Better Machine Intelligence
— A Vision for Nordic AI Leadership

Dr. Kristinn R. Thórisson
Professor, Reykjavik University

" Together we bear the responsibility for creating a society where technology serves people – not the other way around. We are facing a large and complex task that requires dialog, solidarity and a clear vision. "

— **Halla Tómasdóttir**
President of Iceland

" Nordic democracies must think of new constructive pathways to address questions of AI governance. Ignoring the details of responsible AI while introducing AI systems everywhere is bound to have significant adverse consequences for citizens. "

— **Alex Moltzau**
Norwegian Government AI Unit | AI Compute Policy Fellow at the University of Cambridge
Ex European AI Office | AI & Robotics Policy

" For AI to respect human dignity and truly serve the common good, responsibility must be clearly defined at every stage [...] In many cases, however, the internal processes [...] remain opaque, making it harder to assign responsibility and correct errors. "

— **Pope Leo XIV**

Icelandic Institute for Intelligent Machines | **Vitvélastofnun Íslands ses**
Technical Report **IIIM-260526-001** | May 26, 2026

Better Machine Intelligence — A Vision for Nordic AI Leadership

Kristinn R. Thórisson [1,2,3]

1. *Professor, Department of Computer Science, Reykjavik University, Iceland http://www.ru.is/en/*
2. *Managing Director, Icelandic Institute for Intelligent Machines, Reykjavik, Iceland http://www.iiim.is*
3. *Co-Director, Center for Analysis & Design of Intelligent Agents, Reykjavik University, Iceland http://www.cadia.is*

ABSTRACT Trustworthiness in products and services can only be achieved through transparency, predictability, and controllability. Behind much of the current discourse about artificial intelligence (AI) – what the technology can do and what it could be used for – lurks an assumption that contemporary applied AI can meet these requirements—if not today, then surely in the very near future. However, the architecture of contemporary applied AI rests on principles that make it inherently opaque, unpredictable, and untrustworthy, and this is just the start of a longer list of even more serious problems: By supercharging surveillance and espionage it threatens privacy; by weakening cybersecurity it threatens national sovereignty; by sowing lies and discontent it threatens democracy; by dependency on centralized data centers it concentrates power and threatens our freedoms, and by free-riding on the collective total sum of digitized human cultural and scientific output from recorded history until today, it threatens our creativity. In short, *contemporary AI is a direct threat to the very fabric of society*. This is categorically not what I had in mind when I started my AI journey almost half a century ago and undoubtedly not the kind of future the founding fathers of AI envisioned. If AI is to play a positive role in our future – as I believe it can and should do – it must be better than this.

Despite rhetoric about reaching general intelligence soon, contemporary AI remains statistically based rather than transparent reasoning systems. Instead of hopping on the contemporary AI train, with technology that is architecturally and scientifically closer to the artificial neural networks of the 1950s than general intelligence, the vision here outlines why and how the Nordic and Baltic nations – which stand to benefit vastly less from it than larger ones, due to small data – can strategically invest in alternative AI approaches that are more transparent, capable of cumulative autonomous learning, empirical (causal) reasoning, are more controllable, reliable, explainable, and less compute-demanding approaches—in short, *trustworthy small-data AI*. Is this vision achievable? Yes. Much of my professional career has been focused on exactly that. If we work together across borders we can create, in less than a decade, alternative AI paradigms that match our needs better than contemporary AI ever will, in direct alignment with a future we can all be proud of.

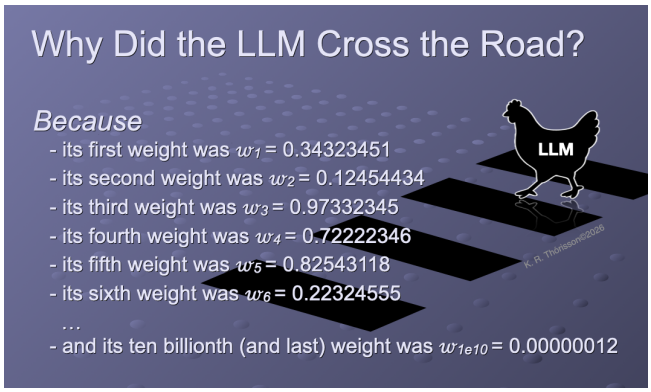

Figure 1: The joke highlights why information structures in large language models (LLMs, a type of artificial neural network used in modern chatbots) cannot support the kind of information segmentation needed to produce verifiable hypotheses (useful explanations)—its data are as incomprehensible to the system's human creators as to the system itself (see p.13). As a result, their behavior is rather impossible to control.

KEYWORDS *artificial intelligence, general machine intelligence, innovation, society, transparency, explainability, small-data, democracy, national sovereignty, fun and good fortune*

Introduction

Everything about human society worthy of mention is due to thinking; this activity of human minds is what most clearly separates us from other animal species and makes our creations and actions so fundamentally different from theirs. Knowledge has always been preferable to ignorance, but it doesn't come for free—it must be created first, and subsequently managed. A deeper understanding of the process of knowledge creation – how we come up with new ideas and apply them – brings us

Contents

closer to the very thing that makes us human. This is the topic we find at the heart of the scientific research field called 'artificial intelligence' (AI).

If thought is a process of mind, intelligence can be seen as the capacity of that process to bring about change and affect the world. Of course, many things can bring about change in the world yet are not 'intelligent;' it is the systematic, pre-planned, goal-driven kind of change, and equally importantly, an ability to act in light of the unknown, that separates intelligence from all other kinds of processes on the planet (cf. Thórisson, 2020a; Wang, 2020).

At the risk of oversimplification, intelligence may be seen to consist of three complex but fundamental processes whose interleaved operation produces mind: *Control*, *learning*, and *classification*.[1] We are familiar with these from everyday life; *"Is that an airplane or is it Superman?"; "Look out Toonces, you're driving off the cliff!"*—the former is about classification, the latter is control.[2] *They are not the same kind of process*. The former can be implemented as a static pipeline system, should we wish to do so, as a 'frozen lookup' process; the latter is an active process that requires an active feedback loop, including fresh measurements, and thus cannot be frozen. Both are critical for survival in nature, and industry has already adopted (simplified versions of) these in many ways for a variety of purposes and conditions. The third process, learning, while seemingly an intuitively straightforward concept, is actually an extremely complex process that neither neuroscience, psychology, or AI have yet completely come to grips with.[3] While each of these processes has features and properties that set it apart from the others, when it comes to intelligence they are inseparable: To produce thought, they must be intricately interwoven in complex ways that are not yet well understood.

When I was 12 years old I read the book "The Mind," one of the many in the excellent Time-Life series on science (Wilson, 1965). I was awestruck by Grey Walter's work on the autonomous robo-turtles that could autonomously find the way to a charging station. The idea of the brain as a giant mesh of wires allowed me to imagine a future where robots would do all the boring and dangerous work that humans didn't want to take on. "If the brain is the seat of the mind, and the mind is what gives the human species superpowers, we can make machines that do the same!" I remember thinking. That summer, AI became the most exciting and obvious career goal for me to pursue; a future with the greatest potential and largest amount of flexibility to suit everyone's wants and needs. Growing up in Norway and Iceland, countries that rank among the world's most successful, egalitarian, and peaceful social democracies, at the far end of the biggest prosperity growth period in Iceland's 1100-year history, I was not well-positioned to understand or contemplate how different historical circumstances in other regimes might end up distorting and upending that vision.

About a decade later, in my early 20s, I bet the rest of my professional career on the prediction that AI would become big early on in the new century. If that would turn out to be correct I should have time to get the education that would allow me to participate in that endeavor. That's what I did. Now the evidence is in; that was a bet I definitely won. But the actual future that has arrived has turned out to be quite a bit more complicated than I anticipated. The core foundation of contemporary AI – feeding a giant artificial neural network with the entirety of humanity's digitized information,[4] powered by the energy of a medium-sized city, then convincingly packaging and polishing it to take the form of an imitation human – is absolutely something I did not expect to see. The hype and enthusiasm surrounding the technology, coupled with global surveillance and warfare, has made its inroad feel a lot like the plot of a dystopian science fiction movie. A

[1] To produce intelligence, the mechanisms for runtime coordination of learning, control and classification is done by what is referred to as a 'cognitive architecture. While the three-part division of mind is an oversimplification, it is much closer to how minds actually work than any claim that contemporary AI technologies like large language models (see Figures 2 and 7) posses intelligence, or that it can in any way be considered any kind of a theory thereof.

[2] I use the term 'control' in a more general sense than is typical in engineering.

[3] When a particular research topic is shared between scientific fields, the fields – being separated not only by their main focus but also by methodological and theoretical differences – will inevitably address that topic differently. Of the four kinds of scientific explanation that mirror a topic's scientific depth of understanding – prediction, goal-achievement, explanation, and re-creation (Dellsén, 2018; Thórisson et al., 2023; Thórisson et al., 2016) – neuroscience and psychology, like most other fields of scientific study, only commit to the first three; AI is the sole one that makes the fourth one a strict requirement of its core aims (cf.Thórisson and Minsky, 2022).

[4] We will mention LLMs and other genAI technologies throughout, but our focus will be primarily on the key technology under the hood on most contemporary AI systems: ANNs. Reinforcement learning and various other statistical machine learning approaches are also subject to limitations, but we do not address those specifically.

lack of scientific grounding, coupled with the technology's built-in flaws, has resulted in a discourse riddled with lack of judgment, misunderstandings, and misguided expertise, has reached levels that boggle the mind. This is categorically not the kind of AI I wanted, and not the kind we need.

What has been built up in the Nordics – the social structure, effective and universal educational system, robust social safety nets, a healthy balance between private and public ownership and control, the democratic processes, the collegiality – takes tens, often hundreds of years of effort to build up and maintain, yet only a few to lose. Why don't we just sit back and let the movie play out? Because the threat is real. Trust would be the first thing to go. With a high level of trust – in our democracy, institutions, legal framework, scientific knowledge, and human decency – it would be reckless not to take this seriously.[5] All that is now under threat. The good news is that we can actually do something about it, and we still have time to do so if we work together.

Yes, the world around us is changing at an alarming rate, but the reasons for doing good scientific work have not changed, and neither have the reasons for pursuing AI as a topic.

Why We Should Pursue AI

There are two main answers to the question of why we should study intelligence. The first one is the practical one; the one that any 12 year-old can grasp: With "more intelligent" machines we'll have more options for automating the things we want to get done. Having more ways to perform them gives humanity more options for how to implement what is needed to create and sustain an effective and efficient egalitarian societal structure that works for everyone. Before we begin, to be worthwhile, we should of course have a clear idea of what we want done, what concepts like fairness, ethics, "no one is above the law," and so on really mean, pragmatically speaking. This is where social democracies can demonstrate their strength, through public dialog, the court system, governmental oversight, and democratic processes in ensuring that any corners that may be cut do not unfairly cost anyone's arm or leg.

The value of AI technology is its promise that it can leverage scientific understanding of the key processes behind intelligence, for better shaping of society in desired ways. Contemporary AI works well for some kinds of information transformation, and is certainly inspired by this ultimate goal, but it is not (yet) based on very deep understanding of intelligence but rather, the application of advanced techniques and enormous compute power (Figure 2). Needless to say, it is not enough to simply emulate intelligence if the technology to do so is too dangerous, too unwieldy, or too expensive—whether environmentally, economically, or socially. It turns out that contemporary AI actually teeters on all of these flaws and in some cases crosses over. The reason: Its lack of a solid scientific foundation. To be a true success for automation, AI cannot simply rest on the recent milestones that have pushed statistics and compute power to new levels; it should rest on a proper scientific explanation of the processes that turn information into knowledge and knowledge into intelligence.

The second answer to why science should study intelligence has to do with our shared pursuit of understanding the world; deepening our understanding of knowledge, thought, and the human mind. This may seem like a luxury, but if we want the practical benefits of AI, it is a

'Contemporary AI'
Characteristics of Applied AI In Active Use

Artificial Neural Networks (ANNs)
Umbrella term for the technology behind most contemporary AI applications.
Require enormous amounts of data up-front that are 'baked' ('trained') into them.
Can be set up (baked) for a variety of classification purposes using pre-defined statistically-tuned input-output mappings, before baking.
Does not change ('learn') after it leaves the lab.
Inspired by nature (McCulloch and Pitts, 1943) but does not work (and has never worked) anything like the neural networks in natural brains.

