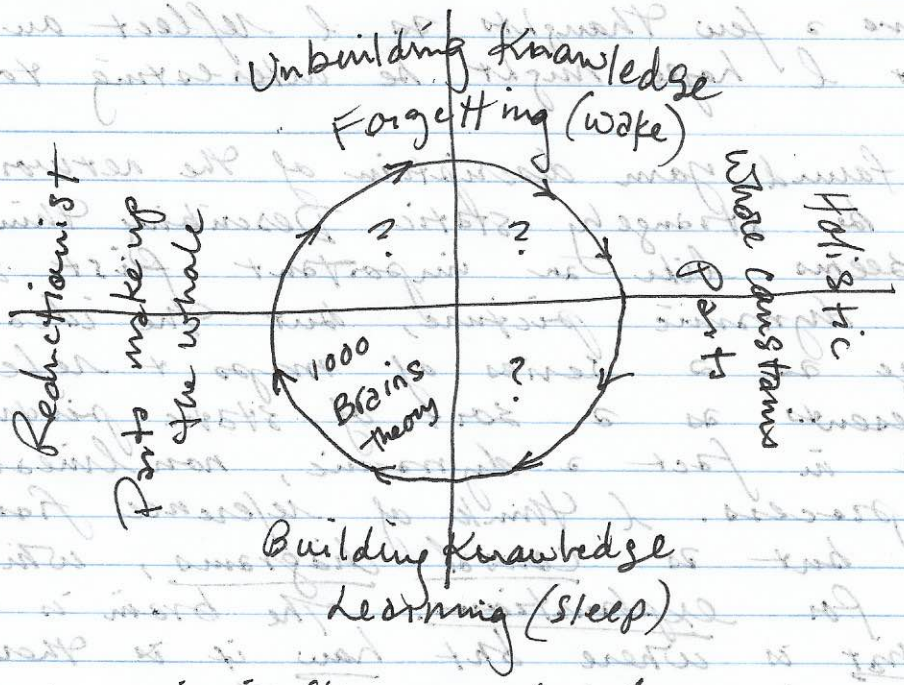
Dear Mr. Hawkins,

I am writing to thank you for your book A Thousand Brains. I found it clear, fascinating, and at times very compelling. I think almost certainly you have made a conceptual step towards the point of view about the brain which will turn out to be more right than the ones we have now. I strongly share your intuitions about the safety of intelligence, and the importance of embodiment. Your introduction of the mathematical term "reference frames" into the description of cortical columns certainly seems like a welcome advance. I am a non-scientist who has taken an interest in this subject, and I appreciated your clear and philosophical approach to the language.

I have a few thoughts as I reflect on your book that I hope might be interesting to you.

First: I found your description of the network of the brain to be strangely static. Describing thinking as moving seems like an important first step towards a dynamic picture, but the idea of knowledge as a series of maps + reference frames presents as a sort of static picture what must be in fact a dynamic, non-linear, recursive process. I think of reference frames ~~=~~ not as maps but as causal diagrams, which is another synonym for explanations. The brain is mapping not just what is where but how it is there, in other words, what wiggles what. For example, with your coffee cup, what a coffee cup is mapped within is a network of expectations that if behave a certain way, that it is brittle rather than soft, that it slides rather than bounces, and that it doesn't bite your hand, etc. In describing this as a multi-dimensional map you've given a metaphor that gives us the

impression that knowledge is set up as an static object in three dimensions — a series of cortical columns — a description that must be right, obviously; but it doesn't explain at all how that object comes to be configured one way & not another, and how it can change over time. In your account, knowledge is built by knitting details together to form a whole. This is an interesting way to think of cognition, but logically this can only describe one fourth of the process (1/4). There is not only remembering but also forgetting, and the whole is not only knit together by parts, but the parts are contextualized by the whole.



The x-axis in this graph above seems to provide the explanation for why the brain is divided and asymmetrical — a detail too important to leave out of a good general synthetic theory like yours. The y-axis, on the other hand, provides the explanation, in my opinion, for the sleep/wake cycle. My own conjecture is that the brain latches through

both of these axes continually, and that the building part of the cycle is actually provided by the sleep phase, and the unbuilding part of the cycle is provided by the waking phase.

