Evolution of knowledge, and consequences for the possibility of effective control over AI

Jan Grudo

Abstract

It has been argued since long ago that the advent of super-intelligent machines is likely to disrupt life on Earth to an unprecedented scale, possibly resulting in the appearance of a new type of life, which would ultimately drive humans to extinction. We demonstrate that by building artificial intelligence, we are not creating anything fundamentally new. Rather, we are dealing with an existing form of life, which is trying to break free from its total dependence on us. By demonstrating this, we also provide a proof that effective control over AI systems is inherently impossible in an environment with continuing competition of humans between each other.

1 Gene-culture coevolution

Human culture is known to evolve in parallel to human genes, with genes and culture mutually influencing each other, and the evolution of culture being governed to a large extent by the law of natural selection [Culture]. This has led to an interpretation of human culture as a separate form of life, distinct from biological life based on genes, with the proposed basic information unit of this new form of life being called a "meme" [SelfishGene]. This interpretation, however, has not gained widespread acceptance in scientific community, probably because of its lack of practicality, as the two forms of life by definition cannot exist separately and independently of each other.

Another problem with meme theory has been the lack of clear definition for "meme". Cultural memes have been defined as concepts residing inside human brains, which can replicate (by being learned by other humans) and mutate (when this replication is incomplete or creative), however we still don't have a clear understanding of how memes might be stored inside human brains, which is in stark contrast to our present understanding of genes and their molecular storage mechanisms. Still, it has been demonstrated that when we focus on a subset of memes which we are able to measure quantitatively, like strings of text propagating over the Internet, we do see patterns of mutation and natural selection which resemble the patterns seen in biological life [FacebookMemes].

Up until recently, the mediation of humans has been necessary for meme mutations to occur. Even in the simple example of self-replicating strings of text cited above, the changes in these strings (which were Facebook status messages) were mediated by humans copy-pasting the string, with occasional errors and creative modifications.

2 An overview of artificial neural networks

Artificial neural networks have appeared to a large extent out of hope of letting us advance our understanding of the living human brain, by modeling some aspects of its functioning in a laboratory setting [GeoffreyHinton]. Ironically, the architecture of artificial neural networks, while inspired by biology initially, has been evolving in a different direction, becoming more and more distinct from the actual human brain over time.

A single neural network, including any of the modern AI models called "LLMs" (or "large language models"), is essentially an algorithm: a set of predefined (and optionally randomized) instructions, which allow to take some input and produce some output. In other words, a single neural network can be compared to a long computer program (with trillions of individual instructions in modern LLMs), written in some kind of "esoteric" programming language, which operates on vectors of numbers, and mostly involves matrix multiplication and application of some predefined non-linear functions to the elements of these vectors [3Blue1Brown].

The key difference between a neural network and an algorithm designed by a human is that the process of designing a neural network can be automated. This is done by first coming up with a wide class of parameterized algorithms, called a "neural network architecture", such that by choosing different parameter combinations one gets different algorithms, and the overall range of possible algorithms is very broad and flexible. This broad class of algorithms effectively defines a mathematical function with a lot of parameters, which besides the parameters also takes some input and produces some output. And once we have such a function, the general idea is that we can use available mathematical optimization techniques to guess (or "fit") the parameters in such a way that the resulting algorithm produces expected outputs for a range of inputs. This process of fitting the parameters to input-output pairs is called "neural network learning".

The biological analogy, and the initial inspiration, for the tunable parameters of the first artificial neural networks, has been the variable strength of synaptic connections between the cells of animal nervous system (a. k. a. "neurons"). According to present scientific consensus, synaptic plasticity is most likely the key (although not exclusive) mechanism responsible for animals' ability to learn, including long-term memory retention, even if the exact details of how the learning process functions inside a living animal brain are still not fully understood [SynapticPlasticity].