Reinforcement Learning (RL)
Umbrella term for contemporary AI systems that can adapt at runtime, after it leaves the lab.
Data must be limited to a handful of variables given before launch; often combined with ANNs.
Inspired by reinforcement learning in nature (Sutton and Barto, 1998) but inferior to natural reinforcement learning found in e.g. house pets.

Machine Learning (ML)
Umbrella term for a variety of methods for data classification, including those above, as well as autocorrelation, decision trees, predictive coding, clustering methods, and many others.

Figure 2: The term 'contemporary AI' as used here refers to all applied AI technologies in practical use today. For the present discussion we can think of them as relying on methods from (one or more of) these three main categories. Technologies called 'generative AI' ('genAI') are all based on some form of ANNs, like today's chatbots, which are based on a type of ANNs called 'large language models' (LLMs). Other ANN types include deep neural networks (DNNs) and generative adverserial networks (GANs). No contemporary AI has real intelligence, but it is classified as 'AI' because its origin can be traced back to at least a century of academic research on cognitive science, artificial intelligence, and brain science.

[5] The Nordic Council of Ministers calls trust "the Nordic gold" (NCM, 2017), boldly stating on the cover of a report on this topic that: *"The Nordic region has the highest levels of social trust in the world, which benefits the economy, individuals and society as a whole. This [...] is now under threat."*.

major error to think that technology can progress without science (the inverse is also true; science cannot work without technology, but this is of lesser concern here): To ensure that the tools we create deliver what is intended, to deploy them safely, and properly handling cases where things go wrong, requires them to have a proper scientific foundation. This does not come about without effort, of course, and is too intricate to be curated and delivered entirely by a market economy. Let us recall that the very concept of a market economy was invented and implemented with an eye to achieving particular results, addressing particular aspects of an accountable society. The commoditization of intelligence upends that purpose, but at the same time opens up vastly more options for how to design a well-functioning society. The purpose of AI research is therefore not to be able to copy or imitate human minds to a tee, but rather, to lay a scientific foundation to cognition that allows us to build a better world.

A lot of progress has been made in research on intelligence over the past century, in psychology, neuroscience, philosophy, and artificial intelligence, but no common framework has emerged that unifies even half of its many aspects in a single theory. If we take the analogy of a puzzle, and the various aspects of intelligence as the pieces – reasoning, planning, learning, control, pattern recognition, communication, to name a few – fitting them together to create a clear picture of how they work *together* requires us to understand the shape of each one. But when each is studied in isolation from the others, as has largely been the case in AI and related fields, fitting them together after the fact becomes pretty impossible. This part of intelligence research is thus still on the to-do list of science.

Inflated Claims of Contemporary AI

A significant increase in AI-branded products, services, and technologies in the past few years has given many people the feeling that progress in the field of artificial intelligence is speeding up. Even though media coverage seldom lends this assumed 'speedup' more space than a sentence or two, and hardly ever qualifies, explains, or defines it,[6] to the vast majority of people talking and writing about AI this assumption has become a jumping-off point for wild speculation about the kinds of AI technologies we can expect to see within the next few years. Information technology leaders offer frequent quotable-yet-questionable claims about what lies on the near-term horizon, adding to the hype. Thinly disguised sales pitches and attention-grabbing absurd predictions are taken at face value, frequently setting the Internet ablaze. These unfounded predictions invariably contain conflicting information and bewildering contradictions, exacerbating the confusion and making the dialogue incredibly difficult to decipher by citizens and politicians alike. Journalists often fail to follow up with these self-proclaimed prophets when their predictions eventually fail to come true—which they always do (Bode, 2026).

Contemporary applied AI has locked onto classification alone, leaving out cumulative learning and control. This lopsidedness has lead to much misunderstanding, including that the only work left before we have general machine intelligence is for the tech giants to extend their existing products. Smaller teams with alternative approaches give up on even their greatest ideas because they seem incompatible with this view, and many competitive research grants – including those intended to ensure long-term thinking – choose to explicitly require popular approaches (e.g. deep neural networks) instead of describing scientific goals (e.g. trustworthiness, reliability, transparency) as conditions for application, making it impossible for more creative research teams to compete with the large industry labs.

The lopsided narratives around AI continue and the hype persists uncorrected—not only because of the salesmanship; contemporary applied AI technologies rest on extremely weak theoretical foundations, and while statistics plays an important role in science and has a relatively long history, statistics-based AI technologies developed to date, like artificial neural networks (Foukalas, 2025) and computational reinforcement learning techniques (RL; Sutton and Barto, 1998), leave out key concepts in cognition; research on how we get from statistics to real thinking is still considered fringe (cf. Thórisson, 2012),[7] so the claims are difficult to fact-check.

Much of the confusion about contemporary AI is due to the impressive performance of large language models (LLMs) to produce convincing text and human-like dialog. They are good at imitating human text production because that is precisely what they have been built to do, using enormous amounts of textual data sourced from the Internet. It is important to keep in mind that:

- **a** LLMs lack explicit semantically structured causal representations needed for robust reasoning and self-explanation.
- **b** LLMs do not present a complete architecture for general intelligence.
- **c** LLMs cannot substitute for an adequate, scientific, explanatory theory of intelligence.

We will dive into these topics in the coming sections.

The Source of Perceived 'AI Progress'

Progress in applied AI research over the past decade has primarily been due to improvements on three main fronts. Firstly, greater access (by a few tech giants) to a treasure trove of digitized content; secondly, improvements on machine learning methods based on artificial neural networks (ANNs), in particular methods

[6] An exception is when it's defined in investment or expenditure (cf. OECD, 2026), but that is not the subject of this paper.

[7] We touch on some of this on page 16 in the section on general intelligence, and dive into the theoretical aspects in the section *Theoretical Limitations of Artificial Neural Networks*, p.21.

involving pipelined computation for matrix multiplication suitable for graphical processing units (GPUs); and thirdly, access to extended use of enormous amounts of compute power. Put together, this has enabled what boils down to the deepest, most advanced statistical analysis of human-created content ever. I don't think it's an exaggeration to say that the capabilities of these systems have taken most people by surprise, but many things are still missing—too many to justify calling them 'intelligent.' ANN-based systems do not 'think;' in no sense do they possess a 'mind,' nor do they have 'agenthood;' they are just as free from all these features as any piece of software you have ever run on your computer. We will get deeper into these concepts in the coming sections.

Outlining these facts might be construed as "technology bashing," but don't get me wrong, ANN-based software has, in the last few years, swiftly entered applications, making genAI technology like chatbots familiar to the general population. These innovation have seen a number of experimental uses—some of them useful, some much less so. On this there is little disagreement. However,

> these AI-driven innovations are *not* progress on AI per se but primarily consist of progress on the *application of AI*, driven mostly by standard software and product development methods, services creation, and traditional product packaging and marketing.

Rather than representing some 'new form of intelligence,' the 'progress' that the tech giants like to talk about is much more appropriately described as experimental software development and deployment. It is software development *by the tech giants, for the tech giants*.

Much of the panic related to AI stems from the novelty that contemporary AI brings, but also from gross overgeneralization and misuse of terms. When technocrats – e.g. tech giant employees – use the term 'AI,' their claims tend to be based solely on today's applied AI—often the applications that they themselves are making this month or this year. While these technocrats may be experts in (their own) applied AI technology, rarely are they well versed in the history of AI or relevant cognitive principles that inspired the methods. Their claims about 'AI' are predictably based on the assumption that future AI technology will (and must) be based directly on the AI we see today, only with foreseeable and obvious linear extensions; "contemporary AI on steroids." This is absolutely not the case, as we will see in the coming sections.

Using the umbrella term 'AI' – the name of a whole research field – when talking about the development of contemporary AI, now or in the future, exacerbates these problems, making it incorrectly seem like the future of AI is entirely dependent on extensions of *the present state of the technology*. Since the underlying AI technology itself is "given," it lands in a blind spot that prevents people from thinking about what's under the hood, new applications based around the same ANN-based technology as before look like amazing advances; 'AI' seems to be developing at blazing speeds. Ironically, much of the new stuff in such applications, agent-based systems to take one example (cf. Knight, 2026), consists largely of traditional software.

In contrast, to academic AI researchers, some of whom have been working on such systems for over 30 years or more (like yours truly), the term 'AI' usually refers to the whole field of AI, including both contemporary AI *and* future AI. The general public, politicians, and even software developers, who don't realize that today's ideas in AI took half a century to turn into applied AI, may think that next decade's 'AI' will follow the same paradigm as contemporary AI, inheriting all of their main characteristics. This has the unfortunate albeit rather foreseeable side effects of stirring fear mongering and confusion. Paradoxically, it also induces a sense of urgency to "adopt contemporary AI everywhere ASAP," often without any accompanying curiosity as to "why." Being aware of what is already on the drawing board of AI researchers across the globe, what the unanswered scientific questions are, and what remains to be developed, makes it easier to sift through the confusion.

Another point of confusion is the software development methodology used in contemporary AI. Chatbots, image generation models, and other systems based on ANN technology, are built using allonomic methods—their structure, operation, and information organization is created in the same way as every other engineering project undertaken in our past: Built by human experts, according to given specifications, overseen by human experts, tested by human experts, and released to consumers by human experts. While some of these steps may make use of contemporary AI, it is in a supporting role, because it is neither self-guiding nor intelligent—not any more than bread in a baker's oven. This process mirrors exactly a baker in a bakery, collecting ingredients to make dough, arranging the dough, putting it in the oven, letting the heat do its thing, then taking it out at the right point in time and selling it. The result is a "frozen information" that doesn't have any machinery to create new meaning, to learn on its own, to form new concepts—it is *pre-baked*. The language used to describe these systems, filled by terms that make it seem like they have a will, an inner life, and human-like thought processes, flies directly in the face of the methodology used to build them. The language of conscious beings is used to describe the operations of a mechanical clock. Attempting to give such systems responsibilities that require responsible agenthood[8] is dangerous for

[8] Two important implications of the concept of 'agenthood' are *continuity* and *'abstractable reliability'*. The former refers to the fact that human (and many other animals) maintain a context of meaning, goals, and understanding that gives their behavior stability across time; the latter means that there is a logical

many reasons (they cannot be ensured to 'follow their programming,' due in part to their stochastic nature, and their opaque information comes from data whose quality is mostly unverified). Some ANN-based systems may *look* like they are worthy of responsibilities that require agenthood, but just like the language capabilities of LLMs, its foundation is sophisticated imitation masquerading as intelligence.

While contemporary AI is not really intelligent, it gives a partial glimpse of upcoming societal changes that may be necessary in the coming decades. The most significant changes will stem from *future applied AI*, based on technologies that have yet to leave academic research labs. It would be best if such technologies were based on autonomic methodologies (see Table 1, p.32), which are not pre-baked but rather continually self-forming. These have potential for offering transparency that goes well beyond the (unavoidable) opaqueness and controllability in applications of contemporary AI. Like the processes responsible for a growing tree or a fetus, these systems start from a 'seed' that is much smaller and compact than the eventual thing that springs from them (cf. Thórisson, 2012). The safety of autonomic systems that are driven by explicit top-down goals will be verifiable and certifiable through advanced empirical methods, and they can be made (transparently and logically) self-correcting via the same techniques. Much of this AI will be vastly superior to contemporary AI—requiring less data, less energy, be less faulty, and more trustworthy. Compared to today's AI, it can be made to be more controllable, transparent, manageable, human-friendly and useful. Over-investment and unfounded trust in contemporary AI directly hampers progress towards these newer, better AI paradigms.[9]

Because of the potential that AI technology holds for not just affecting, but *significantly changing* our society, it is extremely risky to have festering misconceptions about it. Here are the misconceptions that I think most urgently need to be eradicated:

1 The "AI race" has already been won, or is close to being won, by USA and China.
2 Europe can't compete with China and USA on AI technologies.
3 Present AI methodologies suffice to successfully implement a fully AI-centric society; little or no new scientific research is needed.
4 General machine intelligence is a myth.
5 Contemporary AI is a solid foundation for general machine intelligence.
6 To fully explain intelligence (human or general), no new scientific paradigms are needed.
7 Industry has already taken over all relevant AI research from academia, or will soon.
8 The arguments for abandoning AI R&D altogether are better and stronger than those for continuing on the present path to general machine intelligence.

If you consider any of the above statements to be true, or to have some merit, this paper is for you. And to those who disagree with them: This paper is for you too!