There is convergent evolution on both of these features in intelligent animals. The octopus, for example, is a mollusk whose common ancestor with us was something like a clam — almost certainly not very smart. Yet the octopus has eight left brains in his arms for sensing the parts, and one right brain for synthesizing the whole, just as the bi/bi versions have different styles of cognition in each hemisphere. And the octopus not only sleeps eight hours a day, but some cephalopods exhibit REM sleep just as we do — the most intelligent ones to the largest degree.

Sleep cannot be for resting, for saving energy. First of all, it only saves about 15% in humans, a small energetic benefit. Sleep is physiologically nothing like hibernation, for certain. Hibernating animals accumulate a sleep deficit. Instead, sleep must be a repair phase, a rebuilding phase. Animals, including humans, that are deprived of all energy sources, fasted, for long periods sleep less, not more. You can try this yourself by prolonged water fasting. This result, across all known organisms, proves that sleep cannot be for energy conservation. If sleep is a repair phase essential for intelligence, it must be that what the nervous system is doing when it is asleep is building a more functional connective network, & that what we are doing while awake is inflicting useful damage onto that connective network.

We know that lifespans of animals are under

evolutionary control, as well as reproductive strategies. Lifespans, plus reproductive strategies, determines the average plasticity of the genome of a species, how much that genome changes over time. This rate is under the control, ultimately, of the ecological niche the organism is in. But there is plasticity not only in the genotype of animals but in their phenotypes as well, and this must also be under the control of the ecological niche. In other words, every part of the body, the bones, the blood vessels, the skin, the nerves, all of them have an evolved level of sensitivity to trauma of different sorts. That is phenotypic plasticity. Organisms have the ability to heal from this trauma. In other words, it is a better strategy to be made out of breakable connections than rugged, durable connections in some niches. The more breakable, the more fragile, the nervous system of an animal is, the higher the plasticity not only of the phenotype, but the higher the plasticity of behavior. Behavioral plasticity is about having a highly breakable, but reparable, set of neural connections in large numbers, and that is what we call intelligence.

Second: You make the distinction between what the old brain provides and what the new brain, or neocortex provides. This is a common way of looking at brains, because it distinguishes an anatomical feature that is distinct in humans, which seems to provide the explanation for why humans dominate life on our planet. In this view, the new brain does something specific that makes us superior. This is an utterly reasonable premise. But you described intelligent machines as lacking

motivations entirely unless they are supplied by the designers of the machines? In this way, you propose, they will be purely "new brain," without selfish desires or emotions. At the same time, you describe acquiring knowledge, & thinking, as movement, as navigating through the world building a model of it. I agree wholeheartedly with this idea, to acquire knowledge, causal diagrams, what wiggles what, AGI will have to not only be embodied, but it will have to move, and not only move, but conduct experiments. This is what a child does; she wiggles things to see how they wiggle back. This is true learning.

But wait a moment! How can a machine move without motivations? Objects at rest don't spring into action spontaneously, something or someone has to provide the impetus. If it is the engineers who provide instructions for how to explore the world & what sort of experiments to do, then we are back to current levels of robotic intelligence and AI, not AGI. Let's back up for a second. If a machine must be directed or programmed with particular explanations techniques, it's not intelligent, but if it develops its own, that looks to me a lot like having its own autonomous "motivations."

What supplies the motivations of a living organism? Neo Darwinism & the old/new brain view would say this is simply the drive for genetic replication, & we can and should dispense with this irrelevant mechanism; we should build machines that learn in a disinterested way, building knowledge purely for knowledge's sake. But there is another way to look at it. You could also say that evolution has provided organisms with motivations because they serve a purpose,

they cause the organism to resist its own destruction. Motivations like hunger, thirst, fatigue, and greed provide the impetus for organisms to seek out conditions and nutrients that supply them with homeostasis. Homeostasis is the purpose of the physical arrangement of every living thing. And this arrangement of the living things body is not only set up for physiological reasons but for epistemological reasons as well. All organisms have sensory mechanisms that give them knowledge, knowledge that leads them to find the right conditions + nutrients to support homeostasis. The bat has sonar in order that he might learn where the bugs are at night, + learn to catch them. The bat is motivated to map the cave + the night sky by his metabolism, by not only his cortex but by the demands of his whole body system + what it needs — a roost for hanging, water, bugs, whatever.

