Epiphanies About AI, Intelligence and Cognition

Igor Ševo · January 29th, 2025

> *This article is a condensed summary of insights outlined in my treatise on Engineering Intelligence, Minds and Cognition, a 100-page document about discoveries, insights and epiphanies about our path towards AGI and building intelligent systems in general, which I produced as part of my role as Head of Artificial Intelligence in HTEC Group. While the insights in the document come exclusively from my own research, experimentation and years of pondering the phenomenon of intelligence, HTEC kindly provided me with the necessary space, infrastructure and the design team to support the creation of this document. No part of this article or the attached document has been generated, transformed or corrected by AI.*

Automation—the replacement of human labor with machine labor—has been a continuous, gradual and accelerating process for perhaps the past three million years, starting from the first moment one of our ancestors picked up a piece of flint or basalt to reshape a plant into something edible or otherwise useful to today's attempts at replacing highly sophisticated digital work with the work of disembodied tireless digital robots. Only now, however, do we recognize these systems as intelligent. Rarely do we stop to admire the awesome depths of what has already been automated, what we take as a fundamental aspect of our lives today, unthinkable to our ape ancestor: plumbing, electricity, architecture, wireless internet and radio, traffic and trade, the economy, food production, medicine, writing, the law and all other social structures that enable us to live our lives in the way that we do today. We have become so embroiled with our technologies, customs and infrastructure that it is often hard to tell where one ends and the other begins. That line has thus far been drawn by deference to one of two simple arguments: intelligence and consciousness—we are smart, creative and able to solve new problems, while the technologies we use are dumb, procedural, and pre-programmed, we are conscious, have free-will and can recognize our own existence, while the technologies we use are dead, deterministic and unaware of themselves.

The advent of modern large language models, algorithmic systems which can reason, behave in ways we cannot predict and fully understand, is threatening to change that worldview. To put it simply, the systems that underly us and enable our existence are creeping in towards our level of intelligence and we will need to rethink how we understand our own cognitive machinery with respect to the digital one, if we are to coexist with them. We are having, much like in the time of Copernicus, to face the fact that we may not be as special or privileged as we once though. We are having to reevaluate the axioms we never needed to question before. If brains and language models are both machines, biological and digital, what makes us conscious and them not? What is it that gives us freedom of will and denies the same to an algorithm? At what degree of complexity do freedom of will, self-recognition, and 'general intelligence' emerge from simple non-thinking components?

More importantly, how do we engineer and create intelligent systems that behave the way we want and can predict and is that even physically possible? Although I do not claim to have even begun to unravel the complexities of these topics, I do believe I have found inklings of what is to come—revelations about what intelligence is, how we had created intelligent autonomous systems even before the advent of LLMs, how building new intelligent systems will by necessity mean an extension of human cognition and whether consciousness is all that unique a phenomenon. The key question to begin addressing all these seems to be: what is intelligence?

Before explaining each of the insights elaborated in the treatise, I would like to briefly and succinctly summarize the insights presented in it. Some of these may sound aphoristic, but each is developed in detail in the attached 100-page document. I strived as much as possible to provide tangible evidence for each of the claims, providing my own experimental insights wherever applicable, and supporting my claims with results from notable studies across sciences and literature where the current technology did not allow immediate experimentation, or the experimentation would be impossible due to budget and infrastructure constraints. Although the main argument I wish to put forward is that 'intelligence is a measure of generality', the consequences of this for the development, integration and maintenance of intelligent systems, can be delineated into a neat set of aphorisms. Some of these may seem overly abstract to a technically novice reader, but tangible predictions will follow immediately.

- Intelligence is a measure of generality. The more problems a system can solve, the more environments it can adapt to, the more concepts it can integrate and associate, the higher its intelligence. The term AGI (artificial general intelligence) is pleonastic—to be intelligent is to be general.
- Intelligent systems are fractal and scale-invariant. The more intelligent the system, the more its components begin to look like the system itself.
- Intelligent systems are not modular. The more intelligent (general) the system, the more interconnected its components, and the less it can be split into logical modules. Intelligence requires component integration. Consequently, the higher the intelligence of a system, the less it can be designed and the more it must be evolved.
- Interfaces reduce intelligence. The more intelligent the system, the fewer the interfaces and the higher the intrinsic component coupling. Consequently, specialized systems (be they tools or humans), which must interface with the overall intelligent system reduce its overall intelligence. As a system becomes more intelligent, it becomes more integrated and modules and interfaces begin to move towards the