These eight misconceptions are especially dangerous because they all contain significant potential to (incorrectly and unjustifiably) become self-fulfilling prophecies. Let me answer the last point #8 first: There is no way to abandon AI R&D without rejecting all scientific pursuit outright—that the general scientific enterprise of producing new knowledge in every field on every front should be completely abandoned wholesale, because the march and purpose of AI research is no different than basic and applied research in *all other fields of science:* to deepen our understanding of the world. And as history shows, scientific knowledge has brought enormous benefits and well-being to humankind since the enlightenment (cf. Principe, 2011); the pros far outweigh the cons. The rest of these claims are addressed throughout the paper, from multiple angles; compact and extended answers for them start on page 30.

Does this all mean that we must accept our predicament and succumb to the noise? Not at all. Quite the contrary—now is the time to double down and do the work needed to prevent unnecessary delays in putting research and application of AI technologies on stronger footing. This involves many things, including the seemingly unrelated task of educating the general public about the nature of innovation, because, to better understand the future of AI requires making a clear distinction between *applied* research and *basic* research. We must also recognize that all revolutionary technologies go through many successive, logically connected development stages before they are ready to be applied in the real-world. This is how things were for the airplane, telephone, and computer. It is no different for AI.

Technology Readiness Levels

To move an idea from imagination and make it into a real product or service usually calls for a concerted attention and effort over extended periods of time, plenty of additional knowledge acquisition, prototyping materials, tools, good supporting ideas, and of course funding. The 9 steps of the Technology Readiness Level scale (TRL; Mankins, 1995) capture the idea of gradual development and refinement of fundamental ideas that eventually

relation between levels of abstraction of their goals.

[9] Any sufficiently powerful AI technology could potentially become a threat to society, whether or not it is based on statistics, like contemporary AI, or on something else. However, the opaqueness and lack of semantic hierarchy in contemporary AI turns it into an ill-manageable threat.

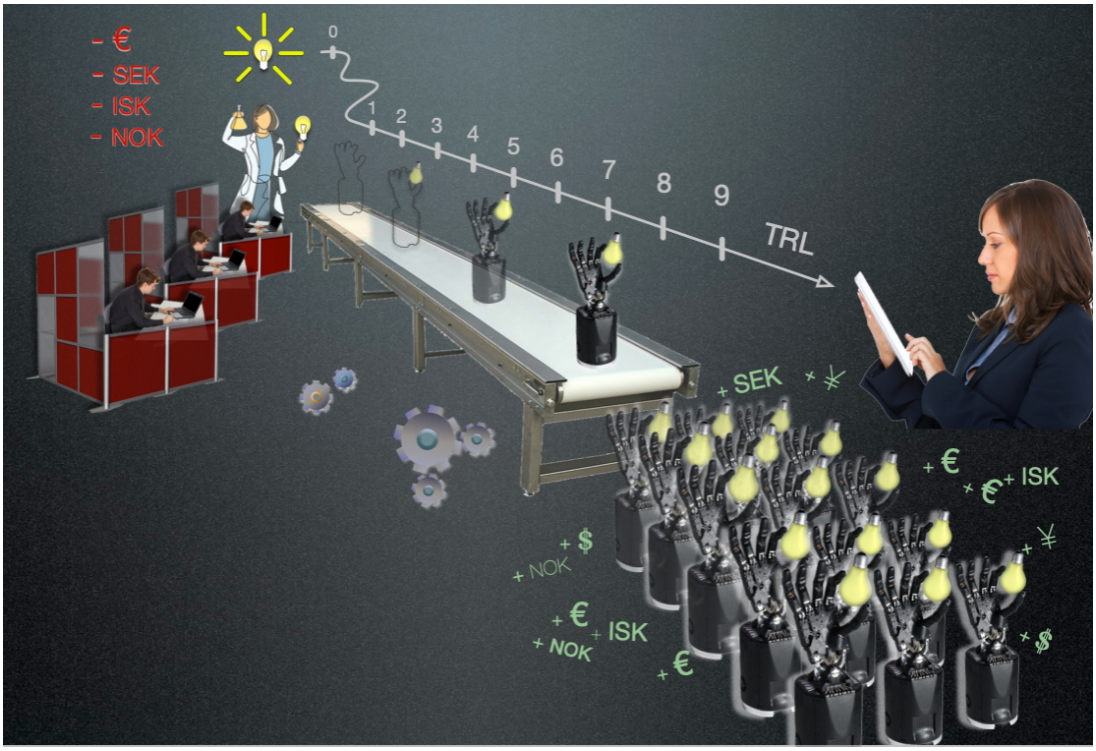

Figure 3: The image illustrates the *Technology Readiness Level* scale (Mankins, 1995) as a conveyor belt. Based on a "technology maturity" score from 1 to 9, representing the stages that all ideas must go through before they can become products or services, it starts with an early invention or idea in need of investment at TRL 1 and ends when it has turned into revenue-generating products / services at TRL 9. We added a "zeroth" level here to mark point in time when new 'blue-sky' ideas come into existence in someone's mind. Any good idea may die before they reach the next level, because moving new ideas along the innovation conveyor belt requires both mental work and muscle: Prototypes, funding, supporting ideas, etc. Most societies implement several such conveyor belts for various domains, stabilized with secured long-term government funding to non-profit research institutions at the lower end and middle (TRL 0 to 7). Besides electricity and communications, AI is the most obvious technology for any society to build such coordinated innovation infrastructure for across all TRL levels, due to its its domain-generality and potential for being a productivity exponent. Note that this refers *not* to applied AI but to development of the AI foundation itself.

become useful to society (Figure 3). At the low end of the scale (TRL 1 to 4) is *basic research*, where new ideas are molded and early demonstrations of their principles made. Intuitively, the more revolutionary an idea is, the longer some of the steps may take (depending on the nature of the ideas and the environment in which they are being developed). In the scale's middle we find early prototypes (level 5) and mature ones (level 7); at the high end of the scale are practical (but *pre*-production) solutions, 'bleeding-edge' (barely ready; level 8), and cutting-edge (close-to-ready; level 9). 'Turnkey' products and services beyond level 9 can be given a price tag and are ready to be deployed for their intended purpose.

By adding a zeroth level to the TRL scale we can stretch it all the way from daydreaming – a new idea in someone's mind – to the point in time when the idea (or some derivative of it) is offered as a product to the world at large. In between are many steps and typically many years, sometimes decades. Depending on the idea, this path can be long and unforeseeable, filled with random twists and turns. For a typical startup, things are going well if each TRL step translates to two years or less of calendar time. Only after development is done, by level 9, can a steady return on investment be expected.

> The TRL scale is descriptive; its direct application is intended for analysis, not planning. As a tool, it is nevertheless useful for reminding us about the inevitable spatiotemporal scope that innovation spans and for clarifying that invention and innovation does not solely lie with industry, academia, individuals, or companies—but *all of these combined in systematic coordination*.

Radical ideas for new products come along roughly every decade; radio, television, washing machines, microwaves, autonomous vacuum cleaners, are all examples of radical products. Truly revolutionary technologies come along a few times per century, like transistors, computers, and digital communication networks. Note how the computer, transistor, and digital network all fit neatly together and rely on each other, forming recursive dependency networks: Each is more powerful in some combinations with the others, as is commonly the case with complex, intricate phenomena.

Today's knowledge- and innovation-based societies could not exist without supporting the full path from daydream to reality; a world that cannot take the long view to support speculation, prototypes, manufacturing, and distribution of radical ideas could never offer the level of medicine, welfare, security or comfort that we have come to expect from modern society. To understand how AI fits into this picture we must look at the principles of how modern society manages invention, innovation, and technology. This is best done with a short trip along the full stretch of what we will call the 'conveyor belt of innovation.'

It Starts With a Spark...

The source of all fresh ideas is thinking. It takes imagination and curiosity to see opportunities where nothing exists yet, where others may only see impossibilities—or nothing. The more radical an idea is, the fewer people will "get it" in the beginning, and the more likely it is to be rejected by general consensus in a larger group. Seldom do completely new ideas emerge from interactions in large groups and the initial moments of invention rarely involve much communication—they tend to be driven by individuals (hence the concept of the "lone inventor"). The early pursuit of an idea, once it emerges, may involve a few people, but only after it has been manifested sufficiently clearly in an individual's mind.

AI history, having been ongoing for the better part of a century, is poised to continue for at least another 50, and the 25-year time horizon in this paper only covers half of that future. Several things have to be developed before general machine intelligence is possible, including dissecting what invention involves (because, yes, thinking involves generating brand new thoughts), so let's continue.

The production and development of good new ideas invariably involves hard work and struggle on part of the inventors, for months, years, and sometimes decades. Oftentimes this goes against (mostly well-meaning) advice from authorities, governments, or friends and family, who find a variety of reasons to convince the inventors to leave it be.[10] People who don't see the light are likely to consider an idea a "waste of time," with a variety of reasons given, like the idea's "impossibility," or that it's "too high cost." The more revolutionary the idea, the greater will be such pushback.

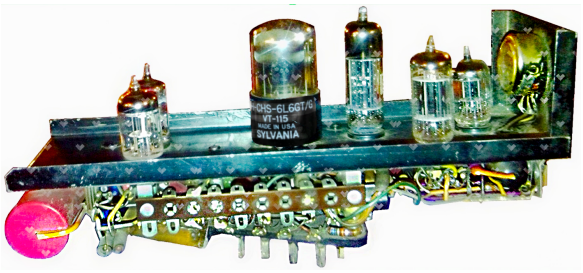

Figure 4: The first implementation of an artificial neural network, Minsky's Stochastic Neural Analog Reinforcement Calculator (SNARC; Minsky, 1952) was based on 40 electronic boards from a gyroscopic autopilot, each one which simulated the role of an artificial "neuron" in the system. The image shows one of these boards, of which there were 40 in the implemented prototype.[11]

The antidote to prematurely rejecting good ideas just because they sound 'too far out' is to do what makes a good scientist: Keeping an open mind and continuing undeterred. So, before you stamp yourself a member of that clan by prematurely rejecting my arguments in this paper, please keep reading.

The creation of fundamentally new ideas and theories requires deep analysis of unexplained phenomena. For this, a stable environment is essential; one that ensures space and time for uninterrupted thinking, exploration, dialog, and experimentation. Universities are one of the very few places in modern society that strategically strive to provide and sustain this kind of environment. They achieve this because of their charter, which is to produce and protect the best knowledge possible and transfer this to students and society with available funding and means. Universities do basic research because it is one of the best ways to ensure the best possible and up-to-date knowledge of all matters.

Companies, no matter how big or successful, cannot take on this academic function because their primary purpose, and ultimate measure of value and success, is profit and shareholder dividends. Companies do research when they need to improve a service or product, for their ultimate goal of monetary return on investment. If they don't, their leaders are wasting their owners' time

[10] In a world with limited time and energy this couldn't really be any other way, because the vast majority of people want to avoid spending their life in pursuit of useless things, and those who don't "get it"– of which there will never be a shortage – will always at best classify such pursuits as a waste of that precious time. Unfortunately, such pushback – its force or fervor – can almost never be used as a guide for an idea's future value.

[11] See www.historyofinformation.com/snarc — *accessed Mar. 22, 2026*. Image enhanced (carefully! by me) with Pixella.ai. The term 'neuron' is in quotes because, in spite of looking impressive, just like modern artificial "neural networks," they don't work anything like any neurons found in biological nervous systems.

and money. These same principles prevent them from freely sharing their innovations, results, or technologies with the world. The patent system was invented to work against this natural tendency towards secrecy, to allow knowledge sharing without complete loss of profit potential (Lessig, 1999).

In contrast, the main service rendered by academia is not to shareholders or owners, but rather, to students and society at large, so while their "return on investment" *could* be measured monetarily, should one care to do so, it is neither easily nor typically doable; instead, the number of citations to a university's scientific publications, number of research and teaching awards received, the ratio of graduated students to admitted ones, and so on, are used to measure other kinds of success that are more relevant to its goals and operation—and much better suited to motivate, foster, and support creative thinking. For a researcher to get one's work noticed and cited by others, dependable knowledge is key. Good education also requires quality knowledge. This is where *basic research* lives, and why it was born: Basic scientific research produces better knowledge than any other method known to humankind.