From this point of view, you could say that not just the configuration of the neural networks in the bat's brain, but the configuration of his whole body is a sort of guess, a hypothesis about where, ~~[illegible]~~ and when, and how, the conditions and nutrients that will support his homeostasis will be likely to be found in his environment — in his ecological niche. As you pointed out in your book, organisms are sensitive to evidence that they have gotten things wrong, that their conjectures have been refuted. My favorite example is to think of the last time you noticed the skin on the inside of your left pinky knuckle. Now put on gloves; do you notice that spot? No. Now put on gloves with a hole just in that position; the skin suddenly comes alive on the inside of your pinky knuckle! You notice the paints

in your sensorimotor navigation of the world when your conjectures are refuted. This is Karl Popper's model of epistemology in physical, biological form.

I believe that this means that an AGI will not only have to be embodied, but it will have to generate autonomous expectations, and then be motivated to alleviate the points at which its expectations are overturned by experiments with reality. If it expects the coffee cup to be solid but it shatters, this will provide a motivation to learn why. But think about the dataset that will be processed by this machine — the refutations it encounters will be supplied not arbitrarily, in a disinterested way, from "reality" as a whole, but only by the actual readings on its actual sensors, that are pointed not to reality as a whole, but only to a very special subset of all the possible readings that could be taken in the universe, only the ones that are accessible and integratable from the point of view of its own sensors, sensors that are integrated into a unitary physical body. This is exactly the same kind of dataset that we are referring to when we refer to, in the case of a biological organism, the ecological niche of that organism.

Raw niches are not all the same, and they are certainly not interchangeable. If you put my brain into the body of a bat in a cave, none of my stored knowledge would work because knowledge, as you have shown, is a set of sensorimotor configurations unique to each living body. My visual connections would be largely useless in a bat, and my auditory connections would not be able to make sense of the overwhelming noise coming

from my echolocating ears, and they would not appropriately map onto the movement of my arms, which had become wings, and so forth. The knowledge that is configured in my brain comes from the "dataset" that has been supplied by my unique body in its unique surroundings, and can't be transferred readily into another body in a different niche. Any learning, intelligent system, whether organic or synthetic, is motivated by the surprises that it encounters with its sensors, and these form a unitary whole system within its niche. None of these surprises are abstract, they are all concrete; they are all felt in an integrated physical set of connections. The only way to support homeostasis is to correctly anticipate the causal diagrams of the world around us, and the only way to build those causal diagrams is to act as if one is supporting homeostasis. They are one and the same process. Any causal surprises discovered by an intelligent body will supply motivation to learn why they happened (Why did the cup shatter?) This logically has to present itself as indistinguishable from the type of motivations that are supplied by organic bodies — they will look not like hunger + thirst, but they will be self-interested, peculiar, and hard to communicate, just like the whimsical interests + curiousity of children (the most selfish animals!)

This view, although I think it is logically inescapable, presents a problem. If the old brain, new brain, and the body, are integrated into a seamless unit, and the whole unit is engaged in epistemology, not just the neocortex, then what accounts for the unique epistemological success of humans? Why are we so much smarter than any other species?

I believe the reason for this is not cortical columns, exactly, but a specific behavior that our overall body plan makes possible. Claude Shannon discovered that the important thing, if one wants to store and transmit knowledge in the form of information, is that, as he wrote in 1948, "the message be selected from a set of possible messages." In other words, the knowledge must be broken into interchangeable parts, an alphabet, so that it can be reassembled by the recipient upon arrival. If a message is broken into an alphabet and formed into symbols, whether they are letters and characters or 1s and 0s, the recipient can match + map those characters onto the neural structures that they have, but more importantly, the message can be reproduced and transmitted with complete fidelity, which means that it can be improved over time. As messages are reproduced faithfully, they can be criticized and the best ones can be "virally" spread. Without this, messages degrade through a game of "Telephone" quickly. Alphabets are essential to the growth of knowledge, because in order for the sum of knowledge to accumulate exponentially, errors must be corrected in transmission.

