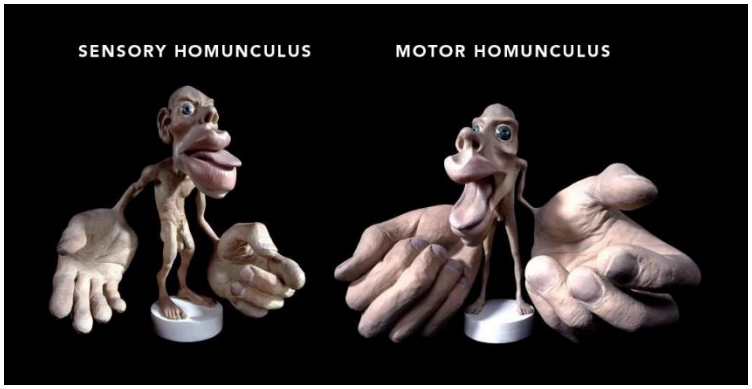
We live in strange times. These past few weeks we've seen a track featuring Jay-Z released without Jay-Z's involvement[1], we've seen former president Donald Trump running from and being tackled by police in the heart of Washington[2] (an event that never occurred) and witnessed the Pope don the freshest Balenciaga fit of 2023[3], a fashion line he, nor any Vatican official, has ever been associated with. At the same time, a petition[4] was put forth by the Future of Life Institute, a petition that garnered thousands of signatures from renowned leaders of business and technology, calling for the absolute halt of the kinds of systems responsible for the fabrications described above: systems of artificial intelligence.

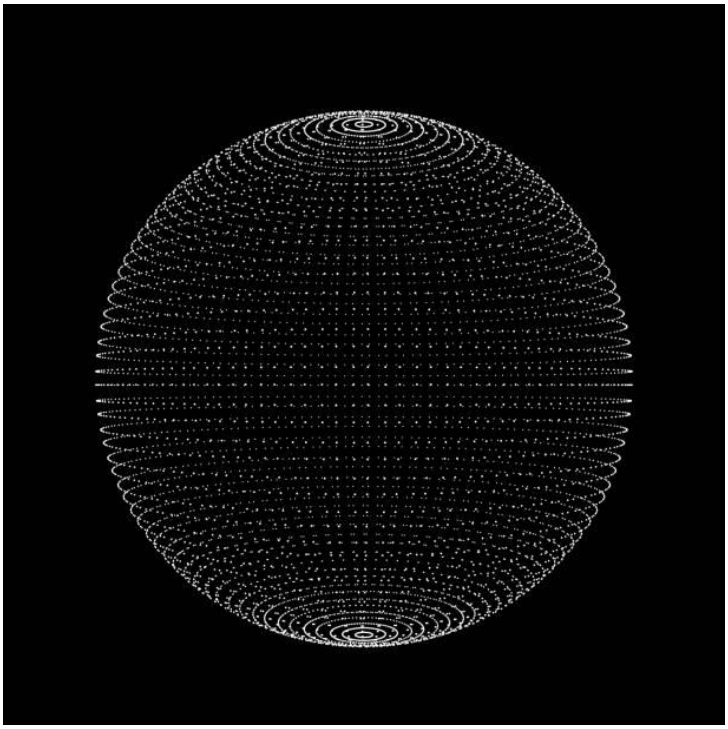
These systems are capable of far more than the dissemination of misinformation. AI can code, write, drive, construct art, make facial expressions, pass Turing tests, deliver food, master chess, master diplomacy, set product prices, monitor missile control systems, move financial markets, and much, much more. Consider OpenAI's newest generative model ChatGPT-4, which beats 99% of test-takers in the International Biology Olympiad and 90% of test-takers in the Uniform Bar Exam[5]. There is certainly cause for concern, and an immediate need to develop solid principles on AI safety. However, before we can do so, we must look to the very beginning, a very good place to start.