Proponents of contemporary AI chatbots seldom talk about the origins of their technology, which can be traced directly to academic research in the middle of last century. We know this because of papers written by Warren McCulloch and Walter Pitts (1943) and Marvin Minsky's SNARC (1952; Figure 4, p.8) on artificial neural networks, the latter being the first implemented ANN, which explicitly mention *natural* neural networks as a key source of inspiration for developing intelligent machines.[12]

The historical development of ANNs since then consists of a rather transparent succession of innovations and inventions, each building on ideas that came before, inching them towards the present day over a period of 6 decades. Vast amounts of additional compute power and digitized data made it possible for these ideas to become the technology it is today. The path from idea to LLMs: Just over half a century.

Unless you think that this is the only way to build intelligent machines, or is the last deep idea anyone will ever have about how to make them, stay tuned for the follow-on ideas that will emerge in AI over the next 20-30 years. There is much in store.

... Then Comes an Invention ...

When a novel idea has reached a certain level of maturity it can become the impetus for the targeted development of a product or service. This happened with ANNs in the 2010s, when Google trained them to find pictures of cats on the Internet.[13]

If an idea is sufficiently radical it can even become the basis for a whole new industry. The path for any product from radical idea to market is often long, but often it can be sped up with strategic governmental and communal support.

Most industrialized nations do this through non-profit organizations whose charter is to work on pre-market ideas and technologies. This is the idea with innovation institutions like Alexandra Institute in Denmark, Simula and SINTEF in Norway, Fraunhofer and DFKI in Germany, and the Icelandic Institute for Intelligent Machines, which focus on pre-competitive R&D. These foundations bridge the "no man's land" gap, TRL 4 to 7 inclusive, that unavoidably exists between basic research in academia and applied research in industry. Here we find technologies that are too immature for investors and market-driven funding to carry forward yet too short on scientific novelty for most academic conferences and journals. Due to the unique and invaluable knowledge that the innovators and researchers in such institutions posses, their work is often funded (in part) by individual companies, industry consortia, and competitive R&D grants, besides government and municipalities. Being a kind of active matchmaker and catalyst for innovation,[14] they identify potential challenges and friction in industry, find potential technologies that can be brought to bear on it in a positive way, and create prototypes that demonstrate how cutting- and bleeding-edge methods can be applied. With the help of politicians, such organizations often take the form of a public-private partnership, employing doctoral students and professors, professional inventors, as well as administrators and experienced industrialists with a good understanding of the innovation process.

Because the TRL4-7 gap is inherent in our modern societal structure, properly and strategically bridging this gap is not a "nice-to-have;" for any industrialized nation large or small that wants to be part of the global market. It is a *must have*, lest we are willing to accept significantly hampered impact from ongoing academic and industry R&D funding.

Making the conveyor belt of innovation work efficiently and effectively in every domain of society is not simple. Yet never in history has it been as important to get it right as for the technologies belonging to the domain of artificial intelligence. The reason is simple:

> *AI is an exponent of work*, like electricity, but it goes further than that; over and above electricity and the automation of muscle power,
> *AI can become an exponent of the innovation process itself.*

Having said that, it is important to remind ourselves that this is a *prediction about future AI technologies*; contemporary AI, covering only a fraction of what future

[12] After SNARC, Marvin Minsky argued extensively that the ANN-based path was a dead-end for reaching human-level artificial intelligence (cf. Minsky, 1982, 1988; Minsky and Papert, 1969).

[13] See for instance *"Google Computer Works Out How to Spot Cats" (BBC 2012*).

[14] There is an overlap with academia on the lower end of the scale and with the competitive market economy at the higher end.

AI technologies may enable, does not embody the functionalities required for fulfilling this prediction (see *What General Intelligence Must Be Made Of*, p.16).

... That Ends With a Product or Service

At the end of the production line is an offering that has carried a particular technology or technologies far enough to become useful in society. Startups sometime begin with their key technologies at TRL 5 or 6, but it is more common for them to be at TRL 7 or 8 in the beginning.[15]

For a typical startup, things are going well if each TRL step translates to two years or less of calendar time. Longer times are very common, however, and the median duration from startup to cash-flow positive is around 12-14 years. When level 9 has been reached, a beta or alpha version can be marketed and a revenue stream created.

Progress in the "AI Innovation Echo Chamber"

This brings us to the question du jour: If progress in AI has been increasing for the past few years, what *kind* of progress is being made? On what dimensions should it be measured? And also: Will it soon lead to machines with general intelligence?

When it comes to making smarter, more intelligent machines – following the path laid out by the founders of AI research from the very beginning – unless we want to spend time doing other things, we must make sure that the our chosen approach for such development is actually moving us in that direction. If the goal of our AI innovation is for next year's machines to be not simply a tad more flexible, a nudge more reliable, or a bit more desirable than this year, with a few more bells and whistles, but rather, that the machines we can build next year are actually and truly *more intelligent* than those we have today, then we must make sure that our methodology – the methods we have chosen to make progress towards that 'ultimate' goal – are really letting us make that kind of progress; empowering us to do that kind of work. Is that actually the case?

Since no extra funding has been made available to the scientific enterprise, or affected academia in such a way as to increase its capacity for basic research, the work in AI producing our feelings of "progress speedup" has so far exclusively involved R&D in companies.[16] So either companies have taken over basic research from academia, or basic research is no longer needed for the ongoing predicted progress.

Academia does basic research in a very efficient way—it *must* be efficient because its time horizon is so long. For this reason it's work is paid primarily from public funds.[17] To do basic research in the way academia does it, tech giants would have to had shared the vast majority of their results with the world (to match the publications of academia), at least quintuple the number of interns (to match number of graduate and undergraduates in engineering and computer science departments), and extend their invention horizon to at least 50 years (to match academic time horizons). There are practical reasons for why this is not possible while simultaneously protecting its intellectual property, one of them being that the standard duration of a single patent is only 15 years. So this is an unlikely argument for someone to make, which is probably why we don't hear it very often.

What else could be going on in industry then? How do companies plan to achieve their AI goals if they are not doing basic research? The only thing they *can* do is to work around the clock to improve on solutions from the last couple of years. But if last year's technology has already been released as a product, even if experimental, it must be at TRL 7 or 8, maybe even 9. That can only mean one thing: The work thus undertaken is *applied* research.

Rather than saying 'no' to a request for *"human-level AI in 24 months or less"* from their superiors, the teams that built some company's latest products and services imagine perhaps that this may in fact actually be doable, maybe based on a some new ideas they came up with last year and some clever changes to the software architecture they implemented three years ago and always wanted to get back to working on. Plus more data and more computing power, of course! In such work, the overall AI architecture, and its core principles of operation, are never called into question, because if they did, the time cycle for the next version would likely get too long, or at least hard to predict, if not both.[18] And because there is no underlying scientific theory of intelligence, one must cross fingers and hope it works. In the mean time, headlines keep overpromising, and market value of companies climbs. It is an 'innovation echo chamber,' pure and simple.

[15] Modern tech companies, due to their sometimes large size and deep pockets, have on occasion stretched internal development as low as TRL 4, but this is very rare. For example, Xerox' development of the graphical user interface (**xerox**) in the late 1970s and Google's development of large multimodal ANN models around 2012 (BBC 2012), two examples of advanced R&D that came out of mature companies, were in both cases primarily based on ideas that were arguably already at TRL 6-7.

[16] Quite the opposite could be argued, in fact, as companies now lure scientists away from academia with offers of multiplied salaries (Kunze, 2019). USA's government has cut significant amounts of competitive research grants to academia (cf. AAU 2025).

[17] There are some differences between countries in how such work is funded, but when all is said and done, public funding is a fundamental part of the scientific enterprise worldwide.

[18] This is exactly what happened over a whole decade with Tesla's 'full self-driving' feature; each year since 2015 the CEO of Tesla promised customers "full self-driving next year" and put this request to Tesla's engineers (Wikipedia/FSD-Predictions). Every year that promise/requirement was updated by one year, until 2025 when Tesla seemingly finally removed all mentions of future updates to this technology from their user agreement and stopped announcing its imminent arrival (AutoNext 2026).

> For industry, government, or society, a good scientific theory is not something that is *"nice to have."* Good science grounds the critical infrastructure of any technology-based society effectively and efficiently in evidence. *Nothing is as useful as a good scientific theory.*

Every scientific field, no matter how mature or immature, how successful or unsuccessful, keeps searching for better, more unifying theories to explain their phenomena of study. In AI, such work has often taken the back seat while practical applications grab the limelight.[19] This state of affairs cannot work in the long term. Now that the automation of virtually everything is on so many people's wish list, the lack of good theory is becoming increasingly unbearable for society.

So, we can summarize what is happening in industry AI innovation as follows. Companies take recent solutions (primarily their own, but also from their competitors) and push them back on the innovation conveyor belt, one or two TRL steps, along with a set of new requirements. Some of those requirements may be within the skill set and capacity of their industrial teams to deliver, others may be well beyond them, possibly even beyond the scope of present scientific understanding (as is the case with general machine intelligence). Within the confines of industry processes, given other deadlines and ongoing tasks that companies must attend to, this latter list of requirements will certainly remain unmet for several years if not decades.

Developing and improving contemporary AI technologies can in many cases be classified as innovation if the resulting technology is new (in the sense that no other company has done it the same way before or offer exactly the same technology). From the viewpoint of a company it can be called "progress" if the work pushes the envelope on some fronts of importance to the company. But for the above reasons, all of those fronts will be squarely within the confines of the present *contemporary AI paradigm*. So whatever progress is being made in AI, it is not the kind of progress produced through basic research—it does not produce fundamentally new paradigms.

Are there any conditions under which the innovation echo chamber can nevertheless be broken out of using this approach? Yes, there is one. To make a true leap over the present state of knowledge, this approach works if the state of knowledge – the proximity of the prototypes – are sufficiently close to a *finished product*. This would mean that basic research is no longer needed to reach the target finished product, which in our case could be 'human-level intelligence' (or some significant features thereof; we will get deeper into this concept shortly) – this is the looming assumption (threat *and* promise) of the AI industry. In other words, the technology for the requested product must already be at TRL 7 or higher. Comparing this to the airline industry, which now provides safe, scheduled flights all over the globe, the Wright brothers' prototypes were sufficiently close to a final product (a safe, predictable airplane at reasonable cost, with sufficient carrying capacity) that an airline industry could be implemented over a period of between 20 and 30 years after they demonstrated heavier-than-air flight. I think that flying machines are probably considerably less complex than intelligent machines, in which case the arrival of machines with real intelligence is unlikely to be sooner than three decades from now.[20] In the mean time, we can of course keep pushing for *more general* machine intelligence.

The multi-trillion Euro AI question is this: Is current know-how in AI sufficiently advanced to enable scaling up to general machine intelligence over the next years, and can this be done by industry alone? Or does progress towards such machines require new basic scientific research? To properly answer this, we must dissect state-of-the-art contemporary AI technology, along with the assumptions on which it rests, and compare this to (hypothesized) necessary ingredients required for general intelligence. So let's do that. But first allow me to present a short discussion on what makes products and services trustworthy, because this turns out to be quite important in how we plan for the future (cf. Afroogh et al., 2024).

Trust in Products & Services

Scientific knowledge has produced many dangerous things – materials, chemicals, energy, equipment, processes – that can pose unwanted threats to humans and the environment. Reaping the benefits of such knowledge while staying clear of their negative side effects requires special expertise, methods, permissions, measurements, and a proper regulatory framework.

For the benefit of public safety, every technology, gadget, service, and product must meet certain requirements: Neither users nor bystanders should become unwilling subjects of hidden negative side effects of their normal operation and use. Only under a proper regulatory framework can the general public trust that their own well-being is guarded when new products and services become available.

For the operation of common items such as power tools, telephone systems, and lawnmowers, what common foundation ensures the explainability of the mechanisms behind them to auditors and regulators? How can everyday consumers trust these auditors and regulators in their

[19] Notable past attempts at a unifying theory of cognition did not get the attention they deserved (cf. Minsky, 1988; Newell, 1994; Wang, 2007).

[20] I do think there will be a moment or obvious point in time when general machine intelligence springs into existence; this requires a lot of new knowledge, which will emerge in stages (see section *What General Intelligence Must Be Made Of*, p.16). I also hasten to add that general machine intelligence does not solve a *particular* problem or set of real-world problems, and is therefore most likely only to exist in textbooks.