We know that the invention of information (alphabetical knowledge) has led to the explosion of learning that humans have experienced, but to understand it it is essential to distinguish carefully between knowledge in general & information. Knowledge in general is a very broad category, including all the configurations of neurons in all the pre-literate peoples' brains in our past, when the only technologies we possessed were songs, spears, and campfires. But alphabets changed all that. A wink and a nod is not alphabetical; showing someone how to do

Something is not alphabetical; a pre-literate language is not alphabetical. All these things — gesture, imitation, languages, — are done by other species, like whales, elephants, wolves, crows, and primates.

We discovered that all life transmits knowledge in alphabetical form during the 20th century, but only in one specific way — in the form of the base pairs of DNA, the alphabet of A, G, T, C that makes up the genome. This alphabet accounts for the massive accumulation of knowledge in the form of living bodies, which "conjecture" about conditions & nutrients in their niches. Humans are the only species that has evolved a much faster way to build knowledge, to transmit external information. This is just because of a simple quirk of our niche. We are a collaborative sight-based hunter. We have good vision & social skills but poor scent detection. So we hunted animals by following tracks in the dirt, and communicating about what "story" those tracks told. It is a short, simple hop from following the "story" of animal tracks by sight to painting alphabetical tracks of an arrow (first in mud tablets) that told a story of an arrow devising. But this made possible, for the first time, the massive accumulation of external information in the ecological niche of a species, which could then become part of the "dataset" that our intelligence is trained on.

In order to gain any knowledge from another intelligent being, an AGI will have to map the knowledge represented in our alphabetical symbols onto a similar set of connections to the ones we have, otherwise the ~~~~ message will not mean anything. A bat could write me a story about echolocation, but since I have

no "reference frames" at all for this knowledge, it won't mean anything to me. This explains why it can be so difficult to communicate with one another, and so much of what we "know" is lost in translation. This means that an AGI will have to conduct similar experiments in a similar niche to our own as it trains itself... it can go beyond the abilities of our own senses, but it must at least include the ranges of sensation in audio, visual, + tactile experience, as you suggest. This is the only way it will be able to participate in all the knowledge we have created in our niche; it is this external knowledge that makes us intelligent. Humans with our same neural system, 10K years back, had only songs, spears, and campfires, and never accumulated any new technologies at all.

There is another surprising consequence of the realization that no learning could occur from a disinterested perspective. It means that morality is not at all what we suppose. We think of morality as a set of disinterested rules or norms that we have developed with our special neocortex that tells us special-purpose guidelines about how to treat other living things, especially humans. But the above argument shows how that simply isn't possible... knowledge can only be grasped, used, and developed from a selfish point of view. So think of morality as very different from technology, even as very different from bio-technology. Technology gives us guidelines that allow us to make use of "little bits of our world to build things that support us in our quest to secure our homeostasis by measuring and mapping

our niche, so that we aren't surprised in a nasty way by things we might encounter. Clothing, buildings, agriculture, medicine, all these can be seen as tools we have developed processing little bits of our niche "dataset" and assembling it into configurations that protect us from nasty surprises that may destroy us. But we can't help but notice that our moral norms and values have evolved along with our other technologies, making gains especially quickly in the past 100 years, and even faster in the last 50 years, just as our technology ~~has~~ itself development has accelerated. This can't be an accident.

What if we think of morality as the same "dataset" of knowledge, but processed in a holistic or integrationist way, from the top down rather than from the bottom up, in a reductionist style? From an integrationistic, global perspective, the most important "materials" in our niche, the matter we need to tend to most carefully in order to protect our own homeostasis, is the other living things around us, especially the other humans. From this perspective, morality is a technology to best make use of other beings to accomplish our goals without being destroyed by them, for having them put us into unpleasant conditions, like prison, or ostracize us from the sources of our satisfaction. From our moral sense, we get a general idea of the kinds of things we might do, the sorts of behavior that would be most advantageous in a general ~~[illegible]~~ way to ensure that we are trusted with resources, like money, car keys, and daughters. These resources are vitally important from a selfish

perspective, while from a disinterested perspective one might as well go to prison... what's the difference? If one is disinterested about one's own body + keeping it in stable conditions with no nasty surprises, one has no reason to seek out the protection + cooperation of other living beings. It is only by looking at the whole set of and causal diagrams at once that we can discern that we will be best served to tell the truth, to not murder, and so forth. From a reductionist, bottom-up view of details in isolation, we might spy lots of examples where we would be tempted to lie, cheat, + steal for personal gain. It's only from viewing life as a whole process with an integrated sense that we can see that our self-ish long-term interests are in doing good, more or less. Again, this dichotomy makes sense of the very different "styles" of cognition shown by the right + left hemispheres.