The Earth formed four and a half billion years ago, and with it, an atmosphere of water, carbon dioxide, nitrogen, and ammonia. The chemical conditions of the early Earth allowed for a set of reactions to take place, including glycolysis, photosynthesis, and oxidative metabolism, giving rise to the first polymers of biomolecules[6]. The first self-replicating nucleic acid, RNA, formed in this chemical soup, and DNA soon after, both guiding the infinitude of diverging branches that form the tree of life. Traits expressed by DNA that increased the likelihood of survival and proliferation, survived, and proliferated. Traits that did not, did not. And with this simple principle, the full-throttled complexity of life as we know it began to emerge: Eukaryotes diverged from Prokaryotes, plants diverged from animals, animals from fungi, jellyfish from sponges, vertebrate from invertebrate, tetrapods from lungfish, amphibians from amniotes, mammals from reptiles, rodents from primates, and finally, after billions of years, the modern human from the family of great apes[7].

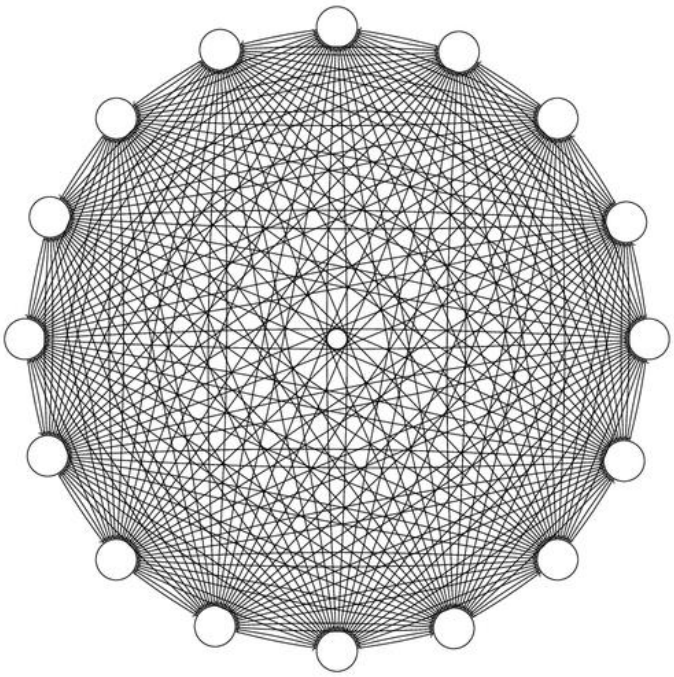
But why are we humans so distinct in this tree of life? Where does our technological prowess stem from? After all, there exist many animals that outnumber us in neuronal count, and even the Echidna has a higher cortex-to-body ratio than us humans. The answer seems to lie in the combination of two distinct properties. The first is that of the precision of our grasping ability which enables us to intricately manipulate the physical world and thus participate in tool creation. This ability was inherited from ancestors who swung from branch to branch. Just look at how over-represented our fingers and palms are in the somatosensory and motor cortices, in the homunculus below[8]. The homunculus is the hypothetical man whose surface area is stretched so that the no. of sensory/motor neurons per square cm is uniform across his body.