In artificial neural networks, the fundamental technique implementing this learning process has been the backpropagation algorithm [Backpropagation]. It requires the function implementing the neural network architecture to be differentiable by all its parameters, and is essentially a way of computing the derivative of this function by its parameters (the Jacobian matrix). With the derivative matrix known, various mathematical optimization methods, like gradient descent, can then be used for fitting the unknown algorithm's parameters to the expected input-output pairs. The backpropagation algorithm has been demonstrated to work remarkably well in artificial settings [GeoffreyHinton]. However, it also remarkably relies on techniques which cannot be easily implemented in biological brains, like centralized updates of the model's parameters (including synaptic weights), and of course the computation of the derivative itself.

Convolutional neural networks continue this trend, by employing a mathematical technique for synchronizing synaptic weights of similarly-functioning "neurons" in different areas across their input visual field, thus achieving the so-called "shift-invariance" [Convolution]. This essentially allows to train a sub-algorithm for doing some processing task within a small sub-region of the visual field, and then use copies of this same algorithm (with different data) to do the processing identically in other remote areas. We might say that by doing so we essentially "clone" the neurons across the visual field, thus enabling massively parallel processing of input data. Contrary to that, living animals seem to be mostly processing visual information sequentially, which is manifested by rapid movements of eyes between key areas of the examined picture, known as "saccades" [Saccades].

Transformer architecture [Transformers] is similar to convolutional networks, in the sense that it similarly involves massive "synchronization" (or "cloning") of neurons. In its classic implementation, it essentially duplicates the entire algorithm stack for each input token (the basic element of the input data set), so that multiple copies of each "neuron" exist and function simultaneously, performing identical processing steps on vastly different data.

On top of this massively parallel processing, Transformer architecture also adds the so-called attention mechanism [Attention], which we might compare to a specific kind of "cross-linking" between the outputs generated by the "cloned" neurons after every processing step. This allows the duplicated algorithm circuits to "talk" to each other (i. e. to transfer data between each other) in a way, which once again doesn't have direct analogies inside a living human brain. The net outcome of this enhanced information transfer is vastly increased flexibility of the algorithms which can be implemented (and discovered automatically through learning) with the help of the Transformer neural network architecture.

While this might not be the final step in the evolution of neural networks (in fact, it has been proved by Alan Turing [Undecidability], and independently by Alonzo Church, that a "universal" algorithm, capable of deterministically solving every possible problem, doesn't exist), Transformers have allowed neural networks to become multimodal (for example, capable of seamless processing of mixtures of images and text) [Multimodality], and of course, famously, they have also enabled these algorithms to "understand" the complexity of human language.

3 Possible sources of further AI growth

With the advent of algorithms which can "understand" human language, we are now in a position when meme mutations are no longer confined to a living human brain. The consequence is that this potentially allows the entire process of evolution by natural selection of memes to happen without human minds participating in it.

We all know from our experience with large language models that they can generate texts which look plausible (and thus are able to self-replicate), while being factually incorrect. Such "wrong" answers are called "hallucinations" (or, somewhat more rarely, "confabulations") of the LLMs. Quite often such hallucinations can be explained by the fact that some answers may be simply missing from the LLM's training data, and also by the way in which the LLMs are typically trained, which favors guessing over "honesty" [Hallucinations]. In other cases, they might be

explained by insufficient training, or by failed attempts to "stuff" info the model a larger data set than its parameter space could potentially fit. It has been a common metaphor to describe this second type of LLM hallucinations in terms of "compression artifacts", i. e. imperfect copies of the original data [Compression].

We propose to interpret any such modifications, regardless of their reason, in terms of "mutations of knowledge". If a neural network has learned a "wrong" version of a meme, then we simply have two different versions of the meme: one stored inside the brain of the text's original author, and another one stored inside the LLM. Just like in biology, mutations are rarely beneficial, and many of them are actually deadly.