- Intelligence requires interaction and self-modification. Static systems which do not self-modify cannot be considered intelligent. The more general the system, the fewer its distinct components, the less pronounced the difference between computation and memory.
- Intelligence is a field, not a localized attribute. A system is not uniformly intelligent across its components. The more coupled the components, the more singular the intelligence. Certain regions of a system may be more general (more intelligent) than others, depending on the choice of partitioning.
- Intelligence cannot be measured by a specific test. If intelligence is a measure of generality, testing its performance on a single kind of task only tests its performance against a specific domain. To test for intelligence, one must test for generality. Human intelligence generalizes over human-relevant problems—it is no more 'fully general' than the current-generation artificial intelligence.
- Truth is alignment between prediction and perception. The more intelligent the system, the more faithful its model of the outside world. The more misaligned the world models between two systems (say, a human and an LLM), the higher the incidence of perceived confabulation.
- Intelligence is not independent of values. Highly intelligent systems cannot be aligned with human values which are not general. The more general a value, the more 'alignable' it is with a high-intelligent system. Specific human values cannot be aligned with higher intelligence, without reducing it (intelligence is a measure of generality).
- Intelligence is a measure of consciousness. Intelligence as a metric is highly correlated with information integration and relative entropy, qualities which have been highly correlated with what is typically understood as consciousness. This, in simple terms, means that the two phenomena are, at least to some degree, overlapping.

From these fundamental discoveries, several important predictions seem to naturally follow.

- Interfaces will not be designed but negotiated ad hoc. User-interface design is going to occur at runtime. The higher the intelligence of the service provider, the more adaptable the user-interface.
- Code, natural language, data and user-interface specification will be merged into a single paradigm, at least for internal model use. In other words, highly intelligent models' internal language will be polymorphic with respect to the four categories.
- Operating systems and companies are merging into a hybrid paradigm. Technology is a mediator for human-to-human communication, but humans are to become mediators of technology-to-technology communication. Humans are going to be components of a socio-digital operating system, much like AI algorithms.
- Real-world application is the ultimate benchmark. Testing intelligent system performance will require measuring against an aggregate meta-benchmark consisting of all individual benchmarks combined. The more diverse the data set, the more reliable the measure of intelligence.
- Hallucinations can be solved by training for behaviors instead of generation. Training for retrieval (RAG) and training for reasoning and self-correction will outperform hallucination mitigation in pre-training.
- General ethics are likely to arise from increasing intelligence (up to a point of dissolution of the term). Specific ethics may be administered through training for policy following.
- Compartmentalizing systems into less intelligent specialized interacting modules is less likely to produce a 'conscious' agent than scaling of intelligence. In effect, producing interacting expert systems reduces generality, which reduces the system's capacity for self-representation.

A single article can hardly summarize a 100-page document, but I do hope to provide at least a basic overview of the insights which I believe may prove immensely important for the development and understanding of intelligent systems.

The process of automation continues as it becomes the process of integration between the biological and the technological. As organisms, we are becoming not only socially dependent on our tools, but also biologically so—our medicine has enabled conditions that would otherwise have been eliminated by natural selection, our laws, technology and social contracts have enabled behaviors that would have easily eliminated our ancestors from the gene pool and these behaviors are being reinforced and Baldwinised into our psyche and biology. We are becoming part of the hybrid collective intelligence emerging from the interplay of society and technology.

While the current generation LLMs are not nearly general enough to solve even the basic automation problems, some of the hurdles of integrating generalized problem solvers are already surfacing. The problems are evident both in our failures to produce proper benchmarks of these systems, reliable RAG performance tests, as well as mitigating so-called 'hallucinations'.

The discrepancy between benchmark results and real-life performance comes in part due to the fact that the more general a problem-solver is, the more difficult it becomes to quantify its performance and value. Much like our education systems must to some degree be standardized and templated, limiting creativity, to allow for standardized testing, we need to curtail the breadth and scope of our testing of AI systems. On the other hand, true creativity, if it can even be defined (in the paper I argue that 'true creativity' is effectively randomness and that what we are really looking for is 'useful creativity'), cannot be quantified. In effect, much like an employer must be satisfied with a worker's performance and individual metrics only serve to inform the employer's ultimately subjective decision, the user must be satisfied with the AI system's performance in the user-study's conclusion. Fundamentally, we are building automated systems to improve human lives without having a quantifiable metric of what constitutes a 'good' human life. Hence, all our efforts to quantify the usefulness of any system, including AI, comes down to the opinion of the user. It seems that a combination of both meta-benchmarking and user-studies is going to be the ultimate determiner of AI systems' success over the period of the next couple of years.