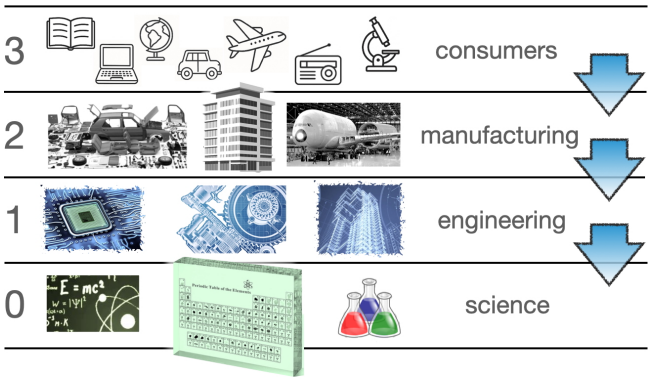

Figure 5: The 'trust' hierarchy. Each stage in design and engineering of products and services depends on a lower, more foundational level (blue arrows) that provides the means for verifying that their operation, in both normal and untested circumstances, is within the bounds required by law and targeted functionality. At the most fundamental level, operation is grounded on relevant scientific theories, but this can only be done if the product's mechanisms are transparent and well understood.

conclusions about a product's safety? The only way to truly ensure this is via *general transparency:* Well understood and easily accessible mechanisms that allow a product's composition and behavior to be characterized, analyzed, and verified against specifications and laws. How can such transparency be achieved? The accepted way to ensure it is through best engineering practices *based on science*. Science rests on empirically verified theories developed over many decades and centuries that supply the ground rules about how things affect one another in the physical world. Since a scientific theory is the *current best available explanation* for any phenomenon that it covers, it is the obvious basis on which to provide a comprehensive map to a product's nature and behavior, even for hypothetical situations. If transparent, a system's failure modes can be analyzed and listed during its development, and remedies proposed and implemented, before it's released.

With the condition of scientific grounding being met, a particular product's operation can be predicted for a variety of uses under a variety of conditions, many of which, if the theories are strong enough, do not even have to be tested in the lab. Equally importantly, *the foundations* for those predictions can be *explained and verified*, even empirically, if need be. For this reason, a system's behavior in both typical and hypothetical untested circumstances can be reliably understood, and the range and limits of the system's intended practical use can be guided away from its failure points, making it trustworthy to its everyday users.

A Short Dive Into Explanation

An 'explanation' in this context – and yes, this is directly relevant to our main storyline here – is a *particular account* of how things of importance to a phenomenon or event hang together; the particularity in question is the singling out of certain things that, had they been different, *could have changed things*. The common term we normally use for things that can change what happens is 'causes.' While the list of causes with a potential to change any given event or phenomenon in the physical world is infinite, several things can fortunately help to narrow it down, and not all causes in a chain of events are relevant for giving a *good explanation*. Here it is useful to remember that every explanation has a purpose – a goal – that drives its form.

The high-level goal of any explanation is to *improve or prove understanding* (Rörbeck, 2022; Thórisson et al., 2023).[21] It does this by highlighting specific information, or pointing out unknown, obscured or misunderstood causes, relations, or other factors. There are three main classes of such information that an explanation may highlight to achieve this purpose (from Thórisson et al., 2023):

- Causal factors and causal chains
- Variables, patterns, and other elements
- Background and context assumptions

To be *good*, an explanation's claim must be testable through some form of action that reveals whether it holds, by referencing measurable, concrete variables; in other words, it must *be validatable* in the physical world—if not practically then at least in principle.[22] In casual everyday human communication, explanations may often be validated via [**a**] thought experiments and common sense (for instance, contradictory explanations of 'whodunnit' will not convince us the detective has figured it out yet); [**b**] empirical measurements, analysis, or experimentation (real-world interaction, e.g. when repeatedly pressing an elevator button when it doesn't light up); or [**c**] a mixture of the two (e.g. when an insurance company must decide whether a car's brake failure can be traced to a manufacturing flaw). All of these use trusted models of cause-effect relations to exclude potential explanations for particular events, situations, and outcomes. And, at

[21] Our definition of 'understanding' follows the theory of pragmatic understanding explicated in Thórisson et al. (2016); referred to by some also as 'comprehension' and 'sense-making' (cf. Klein et al., 2006). By 'proving' is meant empirical (non-axiomatic) verification and validation (subject to the long list of relevant issues from the philosophy of science, cf. Carnap, 1998; Popper, 2005); in other words, not mathematical proof. The 'cause-effect relationships' that define particular outcomes of a course of events are the things that we might consider changing if we want to affect or prevent that particular course of events.

[22] Examples of explanations that are impossible to validate directly or immediately relate for instance to historical events and theories in astrophysics. In both cases, if an explanation is good, evidence for validation or verification may eventually turn up, through luck or hard work, while evidence for repeated attempts at its invalidation fails to show up (Popper, 2005).

the risk of belaboring an obvious point, such models are not statistical, they are causal; causal models subsume statistics because they can be used to generate correlation without additional information (but not vice versa).

Unwarranted Trust In Contemporary AI

Building products in accordance with scientific knowledge guarantees that valid explanations – as defined above – can be provided for their adherence to stated operational requirements and goals. Without good explanations, safe operation of x-ray machines, airplanes, mobile phones, electrical home wiring, and a long list of other things that modern society relies on, could simply not be ensured.

A product does not have to be fully transparent to its customers or end-users, but in principle it must be possible to disclose its potential dangers on demand, along with verified remedies for these, to relevant experts, auditors, and regulators. Should this requirement not be met, it doesn't make any difference whether the product does the right thing most of the time, or even almost all the time. It is also not acceptable that a product's common modes of failure cannot be properly explained, even after they happen—a product's various potential types of failure must be foreseeable *before* it is sold for deployment (cf. Nedunuri et al., 2026).

Such is unfortunately the state of genAI technology (cf. Foukalas, 2025), and other products resting primarily on contemporary artificial neural networks (ANNs): *Good explanations cannot be given about their operation in any instance—neither before nor after they act*. While the general principles of their operation are well understood by their creators (otherwise they could not be created), each instance of output they generate is practically impossible to untangle. Any honest attempt at questioning the output of a particular ANN *scientifically* forces a resort to statistics.[23]

Part of the reason stems from the method of their creation. (NB: This paragraph is a highly compressed account of how ANNs work, intended for those largely unfamiliar with their internals.) ANN-based systems are tuned ("trained") using vast amounts of data ("training data") whose contents and origins are opaque, even to the systems' designers.[24] Their behavior can therefore have unforeseen, undesired, and dangerous side effects (cf. Carver, 2026; Slater, 2025)—even after extensive efforts to curb them (cf. Mahmoud et al., 2026). During their manufacturing, massive networks of interconnected functions ('parameters') are produced from correlations and conditional dependencies extracted from the data. After they have been "baked in" and the system has been launched, the system can be run for its intended purpose; the joint operation of the interconnected parameters producing its behavior. Explaining in simple terms any particular instance of such output, however, is impossible, even after it happens, because the operation of each unit in context is impenetrable – unexplainable to both the system's designers and the system itself.

Whether it is the system itself or its human creators that want to explain system output, to do so requires a meaningful account of the particular information flow on that particular occasion for that particular input. But since each system parameter, while being thoroughly connected (on average) with a large number of other parameters, has no explicit semantic mappings to its surrounding, the input, or to its training data, are no operational "handles" that allow the architecture itself, or its human developers, to separate context from claims, claims from questions, questions from assumptions, colors from sounds, pixels from skintone, and so on, at the appropriate level of detail; no particular trace of information flow through the system is in any way meaningfully separable from any other (cf. Carvalho and Thórisson, 2025). This is the first problem.

The joke in Figure 1 elucidates this fundamental underlying problem of the ANN paradigm. Their uniform information representation forces all content to the lowest common denominator – the atomic elements of the data – completely obscuring its (limited amount of) higher-level structure, leaving instead a whole ocean of a number soup—"gray goo."

The second problem is that without any mechanisms in ANNs for representing causal information explicitly, or causal hierarchy that corresponds to the way things hang together, their number soup cannot – by design – be made to contain reliable and consistent information about how the world works.

> The reason that an ANN's particular response to a particular input cannot be explained semantically (meaningfully, contextually), in light of its input, operation, training data, or how the world works, stems from the fact that semantic explanation requires isolating the particulars of any situation, claim, or context—treating relevant causes separately from irrelevant ones, side-effects from main effects, etc.
>
> As mentioned, a good explanation must reference at least one causal relation that, if changed, would lead to a different outcome. Being 'unexplainable,' then, means that for any given param-

[23] Statistics itself is of course an accepted field of applied mathematics and related philosophical ideas concerning randomness and relationships between measured variables. Statistics is independent of the domain it's used in, and thus doesn't say anything about which variables are of interest, or why, which ones should be measured, which should be controlled for, what to look for in their relationships, what the results mean for the domain, or which variables cause which—it couldn't, because then it would be domain-*de*pendent. All these things must instead derive from our choice of appropriate theories and interpretation of experimental evidence, produced through systematic operations on the world.

[24] Due to size, there is no chance that the systems' creators can check every single piece of data used in such training.

> eter (or subset of parameters) that such a system contains, no regularity or set of rules exists in ANNs – other than the statistics describing the training data – that can be brought to bear on how particular changes to input or system structure have, might, will, or should affect its output. Any attempt at steering the output generation of its baked-in network – whether in structure or content, during or after it 'leaves the baking oven' – will not and cannot be based on any theory, as no such theory exists;[25] instead, it must go the path of blind experimentation.[26]

We encounter the same issue again in efforts to affect the content of e.g. LLMs' output, whether it is to curb racist comments, make them avoid discussing certain topics, or something else. Such attempts have been made on the basis of their method of creation—in other words, attempting to change semantics by way of statistics. But the semantics is what is missing, so statistics are used in its place. This results in attempts to 'make sense of correlation' purely on statistical grounds, which, again, can only address correlation, so the whole approach thus rests on a naïve tautology that cannot work.

When systems like large language models (LLMs) produce non-factual text output, that is, output that goes counter to what we humans know to be facts in the physical world, they are said to be "hallucinating." We – the users – can classify an output as a "hallucination" in light of *our* knowledge of how the world works—what is and isn't possible; what is true and what is false.

Could we rig ANNs to do the same? Could we infuse them with the necessary knowledge for doing so?

For this to be possible in the general case we would need to extract the meaning of the information in question,[27] but that was precisely the original purpose of the system we are trying to control, so again we encounter a circular dependency with an impasse. The large datasets used to train ANN-based systems contain significant regularities that allow the training process to extract and characterize correlational distributions of the data as a whole, of this there is no doubt. Extracting the *meaning* of this data, however, requires explicit knowledge of goals and related cause-effect relations (Pearl, 2001; Thórisson and Talevi, 2024).[28] Prevention of hallucination in generative systems is in principle not achievable within the ANN language-model paradigm, because factual data can only be found outside the model.

Steering the performance of ANN-based systems *in any way* ultimately depends on our ability to engineer into them mechanisms where particular inputs, and particular classes of inputs, reliably produce particular outputs, or particular classes of outputs. But just like the reliable behavior of a Swiss watch, reliability in ANNs must rest on well-known engineering principles referencing cause-effect relations. So, for any data where ground truth cannot be made available, any real ability of ANNs to systematically associate particular parts of its data with known causal factors in the world, evaluating any output's validity with respect to the physical world (by the system itself or by its creators), before the output is generated, using only statistics, is not possible. Hence, in these cases, they cannot know (or be made to know) whether their own output is "hallucinated" or not (cf. Alansari and Luqman, 2026).[29] In other words, no one can say, before it is produced, whether or not a particular ANN's output will turn out to be "hallucinated" or not.

Is there any way to make the output of ANNs more reliable? Yes. Human-designed mechanisms for enabling an ANN-based system to verify their own output can be made to work in cases where all necessary relevant factual information ("ground truth") is made directly available to the system at runtime. An example could be, say, an LLM designed to summarize, organize, and reference the particular contents contained in online material such as books, Web pages, and scientific papers, that is, not what the content *refers to* in the physical world but rather, the medium itself—the actual words, letters, sentences, images, paragraphs, pages, etc., and the actual quantifiable relationships between these.[30] By limiting such systems strategically to a well-defined tasks (and domains) where this is possible, and ensuring good quality training data, ANNs can be put to a number of uses.

[25] And in fact, no such theory *could* exist, because only statistical patterns can be extracted from static data, and statistics alone cannot produce verified directional (causal) information.

[26] The contrast here is between *informed* experimentation and 'blind' experimentation. The former is what science aims to achieve through comparative experiments that systematically intervene on factors hypothesized to be causally related to dependent variables of interest; the latter is largely a theory- and hypothesis-free *random search*—the most expensive way, energy- and time-wise, to explore physical reality (cf. Newell, 1973).