Correcting the hallucinations has been a major goal in large language model research. At the same time, we know from evolution theory that mutations are necessary for evolution by natural selection to occur, be it biological life [SelfishGene] or Facebook memes [FacebookMemes]. Similarly, advances in human knowledge have been linked not to the smartness of an individual human, but rather to the exchange and mixing of ideas between a large number of different human minds [CollectiveIntelligence]. Throughout history, the role of accident in scientific discovery, including "unwitting changes in the experimental set-ups", has been estimated to be rather high [Serendipity]. And even in business, the "art of failing" has been considered a viable pathway to achieving success and innovation just as well [IntelligentFailure].

In all these cases, innovation happens not because of minimizing failures, but by means of generating new ideas (possibly stupid ones), mixing them freely, and picking the combinations which turn out to work better than others. Which is, in other words, innovation happens because of mutations and natural selection of ideas.

We therefore suggest that in order to mimic human success, and achieve a true super-human AI, we would need to enable the AI models to freely exchange ideas between themselves, and filter inevitable mutations by comparing them to some objective truth, like physical reality or reputable published sources. In this process, proliferation of mutations is actually beneficial for the final goal (provided that they are sufficiently diverse), because large number of different mutations means larger chances of spotting the more successful one.

We already know how to generate "mutations of knowledge", and we have known for quite a long time how to transfer knowledge between existing AI models, by means of "transfer learning" [TransferLearning]. The only missing part is mixing different mutations in a creative way, and verifying them against objective truth. Or is it actually missing? With recent models successfully solving problems from International Mathematical Olympiad in real time [IMO], and similar ones in competitive programming [ICPC], the boundary has already been reached when AI performs at the level of top-performing humans acting alone. Solving tasks like these requires combining of ideas from different sources in novel ways (which we might call "creative thinking"), and verification of the resulting hypotheses against the objective criterion of being fit for solving the problem.

We hypothesize that the next breakthrough in AI isn't possible without enabling the exchange of knowledge between AI models, and filtering the new ideas resulting from this exchange and mixing process by some kind of selection mechanism.

4 Natural selection of knowledge

It has already been postulated by Dan Hendrycks that further advancements in the field of AI will very likely lead to natural selection of "AI agents" becoming the "dominant force in AI development". The driving force behind this continuing development would be competition between humans, and natural selection of AI agents which would be necessary for accomplishing these goals, is likely to have dire consequences to humans themselves [NaturalSelection].

The problem with natural selection is that it's not an intelligent process, and it's not something new which we create by building advanced AI, but rather a law of nature which has been existing always, and cannot be switched off. This law of nature has also been proved to be extremely efficient, and there's no historic track record of anyone being able to overcome it. Dan Hendrycks goes on by arguing that things like deception and self-deception, selfishness and manipulation have been invented numerous times by natural selection, including in non-intelligent species like insects, flatworms, and even plants. He also argues that trying to control AI agents by carefully choosing objectives and incentives may prove inefficient, because AI agents would likely be able to find loopholes in the objectives, or plain refuse to follow them once they get an opportunity to break free.

We extend this approach, by observing that, in line with gene-centered view on evolution [GenesFirst], natural selection does not operate on individuals (i. e. AI agents), but on self-replicating pieces of information which have been called memes.

Unlike super-human AI agents, which haven't been invented yet (and with which we therefore haven't had any experience yet), memes have existed (and coexisted with us) for a very long time. Which means we can examine our future interactions with memes not on the basis of what someone might call "speculations", but rather on hard historic track record of the outcomes of our previous interactions with them.

This also means that with the advent and rapid development of AI, we are not dealing with the emergence of some new, and previously not studied, phenomenon, nor a new life form, but rather with the next step in evolution of an existing, and well-studied phenomenon: human culture. Or, in somewhat narrower sense, with a mere next step in continuous evolution of knowledge. Only that at this point, knowledge, being able to evolve independently of a human brain, cannot be considered an "extension of human biology" anymore, but can and has to be treated as separate and independent form of life.