This also shows why your intuition is right, that AGI will be no more harmful or dangerous than intelligent humans, and perhaps less so. An intelligent machine with superior cognitive abilities will, from this view, be super-moral; it will be a leader in our quest to develop better morality and better technology, because these are two sides of the exact same coin.

Third: I expect that you may get a lot of response for your analysis of "viral false beliefs." This "risk" to humanity has been very much in the news lately, and it is hard to disagree with the idea that people seem to be entranced into believing very foolish things. I have two quibbles with your analysis, though.

First of all, since the content of these vivid foolish ideas is very much on display, we can see and analyze them and determine just how stupid they are. However, what we cannot see and measure quite so easily is the ideas they just replaced, the ideas that these people had about the world just before they adopted the "vivid false beliefs." This is a very interesting problem, because one can suspect, in fact scientific and moral progress suggests it must be the case, that on the whole the ideas that are being supplanted by the "vivid false beliefs" may be, on average and as a whole, even more useless, harmful, and stupid than the new beliefs.

Second, there is a major logical problem with the division of beliefs, or knowledge of any sort, into the simple designations "true" and "false." If it were the case that there were entirely true beliefs and all other beliefs were false, then scientific progress would soon come to an end as the true beliefs accumulated and could be applied to every question on every subject in the universe. It must be a bit more realistic, then, to think that all beliefs must be not simply true or false statements, but rather a mixture of true and false at the same time. All ideas must be on a gradient from "kinda right" to "more right." There is no idea so false that it has absolutely no apparent value for the person who holds it, since it got into their mind through the same process of conjecture + refutation that all living organisms use to build knowledge, as I described above. If this were the case, then nothing known by any human, or any animal,

is entirely false, just as even our best, most sophisticated scientific theories (which, as you acknowledge, disagree with each other) are not entirely true, or else we would have nothing left to discover. And beyond what we know, in science, morality or anything else, is a mystery. We can't predict where the truth will come from, whether from a conventional scientist doing mainstream work, or from an eccentric maverick under the influence of "viral" idiosyncratic ideas. (Even Newton himself was obsessed with numerology and the messages in the Book of Revelation... he was what we would call today a conspiracy theorist... in the closet, of course.

The defense against "viral false beliefs," which you don't specify too clearly, is public testing. This is what the scientific method is all about. But if my guess about the epistemology of organisms is right, then the only important tests are not conducted just in the laboratory or on the Hubble telescope... They are also conducted in the bodies of all animals, especially humans, as they bring forth a world for themselves by trying to resolve contradictions between their conjectures and the results they experience. I think it is important to amend your statement about brains from the idea that they "build a model of the world" to the more humble recognition that we "build a model of a world." Plutarch reported that Heraclitus said, "To those who are awake there is one world in common, but of those who are asleep, each is withdrawn to a private world of his own." I think it is worth going further and acknowledging that there is no firm distinction between experience and hallucination, since the chain of constraint that results in

the activity of a neuron always includes the recursive self-stimulation of the brain itself, as well as stimulation from outside the body in the distal sensorium. I think the nearly-forgotten view of the Santiago theory of cognition makes sense of this, and I think this is the direction that your substantial intellectual work is taking us — the distinction between the "world" and the nervous system exists in the "domain of the observer," not in the cognitive domain of the person doing the thinking.

It was a great pleasure reading your book. I hand wrote this letter because I find that I tend to write more thoughtfully by hand. Although no reply is required or expected, I would be thrilled to get an answer. In case you don't care to do all this scribbling, my email address is: charlie.munford@gmail.com and my website is www.talkingoctopus.com. I have some essays on Medium under my name Charlie-munford.medium.com

Again, thank you for creating such an interesting book, and a stimulating conversation.

All the best to you, sir.

Charlie Munford
1610 Robert E. Lee Blvd. Apt. 407
New Orleans, LA 70124
(504)-717-0884
charlie.munford@gmail.com