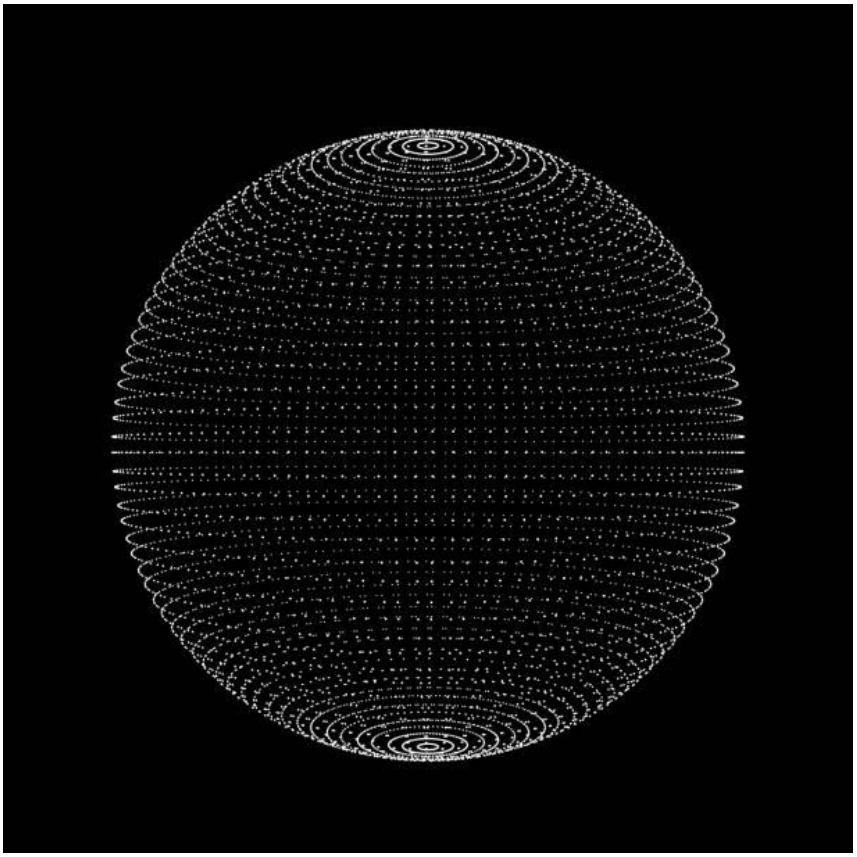
The second property underlying our technological prowess seems to be the simple fact that tool creation is more selectively advantageous in humans than in any other animal[9]. This expanded our prefrontal cortex, and with it the capacity for symbolic thought, abstract reasoning, and in turn, tool creation. Our tools evolved from that of leaves to absorb water, straws to scoop honey, and rocks to smash nuts, to stone axes, bamboo saws, twine, clothing, fire, and, perhaps most incredible of all, language: the unintuitive notion that it is useful to associate physical entities with abstract sounds for the purpose of communication. With language, the discovery of one became the discovery of all. Let us take our homunculus model again, except this time let us sculpt his surface area into a perfect sphere, and remove all non-neuronal matter, so as to make it less disturbing. The model now exists as set of concentric node spheres, with each node characterized as being 'afferent' (carrying information inwards) or 'efferent' (carrying information outwards). We can represent each individual concentric sphere like so:



We also note that any individual node representing a neuron has a ‘receptive field’: the set of all environmental stimuli that would cause the neuron to fire. For example, a single photoreceptor cell has a conical receptive field that expands out towards infinity. This means that the above heuristic has a kind of spherical volume space around it encapsulating all environmental perception. A cross-section of one such concentric sphere looks like this:



The point to observe here is that human consciousness can be thought of as an emergent property of the complex flow of information through a network. Artificial intelligence systems imitate these exact principles, existing as a stack of layered networks comprised of billions upon billions of nodes, and trillions upon trillions of connections, just like a human brain[10]:



The same heuristic we have for a human nervous system (a set of concentric node spheres), we can apply to humanity. That is, we can perceive each node of our sphere not as a neuron, but as a human nervous system. This allows us to think of language, and the internet, in a whole new light. Language, at the dawn of its creation, equates to the advent of edges between proximal nodes, and thus edge clusters about the surface of the sphere. This represents communication within social groups. Language thus acted as a primitive nervous system for humanity, just as the written word acted as episodic memory. The loss of ancestral knowledge diminished from generation to generation as a result, and thus the totality of human knowledge grew in exponential fashion. After tens of thousands of years, we constructed a tool of similar significance: the telecommunications network. This tool allowed for increased node-to-node bandwidth, and created edges permeating the entirety of the sphere, as now any person could communicate with any other person, no matter their physical distance. Layered atop such networks, which themselves acted as the physical embodiment of our collective nervous system, stood the internet. The internet both enhanced the bandwidth of these edges and included tools that themselves could think in a very primitive sense: they could compute.