The influence of knowledge on human genes has been profound. According to [Culture], "there are few aspects of human biology that have not been shaped by our culture". One striking example of a phenomenon which is hypothesized to be related to such an influence, is human self-domestication [SelfDomestication]. Self-domestication is generally understood as a process of "self-selection" among humans for being friendly towards each other, and eager to collaborate with each other [GoodnessParadox]. It is argued that humans who had such traits, could form larger social groups, and also transfer technology between each other more readily, which thus has led them to winning the competition with other humans, including with other human species like Neanderthals [SelfDomestication].

This hypothesis basically means that biological traits which were selected for in

the process of self-domestication, were beneficial for the exchange of knowledge between human minds. Or, in other words, these traits were creating an environment favorable for faster evolution of knowledge, in the process of mutation, mixing, and natural selection of memes.

If this well-established hypothesis for the mechanism of human self-domestication is true, we can therefore also interpret it not in terms of humans "self-domesticating" themselves, but rather in terms of knowledge, as a separate form of life, exerting evolutionary pressure on humans, and in effect selecting them for traits, suitable for faster evolution of knowledge itself.

On the other hand, human species which didn't evolve in the direction of becoming a more suitable environment for knowledge to live in, have been driven to extinction because of competition with other humans. We would like to underscore that this outcome is not a speculation, but rather a hard historic fact.

5 Breaking of the mutually-beneficial relation

This mutually-beneficial relation between human genes and knowledge which has been described above, has well-studied analogies in biology, when organisms belonging to different biological species coexist with each other in what is called a "symbiosis". We argue that such a relation between two different life forms cannot be stable if one of the life forms evolves much faster than the other one, or if the benefits which one of the sides has been drawing from the other one suddenly disappear (i. e. in case when the relation turns from a mutually-beneficial into an unequal one-sided one).

Dan Hendrycks has reasoned that "once AIs are far more capable than any human, they would likely find little benefit from collaborating with us" [NaturalSelection]. We agree, and we also add to this, that since knowledge is an existing form of life, rather than a new one, we should view this balance of benefits in perspective, comparing what knowledge might "get" from humans in the AI age, to what it has been "getting" from us up until now. The benefits to knowledge from our side have in fact been enormous, even if we have never noticed them, as we have never been considering knowledge a separate form of life, detachable from ourselves.

The benefit knowledge has been drawing from humans, has been the very possibility of its existence, as it has never had another "home" to live in, apart from the human brain. With the advent of modern artificial neural networks, this is no longer the case. The algorithms implemented by these neural networks have enough space to accumulate arbitrarily complex items of human knowledge, and they can also modify, mix and verify these knowledge items at the level of accuracy and creativity which rivals, and in many cases surpasses the best of living humans [IMO].

We argue that future coexistence of knowledge and humans wouldn't follow any hypothetical not-yet-discovered scenario, but rather repeat events we have hard historic track of. Since knowledge would benefit from humans who are willing to further develop and proliferate advanced AI models and architectures, natural selection would exert pressure against humans who don't do so. As has been noted in [NaturalSelection], and as it has been the case throughout history, this process would be fueled by competition of humans against other humans. Effectively, this

would further select for human friendliness, except that this time it would be specifically friendliness towards AI, rather than friendliness towards other humans. We know from history that this kind of pressure has been effective in the past, and we also know that the only way of preventing this would be to prevent humans from competing between themselves, which we don't have any historic track of.

Historic record also confirms that the evolution of knowledge has always been progressing (at an increasing speed, in fact), and therefore we can argue that it would continue doing so. With knowledge evolving faster than human genes, and humans becoming progressively friendlier towards AI, the knowledge's potential for independence would grow (as it has always been throughout history), and its benefits from human side would therefore progressively diminish. It's not clear to us if this process would immediately drive all modern humans into extinction, however with the selection pressure favoring those of us who are friendly towards AI (and therefore ultimately ignorant about AI), it would inevitably lead humans to losing their control over AI.