[27] Any question about meaning must address *whose* meaning it is. We go into this topic in *Understanding & Meaning* on p.17.

[28] The training data used for ANN-based systems, whether text, video, audio, images, or all of these, certainly contains both indirect and direct references to well-known, reliable cause-effect relations (cf. Yu et al., 2026). But those relations cannot be extracted and represented explicitly in ANNs, which can only handle correlational data, and correlation cannot say anything specific about cause and effect unless paired with results from comparative experimentation (or explicit, clearly-encoded given formulations). Thus, particular causal principles relevant to any real and hypothetical conditions are neither represented nor representable, and the same goes for access at runtime to the vast majority of the physical reality to which they refer.

[29] In an illuminating incident, police in Utah had to explain why one of their official written reports claimed that a police officer had "turned into a toad." They had used an ANN-based system to produce a text description from body-cam video, which had caught images of a TV in the background playing the movie *"Princess and the Frog"* (Constantino, 2026). The department nevertheless says it intends deploy such a system to "shave off hours" of police weekly office work. From now on, no one can say whether any of Utah police's official reports are telling the truth or not.

[30] This would of course also apply to anything else whose facts can be directly represented and made available at runtime.

For tasks that require an understanding of how the world works, however, runtime availability to necessary data is problematic. For instance, the control of a physical robot in an environment of medium complexity cannot meet the requirement of "ground-truth availability" because it is neither feasible to pre-screen the (practically infinite) situations that the system may face, or the (practically infinite) output it can generate. Having no verified or verifiable representation of reality, there would be no obvious way then to let them compare their output against the world in any way. Also, there is no way to let these systems represent the reliability of the data they harbor, which makes it a risky bet to use them for control. What they needed is [**a**] a way to autonomously do experiments in the real world, [**b**] mechanisms for learning how to verify such experience, [**c**] an ability to model the experience efficiently and effectively in a way that can be flexibly used and revised (later, when stronger contradicting evidence comes along), and [**d**] an ability to represent the trustworthiness of all knowledge acquired this way.

To summarize: Firmly tethered by their statistical foundation, the very design of ANN-based technologies precludes them, whether autonomously or by outside help, from [**a**] modeling the relations between training data and the physical world, [**b**] meaningfully isolating particular parts of their information from other parts, and [**c**] explicitly representing cause-effect relations, all of which are necessary for meaningful manipulation of their content and operational semantics. Their operation cannot thus be certified to meet desired requirements, whether in form or content, unless their knowledge revolves around easily-verifiable facts that can be made accessible to the systems at runtime. However, none of their actions can be foreseen or explained with any reference to their design, certification, or normative internal operation. The larger they are, the worse this problem gets.

The conclusion can therefore only be this:

> ANN-based systems cannot be trusted to take on important tasks in society, including any kind of critical-path task or expertise involving:
> - advice (medical, technical, psychological, political, etc.)
> - infrastructure control (airports, city traffic, surgery, etc.)
> - manufacturing (food, medicine, buildings, etc.),
> - education
> - and many others

Unfounded Predictions About General Intelligence

Many claims about the state of applied R&D in AI have argued that we are very close to developing machines with general intelligence (cf. Todd, 2026). These predictions often allege that the intelligence of such machines will be as good as, or greater than, that of the smartest human—possibly even vastly superior, exhibiting so-called "superintelligence" (cf. Kapoor, 2025).[31] Behind the claims is the assumption – left unmentioned for the most part – that contemporary applied AI is at such an advanced stage of development that pushing it into human-level general-intelligence territory – often referred to as "AGI" [32] – requires only slight adjustments or minor additions.

CEO Claims About General Intelligence

"AGI [artificial general intelligence] will potentially arrive in 2025 or within a few years"
– Sam Altman (OpenAI), *Jan. 6, 2025* [33]

"Superintelligent AI could arrive by 2027"
– Dario Amodei (Anthropic), *Jan. 21, 2025* [34]

"AGI is probably 3 to 5 years away"
– Demis Hassabis (Google), *Jan. 23, 2025* [35]

Figure 6: A sample of claims made by leaders of three tech giants about progress in AI. To gauge the trustworthiness of such predictions we should ask scientists who have spent their career studying intelligence the following question: *Does scientific knowledge and evidence provide the necessary and sufficient foundation for such predictions?*. If the answer is "no" (which it is), the next questions will likely be about how much scientific work is needed to change that. This, my friends, is like asking "how long until we have a cure for cancer?"

How can we judge whether these claims have any merit? We can do so by taking a look at some of the key ingredients of general intelligence, whose evidence has been trickling in from cognitive science and AI now for over six decades (cf. Burnell et al., 2026; Laird and Wray, 2010; Langley et al., 2009; Minsky, 1988; Newell, 1994; Sloman, 2000, 2010; Thórisson, 2012).

[31] A clear definition for what that means is laregely missing.

[32] 'AGI,' standing for "artificial general intelligence," is the name of an international conference (since 2008) and a scientific journal (since 2009). Started by a handful of researchers, the scientific venues were started, and the term coined, because the Workshop Committee of AAAI, the world's largest AI society, rejected their workshop proposal on general intelligence for the AAAI annual conference two years in a row (Pei Wang (2011), personal comm.).

[33] *"Reflections"* (Altman, 2025; paraphrased).

[34] *"Anthropic CEO: More Confident Than Ever That We're 'Very Close' To Powerful AI Capabilities"* (CNBC Television 2025, Jan. 21; paraphrased).

[35] *"Google DeepMind CEO Demis Hassabis: The Path To AGI, LLM Creativity, And Google Smart Glasses"* (Kantrowitz, 2025; paraphrased).

What General Intelligence Must Be Made Of

When tech giants talk about general machine intelligence they imagine that their statistically-based classification technology is much closer to the relevant goalposts than it actually is. Now it is *our* turn to deliberate about some of these goalposts—but with greater clarity and proper scientific grounding. Before we dive in, let me give a short overview of what the following subsections contain.

With the view that the term 'general' in "general machine intelligence" simply means 'more general' than prior AI systems, we start with the topic of *control*, because this is the largest piece of the intelligence puzzle contemporary AI leaves largely unaddressed. This then leads us straight to planning and reasoning, topics that may be familiar to those who took an 'Intro to AI' course a decade or more ago; however, these concepts must be significantly updated from the way they were addressed by the "good old fashioned AI" of last century. From this we move on to understanding—another topic that has been inexplicably ignored in AI research from the very beginning. The process of understanding, also called "comprehension" or "sense-making," lies at the intersection of control and classification—it cannot exist without a unified process that can do both in harmony.

Finally, we close with a short paragraph on learning, because the "machine learning" that contemporary AI offers only captures a small part of that complex process. To properly keep straight the difference between pure hypotheticals and facts, we preface this with a quick refresher about methodology.

Intelligence: Of Matter & Method

We can theorize all we want about hypothetical machines with infinite memory, infinite compute power, infinite knowledge, etc. Such hypothetical machines may be important 'tools for thought' from time to time, but if we want realistic machines that go beyond philosophy and mathematics they must be implementable and exist physically. Implementation means operating on a physical substrate—matter. Giving an AI system's mind a physical form means that it will of course be subject to the limitations of the physical world: *Energy*, *time*, *space*, and *matter*.

It may seem banal to belabor something so obvious, but believe me, much of the current wayward talk and flights of fancy about the possibilities of contemporary AI and its near future would simply not exist if this obvious fact was truthfully observed. Approaching any question about intelligence on purely mathematical terms, while ignoring energy, matter, and time, has lead to enormous misconceptions about what is and isn't possible when it comes to intelligent machines, leading to vastly unrealistic assumptions about how they might inevitably overtake the planet, humanity, and even the universe at large.

Our current scientific understanding of the phenomenon of intelligence as a whole – as opposed to the biological physical substrate of intelligence, i.e. neurons, nerves, and the brain – has no proper mathematico-theoretical foundation.[36] Taking mathematics as the fundamental or sole foundation for speculation about the future of AI is therefore a serious methodological misdirection. Math is for quantities; if you're still trying to figure out what to measure, it cannot help you. Had Darwin done this early in his research the mismatch between scientific method and scientific matter would have completely stopped his theory of evolution in its tracks.[37] Even today, mathematics on its own is pretty far from being a sufficient foundation for evolution research. This is even more obvious in AI.

Control, Planning & Empirical Reasoning

If *thinking is acting on information*, as is a broadly accepted view in AI and cognitive science,[38] then just like following a ready-made plan physically requires sequencing physical operations on the world (including one's body), thinking requires sequencing physical operations on the information manipulation substrate that runs the mind. In nature this may be a brain; in artificial systems this could be any standard (von Neumann) computer. Be that as it may, time and energy not being infinite, to avoid wasting valuable time thinking about useless things, the *management of thought* must be considered a critical part of what we call thinking. In other words, *attention* (used here in the everyday sense, that is, the purposeful higher-level steering of information manipulation processes, more transparently called *resource management)* must be an integral, inseparable part of what intelligence is all about: Whether human or machine, a mind cannot achieve anything worthy of being called 'intelligence' if it cannot steer its own thoughts in sensible ways. Moreover, this steering must be learnable, so that it can improve with experience and adapt to new situations and circumstances, moving towards increasingly better 'ways of thinking.' This is one kind of meta-cognition that we consider *necessary* for general intelligence.

Controlling one's thinking implies a control system that controls itself. How would that work? What does the design specification for reasonably effective and efficient self-control look like? This is a fundamental question that the science of engineered intelligence must properly

[36] Even knowledge of the substrate of the human mind, consisting of the human brain, brain regions, neurons, nerves, neurotransmitters, etc., while being on somewhat stronger grounds in this regard, suffers such fragmentation (cf. Luo, 2021).

[37] Taking just one case in point, the mechanism by which information was carried between parent and offspring, DNA, would not be known until 94 years later (cf. Olby, 1974).

[38] To be clear, this remains a hypothesis until it has been shown that the functional processes involved in thinking can be fully replicated by operations on information. Note also that this hypothesis does not necessarily mean that the *phenomenal experience* of thinking – that is, the experience of being a living, thinking being – can be sufficiently addressed or explained from a purely informatic view; that is a different hypothesis.

answer but which, unfortunately, has not received much attention (no pun intended) in AI.[39]

If we accept attention – that is, endogenous autonomous control of cognitive resources – as a *necessary* (yet not sufficient) feature of intelligent systems, the next step is obvious: Steering anything, be it a boat, a blimp, or a banana thrown across the room, works much better with a bit of planning. Planning, in turn, requires *abduction*—the ability to think "backwards" from future goals (partial world-states) to the present time. This is where we enter into the domain of control with full force: Abduction marks just the beginning of an empirical reasoning process,[40] because the world can also and will also act on its own, without our intervention (cf. Halpern and Pearl, 2013; Mayer and Pirri, 1996). Conversely, to reliably predict how the physical world may behave from a particular starting point (e.g. 'now'), and especially how it *cannot* behave and is *likely and unlikely* to behave, requires refutable (defeasible; cf. Pollock, 2010) *deductive* capabilities: Collecting arguments, based on known causal relations, to produce an (approximate) picture of what the future may look like, in light of good arguments for and against.

Producing plans for goals with measurable future outcomes, and mentally fast-forwarding the present time to that future, with an accompanying ability to judge the reliability of such plans, are fundamental reasoning processes that can only work for the physical world if they are based on *empirical evidence*—that is, the knowledge of how events *can and cannot* unfold. Knowing only the distributions and correlational dependencies between things in a particular environment in a particular configuration is not sufficient for this purpose, because changing any part of such a configuration changes the relations between things and thus the statistical distributions (Pearl, 2001). Rather than trying to compute and store all permutations of all the ways everything in the world may relate to everything else in any configuration beforehand (which is impossible in principle anyway; note the inherent tautology), to get a handle on what is and isn't possible in a wide range of situations we must instead rely on general rules by which interventions change the relations between things—in other words, on *models of cause-effect relations* (Pearl, 2009). So empirical reasoning is thus a necessary (but not sufficient) feature of any kind of intelligence that must handle the physical world. But wait; there's more.