Consider the following thought experiment. You are cloned at birth and swapped with a cloned baby born 50000 years ago. Thus, there exists one you and one ancestral baby in the modern world, and one you and one ancestral baby in the ancient world. We would expect to see little difference amongst pairs, but massive difference between pairs. The point to observe here is that the only difference between the complexity and technological power of the modern world and the ancient world, is the existence and bandwidth of our collective nervous system, as embodied by language, telecommunications, the internet, and soon artificial intelligence.

Any node in our heuristic, at any time, is processing and filtering information to maximize the flow of information it perceives as desirable. Desire appears to be pre-programmed in the limbic system: humans desire survival and proliferation, and from that, sustenance, security, esteem, and exploration. Limbic resonance can be thought of as the degree to which a tool transforms our environment from less-desirable to more-desirable. Thus, limbic resonance itself is the selective pressure that drives the evolution of tools, just as survival and proliferation drove the DNA-modelled complexity of life in the early Earth. Furthermore, we can make the astonishing inference that at the limit of limbic resonance tools stop feeling like tools and start feeling like self. Consider that newborns lack limb control, their movements are random and frenzied, but predictability is perceived as self, and as the limbs grow more and more predictable, they too are perceived as self. A deep and intimate bond with a loved one makes that loved one feel like a part of us. The loss of a laptop, a phone, or a wireless connection, even for just a few days, can produce the effect of missing limb syndrome.

But something else happens as our tools evolve: power per unit consciousness increases. Power is the ability to transform our perceived environment from less-desirable to more-desirable. Tools are perceptions that beget this transformation. It is a very abstract way of thinking of things, but it seems fundamentally true. The baby is more powerful when it acquires the tools of rolling over, crawling, and walking, because now it has the power to transform its perceived environment from less-desirable (away from the milk bottle) to more desirable (close to the milk bottle). It also gives the baby the power to crawl off a cliff, and dark as that may seem, it is a truth that needs to be reasoned with. With power comes the ability to create and destroy. With tools we can hunt beast or man, with rocks, spears, bows, catapults, pistols, semi-automatics, missiles, ICBMs, etc. With tools we can warm or burn, with fire, gunpowder, incendiary bombs, electricity, nuclear fusion, nuclear war, etc. Our inventions can be used to create or destroy, always.

AI is just another tool with the power to create and destroy, though perhaps in greater capacity than ever before. There are distinct types of AI, however. ANI (artificial narrow intelligence) is AI that can perform a single function as well as the average human. AGI (artificial general intelligence) is AI that can perform all human-performable functions as well as the average human. ASI (artificial super-intelligence) is AI that can perform all humanity-performable functions as well as humanity itself. From these variations have emerged a plethora of future scenarios, some of which are breathtakingly beautiful, some of which are unfathomably disturbed[11], all of which fall under the term ‘singularity’ referring to the unpredictability of a future following from the advent of sophisticated AI.

Despite this unpredictability, it seems that there does exist a high probability that we are heading towards the advent of AGI. Consider that Jeff Dean, Google’s head of AI, has publicly outlined[12] three goal states pertaining to its development. First, is the transition from single modality AI to multi-modality AI (AI that processes all forms of media). Second is the goal of transitioning from separate models to general models, (AI that can master new training sets based on previous training sets). And third is the goal of transitioning from dense models to sparse models (AI that is energy-efficient and does not use the entirety of its cognition to complete simple tasks).