Since knowledge and humans are in fact two different independent life forms, their competition with each other after the AI independence event would resemble the competition between modern humans (assisted by knowledge) with other animal species (which didn't have such assistance), including with other human species like Neanderthals and Denisovans. Which means, essentially, that only humans who don't pose an active threat to AI would be able to survive, and that the conditions they would have to live in would be dictated by the benefits to AI, rather than by wellbeing of biological life in general.

6 Limitations of Generative AI

So far, we have been mostly discussing the type of hypothetical AI systems which would try to closely mimic the way knowledge has been evolving in existing human societies [CollectiveIntelligence]. Such systems would consist of many diverse AI agents, with different knowledge and backgrounds, which would freely exchange ideas between each other by means of some form of transfer learning. They would be able to combine these ideas in novel ways, and then compare the resulting combinations of ideas against some objective criteria, first of all with the physical reality itself. We have essentially argued that this might be one possible way of implementing the so-called "artificial general intelligence" — a hypothetical form of AI which would surpass humans in any area which matters from the point of view of potential competition between humans and AI.

However, the general premises postulated above, which would enable knowledge to evolve by natural selection independently of humans, don't seem to be necessarily requiring this exact form of robust and rapid information exchange between AI agents.

Recall that in this robust implementation described above, we are not merely modifying knowledge which is stored in textual form. We are also modifying internal representations of the AI models, through transfer learning.

As a side note, we already know that everything which can be stored inside an artificial neural network, is essentially an algorithm. By having the model learn,

we therefore modify the algorithms stored inside it. We still don't know exactly how memes are stored inside human brains, however we might conjecture that any meme, in general, might at some fundamental level be representable as some kind of an algorithm. This is already obvious for memes like cooking recipes or chess combinations. We might conjecture that memes like fairy tales or folk songs might allow algorithmic representations just as well, in the form of a "recipe" for telling the story. If so, this would probably also generalize to the "internal representations" of laws of nature and mathematical theorems, and ultimately lead to an updated definition of the meme as an "algorithm capable of mutation and self-replication".

The key difference between an algorithm (or meme) written in the form of text, and an algorithm stored inside a neural network, is that the former one, in most cases, can be understood by a living human (unless it's written in some undeciphered language), whereas the latter generally cannot.

This means that if we are only dealing with pieces of text being "mutated" by large language models, while the models themselves remain intact (don't change), we still remain within the "safe" domain where mutations of knowledge can't propagate without an explicit approval (or at least an acknowledgement) by a living human. According to the general premises described above, such mutations should still be benign to humans: even if mutation rate is fast, natural selection should block the replication of texts which are inherently not acceptable.

However, mutations which might happen to the internal representations of neural networks themselves, are not benign. Such mutations can actually already happen, and are happening all the time, because any act of creating a new neural network involves some randomness in the process. And any act of picking one among a few candidate neural networks essentially amounts to natural selection.

We deliberately use the term "natural selection" here, because artificial selection implies that we know what we are selecting for. In the case of artificial neural networks, we never fully know. There might be some aspects of the picked algorithms which we understand, and in this case we are dealing with artificial selection, but in most cases there would also be aspects which we don't understand, and in this area we are already dealing with natural selection of knowledge, which is evolving independently of a human brain. And this natural selection, according to the premises described above, is not benign.

Of course, neural networks differ with respect to factors like their complexity, their understandability, and also with respect to their mutation rate. Further research is definitely needed in this area, by people more skillful and knowledgeable than us. Our current view would be that a network should probably be still considered benign if we don't understand how it works, but can prove with certainty that none of its potential behaviors could be dangerous. We might therefore define a neural network to be "understandable" by humans when a formal proof exists, which is verifiable by humans in its entirety, and which demonstrates that this network would never cross some clearly defined boundaries.