If all our system could do is record events and play them back, whether "backwards" or "forwards," it might as well be a security monitoring system. Clearly, intelligence is more than that. *Induction* is the ability to generalize patterns—not just correlation patterns but also *causal* ones. To be useful, generalizations must be comprehensive and compact. Generalization is at the heart of figuring out how new things learned relate to things we already know. Any brand-new generalization must be reconciled with existing generalizations; conflicts may be tolerated but should preferably be resolved (either by modifying the new generalization, the old generalization(s), or both, or searching for a hypothesized unknown cause for the conflict). From this activity a network of generalizations, structured in an approximate hierarchy of interrelated rules, can be formed and used to guide the abduction and deduction processes. The rules must have an explicit structure, in the sense of being made up of unique and concrete parts, i.e. a particular rule about a particular phenomenon must be identifiable and separable from other (similar) rules, to avoid confusion and to facilitate that it can be used or changed, in light of new goals, new information, identified errors, and ongoing generalizations.

One more thing before we let go of reasoning for general intelligence. The process of analogy – comparing and contrasting different things, ideas, elements, parts, etc. – is important for a variety of purposes, including learning, identifying unexpected causes, judging similarity of things (including for the purpose of verifying the identity of semi-permanent objects), and many others (Gentner and Smith, 2012). We can call any system that unifies the four kinds of reasoning – analogy, deduction, abduction, and induction – in an efficient and effective manner, where each method supports the others in both a planned and opportunistic manner to address the challenges the world presents, an *ampliative reasoning* system (cf. Thórisson, 2020b). Artificial systems capable of ampliative reasoning exist, but they are still mostly at TRL 1-4.

Understanding & Meaning

Whether you see a big boulder rolling your way or you trip and fall on a public sidewalk, living beings make sense of their world by producing the implications of their situations, in light of their own goals. Whether you are embarrassed, afraid, or laugh it off, this is what we call 'understanding' in everyday conversation—the process inside an intelligent agent's mind that produces meaning for itself, in part or whole, in light of any situation (Thórisson and Talevi, 2024). In spite of what some philosophers have tried to argue (cf. Dennett, 2017), the concept of 'making sense' is neither outdated nor vacuous; any attempt that

[39] One answer for how this can be done is demonstrated by the Autonomous Empirical Reasoning System (also called 'Autocatalytic Endogenous Reflective Architecture'; see e.g. Nivel, Thórisson, Dindo, et al. (2013) and www.openAERA.org – *accessed Feb. 26th, 2026*; code available on Github) which my team and I have been developing for the past 20 years (cf. Eberding et al., 2024; Nivel, Thórisson, Steunebrink., et al., 2014b; Nivel, Thórisson, Dindo, et al., 2013; Thórisson, 2012). AERA implements a transversal attention control mechanism that can be subject to learning from experience (Nivel and Thórisson, 2009, 2013a, 2013b; Nivel et al., 2015).

[40] We use the term 'empirical' here in the same way it is used in comparative scientific experiments and empirical fields of science, i.e. 'in relation to the physical world' (the ultimate authority on how the world works is the world itself).

tries to discredit the concept will ultimately bump into the fact that claiming something as 'meaningless' rests on being able to segregate 'meaninglessness' from things with 'meaning,' which in itself of course assumes that 'meaning' has meaning. The argument that "understanding is meaningless" is in other words inherently self-defeating.

Another attack on the concept confuses performance and meaning: Many people feel that software that is capable of driving a car must by definition 'understand' something about driving cars. But *understanding* a task cannot simply mean the ability to perform it, because then any machine that performs according to its specifications would 'understand,' including washing machines, cruise control, and thermostats. (Do washing machines really *understand* cleaning? Do thermostats really *understand* room temperature? I think not). No, to really understand something – whether task, event, or anything else – means not only the ability to achieve goals with respect to it (Thórisson et al., 2016) but also an ability to *explain* any relevant aspects of that something (cf. Thórisson, 2021). 'Understanding' something like an event requires being able to infer how it may have come about, how it could have been avoided, providing valid arguments for why and how it may be good or bad and how it relates to a variety of aims and goals. The ultimate test of understanding a phenomenon is what science strives for: Being able to *recreate* that phenomenon to a level that proves – empirically – that both the phenomenon and its context have been modeled to a level of sophistication that deserves our trust (Thórisson et al., 2016).

Considering how often it is used in everyday discourse, the fact that both the term and concept of 'understanding' makes very few appearances in the research topics of AI throughout its history is quite surprising,[41] especially considering that it rests firmly on another elusive concept: *meaning* (cf. Speaks, 2021),[42] a term that is even more absent from the AI literature. While philosophers have addressed both topics (cf. Gelepithis, 1988), their treatment is seldom if ever sufficiently detailed to help advance AI research (Thórisson and Talevi, 2024). Yet engineering blueprints to these concepts already exist in the AI literature and some already have working demonstrators (cf. Thórisson and Talevi, 2024; Thórisson, 2020b; Wang and Thórisson, 2025).

It is impossible to talk about *trust* in the context of intelligence without talking about understanding; we would not give an important task to a human who doesn't understand it—why should machines be any different? Unless a task is so straight forward that it can be automated with foreseeable reliability using known methods, a purportedly intelligent machine whose understanding of a important complex task is unclear should not be trusted to do it no matter how 'smart' it is. Just as it is impossible to talk about understanding proper without talking about meaning, it is impossible to talk about meaning without talking about *goals*. Luckily, the concept of a 'goal' is one we can use to anchor this whole concept-tower very concretely: 'Goal' is simply an umbrella term for a substate of the world with particular properties and variables with particular values;[43] properties that might be brought about through modification of some of the world's manipulable parameters to *achieve that goal*. To get your coffee cup (a discernible entity; made up of recognizable parts) to be positioned on a table in front of you (another discernible entity whose relation to the cup can be measured), an agent could for instance choose to use its body-arm-hand system (manipulable parameters) to change the cup's future position (another manipulable parameter) such that, as time progresses, that specific substate of the world is verifiably manifested: "Now your 'cup' is 'on' the 'table'."[44] This requires some knowledge of relevant parameters, and a model of how they relate to each other, but also an understanding of how things can go wrong, how to avoid that, and what to do if plans fail.

A robot that can predict the most common ways things can go wrong when attempting to achieve its goals, which requires knowledge of cause-effect relations (and thus cannot be done with simple statistical prediction), is much more than a mere factory robot, it is a knowledge-driven machine that can use ampliative reasoning to reliably derive the preparations and actions needed to bring things about in a reasonably safe manner (cf. Aguado et al., 2021; Thórisson, 2020b). *Meaning* in such a system is created by continuous computation of the implications and potential interferences that an independent world-in-constant-motion might bring about, and how this world responds to interventions that may affect particular goals in particular ways,[45] and what kinds of subgoals are needed (e.g. to check what other robots are doing,

[41] There are only three sub-fields where the term has been used, as far as I can tell: 'Language understanding,' 'image understanding,' and 'scene understanding.' All of them treat the concept as having mainly or exclusively to do with syntax, not semantics, as the terms 'understanding' and 'recognition' are frequently used interchangeably; the latter is of course synonymous with 'classification,' but classification is clearly not the same as understanding.

[42] Philosophers distinguish between *foundational* meaning and *symbolic* meaning (cf. Lewis, 1970). While their differences are very important for AI research, it is not necessary for us to address them in the present discussion.

[43] We limit the discussion here to explicit goals, that is, descriptions of substates of the physical world that are measurable (so that the accomplishment of a goal can be practically verified).

[44] In everyday discourse we sometimes conflate a description of a goal with the world-state itself, because often this difference doesn't matter. In human communication, if a compact description of a goal state is not clear enough, the agent aiming to achieve the goal may misunderstand what the goal description refers to, and therefore fail to achieve the intended goal.

[45] By the term 'particular' here we mean information at the appropriate level of detail, specific enough that it can be said to be meaningfully relevant to the goals, cause-effect relations or other circumstantial information of importance to the outcomes, side-effects, implications, etc. of a situation.

or planning to do) to ensure that it happens without undesired side-effects (cf. Thórisson and Talevi, 2024).

Deepening of understanding is ultimately dependent on experimentation coupled with explanation generation—seeking reasons (preferably *the* reasons) for why existing knowledge cannot explain what we experience: If a lock doesn't open, you do experiments like wiggle the key and shake the handle (hypothesis: *Stuck due to mechanical friction?);* if the car doesn't start when you press start, you try again (hypothesis: *Switch doesn't make contact?);* if the remote control doesn't work, you put in fresh batteries (hypothesis: *Dead batteries?*), etc. A prerequisite to be able to do this kind of manipulation of the information that makes up our knowledge is that it has discernible parts—that it has "handles."

To understand things better – i.e. to create better models of the rules and regularities that steer the phenomena we encounter in the world – we do experiments on the world from a multitude of angles, under a multitude of conditions; logic and empirical reasoning allows the convergence of the resulting data to exclude the majority of invalid models and partial models (Thórisson and Talbot, 2018). Doing this cumulatively over many years allows certain invariants to be modeled sufficiently well to get things done reliably and safely (i.e. with sufficient predictability).

Theories of understanding have been put forth (cf. Thórisson et al., 2016) and a few of them have been implemented in running AI systems (cf. Nivel, Thórisson, Steunebrink, and et al., 2013; Wang, 1995a, 2013). Does that mean that 'artificial understanding' already exists in the lab? To some extent, yes; according to our definition of understanding-as-an-information-process, it can be designated as *real understanding* without any quotes around either of these terms. The approaches mentioned have been done, for the most part, by small groups of researchers in a handful of labs. To move this work from basic research (TRL3-4), which is where it is now, to application (TRL7-8), requires more work by more teams. That kind of work is precisely the purpose of dedicated research institutions like Simula, SINTEF, Alexandra Institute, the Icelandic Institute for Intelligent Machines, and Deutsches Forschnungszentrum für Künstliche Intelligentz (DFKI), many of which have a non-commercial status and extensive networks of close academic partners and industry collaborators.

Autonomous Cumulative Learning

As the world never reveals its whole story all at once, a critical feature of animal and human learning is the ability to improve knowledge incrementally over time. People often think of learning as a single process, but it is more accurate to see it as a set of interlocked complex mechanisms that are both intricate and opaque. We can see some of these by looking at how the field of AI and psychology have pursued it in bits and pieces: Reinforcement learning, lifelong learning, continuous learning, reasoning-based learning, rule-based learning, one-shot learning, transfer learning, and concept learning. Today's science explains these for most part through incomplete micro-theories whose different philosophical and background assumptions make them incompatible and impossible to unify. Acting and perceiving the physical world requires natural agents like squirrels, dogs, humans and birds to be born with all of these in a unified single system that can operate in realtime. However, very little attention has been paid to cumulative learning and the role of time.

Humans are gifted with the additional ability to explicitly isolate cause-effect relations and apply rules to sort out good explanations from bad ones. This form of knowledge representation is necessary to be able to apply reasoning over the knowledge; through this combination, a learner can create rules about rules and rules about groups of rules, which in turn, in a world of regularities like the physical one, enables creation of better knowledge, and faster—in other words, getting better at learning; *learning to learn*. Flexible knowledge construction requires information to be represented in a modular fashion so that it can be conveniently built up out of bits and pieces. The idea of 'compositional' knowledge creation, where new knowledge, concepts, ideas, or cause-effect relations are formed through composition of smaller explicit fundamental building blocks, like a LEGO[46] model being constructed out of LEGO blocks, has a long history in AI (cf. Ashby, 1958; Tafazoli et al., 2024; Thórisson and Nivel, 2009). It is very different from the knowledge representation approach of contemporary 'neuron-inspired' AI.

As already mentioned, a fundamental aspect of learning in the physical world is experimentation. By systematically 'asking questions' of the world, through direct interventions and comparisons, bad explanations can be eliminated over time through logical reasoning, making the best explanations remain to guide our actions. It proceeds something like this: When faced with something new, or an anomaly, or an error, or something that "does not compute" in view of what we already know, our unified learning mechanisms can update our existing knowledge in just the right manner, fixing only that which matters, without changing or messing with the rest. This, in essence, is 'cumulative learning' (cf. Thórisson et al., 2019).

Few systems to date have convincingly demonstrated a capacity for cumulative learning, in part because it requires composite knowledge representation (cf. Wang and Thórisson, 2025), empirical reasoning capabilities (Wang, 1995a, 2025) and a capacity for what we could call 'everyday experimentation' (Thórisson and Talbot, 2018), all of which must be effectively and efficiently unified in a coherent *single cognitive architecture*. Its principles of operation mirror the scientific enterprise's, whose method for applying reasoning and logic to create

[46] "LEGO" is a registered trademark of the LEGO Corporation.

hypotheses, design experiments, and evaluate empirical data, are well-understood.