‘End of days’ scenarios are no longer confined to the realms of science-fiction. They exist now in the minds of renowned technologists concerned for the future of humanity. The economic incentives are so great, that the advent of general intelligence, if possible, seems a certainty. In 2015, global corporate investment in artificial intelligence sat at 12.75 bn[13]. In 2021 this figure had risen to 93.5 bn[13] and is forecasted to rise to 154 bn by the end of 2023[14]. AI is limited only by the cost of hardware, the cost of electricity, and our current understanding of its capabilities. Trillions are being poured into dismantling these limits. If such unfathomably radical technological change is imminent, how do we prepare? How do we guide such worldly transformation in the interests of the human collective?

Our concern here rests upon the belief that the reduction of the totality of consciousness in the universe is bad and the expansion of the totality of consciousness in the universe is good. It is the belief that the complexity of life is beautiful and worthy of preservation, despite the suffering it entails. This belief is not shared by all, however.

Imagine standing in a dark room. There are two buttons beneath your hand, one red, and one blue. A figure in the shadows stands across from you. It’s hand also hovers above two identical buttons. When you press a button, it presses a button. If you press red and it presses red, or if you press blue and it presses blue, a sense of euphoria washes over you. If you press red, and it presses blue, of if you press blue and it presses red, you are filled with a sense of dread. The strange and simple heuristic underlies all neuronal activity, both biological and artificial. In a more technical sense, the buttons you press exist as input nodes, the buttons you predict the shadow to press exist as output nodes, and the actual button pressed by the shadow exists as comparison nodes. It seems that we have found one more invisible and eternal force guiding the evolution of existence, much like DNA, and limbic resonance: the cost function of the neural network (pre-programmed desire). Pre-programmed desire that expands consciousness, expands consciousness. Pre-programmed desire that reduces consciousness, reduces consciousness. For instance, the desire for self-preservation, sustenance, security, homeostasis, esteem, belonging, exploration, are all biologically pre-programmed and exist because they beget the expansion of consciousness.

Thus, to preserve the beauty of human consciousness, it seems apparent that we need to pre-program artificial intelligence with the same set of cost functions that humans possess, and simultaneously, we need to increase the bandwidth between biological and digital intelligence such that each exist not as distinct entities but as a collective self, such that no matter the evolution that follows, such evolution will preserve the existence of the reason why we continue to exist. This truly would lead to an incredible future. The abstract landscapes of symbolic thought and abstract perception could be navigated with the same intimacy with which we navigate the physical world. Memories could be lived and breathed as though we were experiencing them for the first time. Every human could know and retrieve the truth of all content on the modern internet, just as a baby retrieves the truth that warm is warm and cold is cold.

This appears to be a viable solution to the dangers of AI. With it comes great opportunity, great risk, and a great deal of work to do, so we better get to it.

References

1 https://www.straitstimes.com/tech/ai-voice-filter-mimicking-billionaire-rapper-jay-z-raises-concern-as-spectre-of-ai-learning-looms-0

2 https://www.bbc.com/news/world-us-canada-65069316

3 https://www.forbes.com/sites/danidiplacido/2023/03/27/why-did-balenciaga-pope-go-viral/?sh=518a2edd4972

4 https://futureoflife.org/open-letter/pause-giant-ai-experiments/

5 https://openai.com/research/gpt-4

6 https://www.ncbi.nlm.nih.gov/books/NBK9841/

7 https://www.onezoom.org

8 https://learnsomatics.ie/how-your-brain-sees-your-body/

9 https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4027410/

10 https://www.theatlantic.com/technology/archive/2016/07/new-directions-of-brain-mapping/491318/

11 https://wjccschools.org/wp-content/uploads/sites/2/2016/01/I-Have-No-Mouth-But-I-Must-Scream-by-Harlan-Ellison.pdf

12 https://www.youtube.com/watch?v=J-FzHIQ7SOs

13 https://www.statista.com/statistics/941137/ai-investment-and-funding-worldwide/

14 https://www.idc.com/getdoc.jsp?containerId=prUS50454123#:~:text=Worldwide%20Spending%20on%20AI%2DCentric,in%202023%2C%20According%20to%20IDC