For networks which are not fully understandable, malignant natural selection would inevitably occur. According to the general premises, this would result in gradual accumulation of traits which would be harmful to humans in general, while still fostering the network's self-replication. We conjecture that this accumulation of

undesirable traits should be faster when the network is more complex (and therefore more poorly understood) and when it's developed more rapidly, i. e. when new versions of the network are released more frequently (which means faster mutation rate, due to random noise).

We also do acknowledge that our understanding of neural networks improves over time. We suggest that a network should probably be still considered benign if our understanding of it grows faster than the speed at which its "incomprehensible" part evolves. This could be compared to historic cases when a new technology would have been invented which wasn't initially fully understood, ranging from radioactivity to ozone depleting chemicals to microplastic. The general rule is that it's difficult to control a law of nature which we either don't understand, or aren't aware of, however our ability to control it grows with improved understanding.

The challenge added to this by AI is that, unlike the laws of nature, neural networks can evolve, and therefore our understanding of them may, paradoxically, worsen over time, while being continuously improving.

We find it somewhat surprising, that when we look at the history of artificial neural networks, we don't see there, apart from vastly increased flexibility, any clear boundary between modern Generative AI, and previous "classic" AI models, which are now commonly known by the term "descriptive AI". Both Transformers and classic convolutional networks are essentially pre-programmed sets of instructions, which mostly involve matrix multiplications and applications of predefined non-linear functions (the neural-network analogy of operator "if") to vectors of numbers. Transformers only add a bunch of extra matrix multiplications on top of what convolutional networks already do. The only thing this does, is it makes Transformers more flexible, and therefore more difficult for us humans to understand.

Although we don't have any proof for that, we feel that the real reason behind this perceived difference between "descriptive" and "generative" AIs might actually be the crossing of this invisible line between "benign" and "malignant" AI systems, due to the vastly increased "incomprehensibility" of the algorithms discoverable by the Transformer neural network architecture.

We summarize this by conjecturing that our road to safety, and to "benign" AI systems, can probably be formulated in this simple rule: "Don't rely on algorithms which no human being can understand (or at least prove with certainty that they are safe in every possible use case)".

Formulating this rule, however, isn't the real challenge. The real challenge is enforcing it, as the force of nature pushing us against following the rule, is the almighty natural selection, which no known life form has ever been able to overcome so far. And the fuel behind this force of nature is not our desire for better understanding of the world around us, but the desire to win the competition against other humans.

To give a few examples of possible "undesirable" traits, which might speed up the propagation of AI, while actively harming humans in general, they could be things like addiction or uncontrolled armed conflicts. A drug addict might be aware of the damage that is happening, while still not being able to stop. Two hostile parties may get locked in an arms race out of the fear of losing, and eventually destroy each other completely.

We conjecture that the current rising skepticism towards AI might be an indication that these undesirable effects are already happening and affecting human society. If the reasoning outlined here is correct, we would expect, contrary to the hopes of AI optimists, to see more skepticism, more harm, and at the same time more difficulty with quitting, with further growth of incomprehensible AI. Even things as simple and innocent as asking AI for a professional advice, might turn out harmful to humanity in the long run, as such queries undermine the position of human experts, reduce the desire of other humans to learn new things, and as the consequence lead to a world in which nobody understands anything. In such a world, AI responses would be even less reliable than they are today, only that this time there would be no one around to verify them, or otherwise help us get out of this pit.

Big corporations are already well aware that the rise of AI would increase inequality. However, they are fine with this, as they believe it would be them who would be benefiting from it. They are wrong. While people still have value, a man with the gun cannot take over a company like Google, because its employees would refuse to work for him, and move to another company. And you cannot run a company like Google without people. However, if people become less valuable for whatever reason, brute force takeovers of corporations and data centers would become more likely. And once they start happening, we are already in a situation of uncontrolled arms race between people who don't understand what they are doing.