Beyond a couple of years, it is difficult to tell the future evolution of artificial intelligence without engaging in significant speculation.

Nonetheless, basic insights into intelligence as a more general phenomenon do give us the means to make an educated guess. If AI becomes sufficiently intelligent to carry out our work, it will likely have also become entangled with our economic and social structures. The value of human intelligence is likely to increase in the short term, as only the intelligent will be able to understand and predict this system's behavior, as well as integrate with it. Simply put, not every human will be the 'human in the loop', but only those who can integrate well with the overall evolving intelligence. The distinction between artificial and human intelligence is slated to disappear. Society is an evolving landscape of hybrid intelligence predominantly consisting of human, digital and emergent social intelligence. Those who cannot integrate, be they low-performing humans or outdated specialized algorithms are going to be left at the outskirts of the continuously integrating collective intelligence and those who integrate well will, by necessity, abandon the traditional human modus operandi and become part of the system. The outlook, from today's perspective, is looking bleak for both sides, but there may be a kind of a third solution in which humans exist to support the collective mind almost like the human gut microbiome supports the brain, all the while remaining separate from the collective consciousness and existing as individual entities.

The real question here is what is to be done by the regular small person and what ought a small business' strategy be? If the scaling continues, the simple and the unfortunately unspecific answer is adaptation. Technology, while solving a given group of problems, has always created new and usually more complex problems for us to solve. In fact, as the intelligence (generality) of our technology increases, so does the generality of problems it creates for us. For that reason, the perceived entropy, the indeterminacy, of our future seems so high. We are faced with a group of problems which cannot be solved through specific solutions, but through sheer behavioral adaptation. The world will simply become faster, problems will be solved faster and new ones created faster, jobs will disappear faster, and new ones created faster. Those who are slow will be left behind; those who are faster will be left behind later. The unfortunate answer to responding to intelligence is intelligence.

Amid this shifting landscape, those who are, in the short term, profiting the most are those who are advertising concrete predictions with confidence and determination. Yet, a gambler who roles a seven was neither right nor prescient, but simply lucky. We are competing against a future which is increasingly more difficult to predict. Integrating with those who already have the power and ownership of what is to become the seed and the building block of the future collective intelligence seems to be the best way forward. It is not about knowing how to use these systems, it is about using them and associating with them. It is not us who are going to integrate highly intelligent systems; it is those systems that are going to integrate us. It seems, however unsettling this conclusion may seem, that all one must do is make oneself dependent on and embroiled with those systems sooner than everyone else. In other words, one must invest money, time and intellectual resources into those technologies that will evolve to be part of the collective intelligence. Identifying which technologies, companies and countries seem to be the best candidates for such association is difficult for two major reasons: first, because we cannot predict the behavior of such a future system and second, because it is simply too early to tell the winner.

Most of the current generation implementations are based on an incredibly well-known set of foundational papers from machine learning, data science, computer science and artificial intelligence. Most solutions rely on fundamentally similar architectures and most well-informed researchers are aware of what the next major steps in architecture and training methodology should be to improve the current generation systems. The differentiating factors are of the financial and infrastructural kind. For that reason, the most prudent course of action seems to be to stay well informed about and connected with those actors with the most access to computational resources—they will be dictating the pace of social progress and pace of integration. In the long term, those who have the means to enable general solutions will assimilate those who don't. However, for the time being, preparing the existing infrastructure for the introduction of more general (more intelligent) systems seems the right course of action. Effectively, we ought to be building systems which, due to relying on current generation AI, neither provide value nor demonstrate performance today, but are likely to excel if the performance of the underlying AI increases. This strategy, if AI turns out not to scale in the expected way, is a severe risk in the short-term. However, the payoff in the long-term, should intelligent digital systems scale as advertised, may be a significant market advantage against competitors who have not invested in this AI-embroilment early enough.

For the regular folk, none of these prospects seem to be bright. It is unlikely that the world will become more lenient towards the unintelligent and uninformed. In fact, human intelligence seems to be the best defense mechanism against the ever invading digital and corporate kind. Much like in financial investing, the prudent decision seems to be to simply follow what the major actors are doing—they dictate the game and set the rules anyway, whether the game is real or not.

Some of the most profound revelations I'd had over the years seemed obvious in hindsight. I would certainly hope that the same obvious truthfulness would also persist with the aphorisms laid out here, but I would be content with them just having sparked interest in the reader to read through the more thorough treatment of the topic made in the full paper on Engineering Intelligence, Minds and Cognition.