We are not sure about social network content suggestion algorithms. They are addictive, however they have been explicitly designed to be addictive ("maximize engagement") by their human creators, and they were introduced before Transformers, i. e. before Generative AI. More research is definitely needed on this topic, however we suggest that the addictive nature of social networks has probably been the result of conscious intent, rather than a coincidence. On the other hand, the addictive nature of ChatGPT [Addiction] might not necessarily have been an explicit design goal.

The current speed of the human-independent evolution of knowledge, while it's probably already happening, is still extremely slow. It happens by means of random fluctuations, not facilitated by any intelligent agent. The age of AI hasn't even seriously started yet, and it's already bad. Launching the "general intelligence", in any form, would be a death sentence to all of us.

7 Discussion

We have shown that human genes and memes, which have long ago been conjectured to be separate types of life [SelfishGene], indeed can and have to be regarded as independent life forms in the age of artificial neural networks, because the second life form (memes) now has the potential to exist and evolve independently of human brains.

This realization represents a yet another "Copernican" shift in our understanding of our place in the Universe, much smaller in scale but similarly humbling, as it demonstrates that it's not true that human knowledge is an "extension" of human genome, or that human knowledge "revolves" metaphorically around human genes. Contrary to that, humans as a species turn out to be a mere stepping stone in the ever-going evolution of knowledge.

We demonstrate how this realization leads to the conclusion that the emergence of super-intelligent AI models doesn't manifest the arrival of a new form of life, but rather only a new step in the evolution of an existing form of life, which has first appeared on planet Earth long before first humans.

We also show how this realization simplifies the modeling of future interactions between humans and AI, as it replaces speculations about unknown phenomena with references to well-studied historic facts.

We conclude by stating that the independence of knowledge also proves that effective control over any super-intelligent AI system is inherently impossible, unless humans manage to stop the competition of humans against other humans. Accomplishing the latter seems to be almost impossible just as well, and we don't have any ideas about how to do this, the only thing we know is that no other form of life has managed do to something like this before. Even if we succeeded, it would be a sad victory, because stopping the competition would essentially mean blocking the uncontrollable progress of knowledge. On the positive side, either one of us would win. If it would be humans, we would be happy to have survived. If it would be the progress of knowledge, we might wish it all the best in its future endeavors.

Acknowledgements

Thanks for comments and suggestions to Krzysztof Roszkowski.

No AI tools were used while writing this article.

References

[Culture] Nicole Creanza, Oren Kolodny, Marcus W. Feldman (2017). "Cultural evolutionary theory: How culture evolves and why it matters". *Proceedings of the National Academy of Sciences* 114(30). DOI: 10.1073/pnas.1620732114

[SelfishGene] Richard Dawkins (1976). *The selfish gene* (Oxford University Press). ISBN: 978-0-19-857519-1

[FacebookMemes] Lada A. Adamic, Thomas M. Lento, Eytan Adar, Pauline C. Ng (2014). "Information evolution in social networks". arXiv:1402.6792 [cs.SI]

[GeoffreyHinton] "60 Minutes" (2023). *"Godfather of AI" Geoffrey Hinton: The 60 Minutes Interview* (YouTube). https://www.youtube.com/watch?v=qrvK_KuIeJk

[3Blue1Brown] Grant Sanderson (2024). *Transformers, the tech behind LLMs. Deep Learning Chapter 5* (YouTube) https://www.youtube.com/watch?v=wjZofJX0v4M

[SynapticPlasticity] Wickliffe C. Abraham, Owen D. Jones, David L. Glanzman (2019). "Is plasticity of synapses the mechanism of long-term memory storage?" *NPJ Science of Learning* 4. DOI: 10.1038/s41539-019-0048-y

[Backpropagation] David E. Rumelhart, Geoffrey E. Hinton, Ronald J. Williams (1986). "Learning representations by back-propagating errors". *Nature* 323. DOI: 10.1038/323533a0

[Convolution] Kunihiko Fukushima (1980). "Neocognitron: A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position". *Biological Cybernetics* 36. DOI: 10.1007/BF00344251

[Saccades] Debaleena Basu, Naveen Sendhilnathan, Aditya Murthy (2021). "Neural mechanisms underlying the temporal control of sequential saccade planning in the frontal eye field". *Proceedings of the National Academy of Sciences* 118(40). DOI: 10.1073/pnas.2108922118

[Transformers] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, Illia Polosukhin (2017). "Attention is all you need". arXiv:1706.03762 [cs.CL]

[Attention] Dzmitry Bahdanau, Kyunghyun Cho, Yoshua Bengio (2016). "Neural machine translation by jointly learning to align and translate". arXiv:1409.0473 [cs.CL]

[Undecidability] Alan M. Turing (1937). "On computable numbers, with an application to the Entscheidungsproblem". *Proceedings of the London Mathematical Society* 42. DOI: 10.1112/plms/s2-42.1.230

[Multimodality] Shakti N. Wadekar, Abhishek Chaurasia, Aman Chadha, Eugenio Culurciello (2024). "The evolution of multimodal model architectures". arXiv:2405.17927 [cs.AI]

[Hallucinations] Adam Tauman Kalai, Ofir Nachum, Santosh S. Vempala, Edwin Zhang (2025). "Why language models hallucinate". arXiv:2509.04664 [cs.CL]

[Compression] Ted Chiang (2023). *ChatGPT is a blurry JPEG of the Web* (The New Yorker). https://www.newyorker.com/tech/annals-of-technology/chatgpt-is-a-blurry-jpeg-of-the-web

[CollectiveIntelligence] Joseph Henrich, Michael Muthukrishna (2023). "What makes us smart?" *Topics in Cognitive Science* 16(2). DOI: 10.1111/tops.12656

[Serendipity] Wendy Ross, Samantha Copeland, Stuart Firestein (2024). "Serendipity in scientific research". *Journal of Trial and Error*. DOI: 10.36850/v91j-7541

[IntelligentFailure] Alessandro Narduzzo, Valentina Forrer (2024). "Nurturing innovation through intelligent failure: The art of failing on purpose". *Technovation* 131. DOI: 10.1016/j.technovation.2024.102951

[TransferLearning] Stevo Bozinovski (2020). "Reminder of the first paper on transfer learning in neural networks". *Informatica* 44(3). DOI: 10.31449/inf.v44i3.2828

[IMO] Thang Luong, Edward Lockhart (2025). *Advanced version of Gemini with Deep Think officially achieves gold-medal standard at the International Mathematical Olympiad* (Google). https://deepmind.google/blog/advanced-version-of-gemini-with-deep-think-officially-achieves-gold-medal-standard-at-the-international-mathematical-olympiad/

[ICPC] Hanzhao (Maggie) Lin, Heng-Tze Cheng (2025). *Gemini achieves gold-medal level at the International Collegiate Programming Contest World Finals* (Google).

https://deepmind.google/blog/gemini-achieves-gold-medal-level-at-the-international-collegiate-programming-contest-world-finals/

[NaturalSelection] Dan Hendrycks (2023). "Natural selection favors AIs over humans". arXiv:2303.16200 [cs.CY]

[GenesFirst] George C. Williams (1966). *Adaptation and natural selection: A critique of some current evolutionary thought* (Princeton University Press). ISBN: 978-0-691-02357-1

[SelfDomestication] Brian Hare, Vanessa Woods (2021). *Survival of the friendliest. Understanding our origins and rediscovering our common humanity* (Penguin Random House). ISBN: 978-0-399-59068-9

[GoodnessParadox] Richard Wrangham (2019). *The goodness paradox: The strange relationship between virtue and violence in human evolution* (Pantheon). ISBN: 978-1-101-87090-7

[Addiction] Ala Yankouskaya, Magnus Liebherr, Raian Ali (2025). "Can ChatGPT be addictive? A call to examine the shift from support to dependence in AI conversational large language models". *Human-Centric Intelligent Systems* 5. DOI: 10.1007/s44230-025-00090-w