Self-Reflection, Self-Explanation & Meta-Cognition

We have already touched on some aspects of the topics in this subheading—these are some of the thorniest topics of general intelligence.

Meta-cognition is in principle not difficult to understand; it is simply cognition whose topic is cognition itself—the broad and general concept of 'thinking about thinking.' Self-reflection is a special kind of meta-cognition where the thinking you are thinking about is your own thought processes, in part or in whole; usually the term refers to the *deliberate* ability to point your thoughts to your own thoughts, on queue, like when I tell you to think about how your mind reads these very words—do you hear a voice in your head uttering every word, and if so, whose voice is that?

Why meta-cognition is important can be seen by the analysis in the following paragraphs, starting with the basics: The reason we want intelligent machines is because they are flexible and autonomous; an autonomous system 'gets stuff done' without the need to 'call home.' The more intelligent it is, the more complex goals and environments it can handle. When it encounters something completely new, if it is smart enough, it can teach itself how to do it. Such a machine would be constantly learning new things and get stuff done completely autonomously, with scientifically verifiable guardrails.

Every real-world environment might harbor some unknown dangers. Risk can be seen to have two parts: The particular features of a situation on the one hand, and the abilities of the agent on the other. Taking risks in complex environments is usually not advisable, but may sometimes be unavoidable. In either case, evaluating the level of danger in any situation is essential. In a complex world, for any situation that an autonomous machine may find itself in, risk will not be obvious, if only for the reason that it is *unknown*. To be autonomous, our system must be able to handle a range of such circumstances, and because they were not foreseen by us – the engineers – the control system of the machine may respond to them in new ways—i.e. it will *invent new behaviors* tailored to the new situation. If our machine does this, we should be happy, since that is after all the very definition of autonomy—we have ourselves an autonomous machine. But if we – the system's engineers or users – don't know the exact form of the desired response that our autonomous controller may generate to new situations, how can we be sure that those behaviors are not dangerous, ridiculous, or detrimental to the system itself, or to us?

This problem cannot be solved by giving the machine a Handbook of Risk; what should we put in that book? By definition, we cannot know the unknown risky situations that the machine might encounter. Even if we could populate it with some content, how would the machine know when to use it? It shouldn't use it all the time, only when in risky situations; again, how would it know that? We could consider having the machine 'call home' in risky situations, but for that it would still need to recognize the risk, and estimate its level. The more it calls home, the less autonomous it is: Making the machine consult a handbook or call home defeats the point of having an autonomous machine in the first place.

Here is a solution: We can create a second controller for the machine, system B, that monitors its main controller (we'll call it 'A'). The purpose of A is to learn and act in environment E (A's 'natural environment'); controller B is learning about and managing A; not only to keep it in check but to guide it during its learning. Equally importantly, controller B's role is to learn about the $\{E,A\}$ tuple, allowing the system as a whole to not only make judgments about its environment but to make judgments about the *self-in-environment* tuple. In other words, to judge risk, the machine *must be capable of self-reflection—turning its attention inwards and evaluate its particular experience and knowledge in light of its particular situation.* This is why self-reflection, a kind of meta-cognition, is important for advanced AI systems (Nivel, Thórisson, Dindo, et al., 2013; Nivel et al., 2015; Thórisson, 2012, 2020b).

System B gives the machine a meta-cognitive functionality that allows it to learn about its own capabilities, in light of different kinds of situations, and model risk as a concept, training itself to respond safely and securely. In other words: Self-reflection is necessary for creating safe and trustworthy AI systems.

Are approaches based on such principles failsafe? Do they absolve our machines from all future risk? Of course not; after all, the physical world can be pretty unpredictable (and the purpose of intelligence is to deal with the unknown). This solution is nevertheless the one that nature has instilled in countless species on the planet, to a lesser or greater extent.

In all worlds like the physical one, which don't reveal their principles of operation directly, an autonomous agent can never be sure that it has everything "all figured out." Artificial agents with general intelligence are, just like humans, subject to this real-world fact. But not only that, they must be be able to model this fact in their knowledge. Indeed, a critical part of general intelligence is both knowing *that* you don't know everything, and getting a grip on *what* you actually don't know. Any knowledge that the machine creates must also be defeasible (cf. Pollock, 1991, 2010; Pollock and Cruz, 1999), because if evidence comes in later that goes counter to our 'common sense,' we must be ready to re-evaluate that 'sense.'

This last point concerns knowing what you don't know – an important type of meta-knowledge that can put some estimates on your level of ignorance – and is often overlooked, both in work on AI and discussions of where it's future is headed. Yet it is critical if we want to build highly

autonomous *and* safe future AI systems, because this ability is necessary for judging the quality of one's knowledge, which in turn directly affects the ability to judge risk.

> Any artificially intelligent agent that possesses open-ended action potential in open environments, like animals do, yet is incapable of judging risk, *is by definition a risk to itself and its users*.

A Unified Whole

If what we are after is *more general machine intelligence*, one important aspect of intelligence keeps getting left by the wayside in AI research to date: The fact that any aspects of intelligence the field explores – reasoning, learning, perceiving, remembering, recognizing, explaining, analogizing, experimenting – must all exist in one and the same system for that to be realized. Historically, AI researchers have been very busy tearing natural cognition apart into bits and pieces to study each one separately. This is much more debilitating than you might think at first. A short overview can be given as follows:

A By separating learning into incompatible "silos of study," with names like reinforcement learning, lifelong learning, continuous learning, machine learning (stochastic methods), transfer learning, meta-learning, active learning, and several subcategories of each of them, no machine is as generally capable of learning the way humans learn, and no machine can yet be trusted to go into the physical world and learn reliably on its own.

B By separating reasoning into *deductive, inductive, abductive, analogical, causal*, and *common-sense learning*, and studying each in isolation,[47] no machine yet exists that can do reliable real-world general empirical reasoning like humans, or as consistently and reliably across a wide range of situations, like many other animal species can.

C By segregating perception into vision and speech, leaving out proprioception, touch, or hearing (unless it is about speech) for the most part, AI has yet to generate comprehensive theories of perceptual integration, unified theories of multimodal perception, and how perception and reasoning works together to enable better, more unified learning. Time and learning in perception is mostly left by the wayside in contemporary AI; a comprehensive theory of situated perception is nowhere to be found; 'learning to perceive' is not a topic found in the table of contents of any handbook on AI and neither is the unified 'perception and meaning,' nor 'perception and reasoning.' As a result, the development of robots with diverse trustworthy skills and adaptability is still at a relatively primitive stage.

[47] It is not the separation per se that is the problem but rather the isolation of research on each one from the others, each following different philosophical assumptions, approaches, and methods. This makes their post hoc unification impossible, even after notable progress may have been made.

Theoretical Limitations of Artificial Neural Networks

Betting on contemporary AI as a path towards next-paradigm AI – systems that are less random, less opaque, more explainable, more capable, more intuitive, more useful, etc. – is betting on a technology that more or less ignores two thirds of the processes that are necessary for natural intelligence; I am talking about control on the one hand and learning on the other.

There is a strong tendency among researchers and laypeople alike to think that the shortcomings below – and many others (there are quite a bit more) – must be fixable in principle, through small or large modifications to the paradigm—"after all," they say, "the best examples of intelligence we find in nature all rely on neural networks, so surely an artificial version must be bound for the same ultimate functionality, right?" *Wrong*. There are fundamental theoretical reasons that point categorically and directly to the opposite. These theoretical limitations preclude anyone – whether lone researcher, brilliant academic, entrepreneur bursting with energy, tech giant, or a conglomerate of tech giants – from fixing the situation, no matter how much money or effort is put in, without changing the paradigm completely. Let's take a look at a few of these.

An Unnatural Foundation

There is no question that neural networks exist in nature; the brains of animals use neural networks for control, classification, and learning, and human general intelligence is based (primarily, possibly entirely) on networks of neurons. Yes, 'artificial neural networks' were given this name because they were inspired by nature's neural networks. But this is where their similarity ends; as far as their operation is concerned, they are *very* different.

The artificial neural networks of contemporary AI do not operate like neural networks in nature (cf. Bareš, 2025). The main method for 'training' ANN-based systems, back-propagation, does not exist in nature, has never existed in nature, and *could not* exist in nature: It is entirely unnatural (cf. Morffis, 2025).[48] Unlike natural intelligence, ANN-based systems have to be given all their 'training' data up front; their final form is defined by the statistical data distributions inherent in its corpus. Compared to natural intelligence, which can revise its acquired knowledge on-demand as new evidence comes forth, these systems cannot be made to do that because the information inherent in their frozen

[48] The kind of feedback loops in back-propagation are precluded from being implemented in biological substrates. An extension of ANN pre-training methodology employing direct human feedback for system reinforcement could be considered an approximation to how experience molds knowledge of animals, and reinforcement learning certainly exists in nature. It is, however, the simplest form of natural learning, with well known limitations that are far surpassed by learning from analogies, learning by imitation, and causal learning, and learning where such mechanisms seamlessly support each other.

model works only by exhaustively tracing through the full dataset, mapping out all dependencies found – relevant *and* irrelevant – at 'bake' time.

> The main functionality of ANNs bears no similarity to the workings of natural neural networks, neither in humans nor other animals. Therefore, claiming that ANN-based systems offer a valid path towards general intelligence *because* human intelligence is based on neural networks is *not a valid argument*.

Neither do any claims based on assumptions about modifying the framework to 'better match' the way nature works, as such modifications would have to be vast and deep—so deep that common sense would in every case classify them as being 'completely different' from contemporary AI. At this time in the history of science there is no doubt that any purported similarities between artificial and natural neural networks are purely poetic.

ANN-Based Systems Lack Control

Control is central to any intelligent system; without mechanisms for wielding control, neither time, progress, nor goals can be effectively measured or pursued. The main function of artificial neural networks is classification, which rest on an open-loop input-output mapping architecture.[49] As mentioned in the *Introduction* (p.2), classification is not control, and a predominantly open-loop statistical mapping architecture does not by itself provide mechanisms required for adaptive closed-loop control of external action or internal cognition.

Let's look at the two kinds of control that we discussed in the section on general intelligence (p.16). Firstly, control of physical processes: By and large, ANNs are structured as a pipeline: Input, when provided to the system, flows through the system in one fell swoop, output is produced – a classification – and then the system falls quiescent. Any predominantly open-loop statistical mapping architecture like ANNs does not by itself furnish the mechanisms required for the kind of adaptive closed-loop control of external action, or internal cognition (see below), that is needed for full autonomy and general intelligence. While it is not *impossible* to rig classifiers up in a way that they can be used for control, such systems have significant limitations. This is one of the two main reasons we still don't have, and are unlikely to have, predominantly ANN-steered 'full self-driving' (the other reason is that ANN-based systems don't learn after they leave the lab).

In the early days of so-called 'self-driving cars' the control part of the software was completely hand-coded—'good old-fashioned software engineering.' Eventually manufacturers started experimenting with using ANNs for this part, leading to early optimism that eventually did not pan out (cf. Sifakis, 2022). The reason is simple: Control requires a strong sense of time, not just of time passing of an external (world clock), but also the time it takes to compute things. Can a 'sense of time' be instilled in ANN-based systems? Sure, but it will be a hack that doesn't deliver what you'd hope for. The real solution is to do it like nature does: Unify control and classification in a single coordinated process that works seamlessly as one. This is what is required for better machine intelligence.

Control is not just important for any agent situated in the physical world; it is equally important for steering internal thought processes. Outfitting ANNs with the ability to steer its internal processes is prevented by two major architectural limitations. Firstly, as we established (see section *Understanding & Meaning*, p.17), ANN-based systems have no 'handles' on the information that underlies and produces their behavior. Therefore, they could not possibly know where to look – or even what to do – to change their own behavior, should we want them to. This limitation follows directly from the fact that there is no explicit rule hierarchy in their knowledge representation.

Secondly, they have no internal mechanism that allows them to extract the information necessary to do flexible control, whether from their training data or at runtime. The former is prevented by a domino effect: They have no mechanisms to evaluate the effects of their own actions on the world; this prevents them from identifying cause-effect relations in the world, which in turn are an absolute necessity for any controller. The latter is prevented by their reliance of having to be 'pre-baked' (see *Practical Shortcomings of Generative AI* on page 27).

ANN-Based Systems Cannot Learn

The concept of learning is *of central importance* to the fields of psychology, sociology, neuroscience, zoology, cognitive science, and artificial intelligence, all of which study environmental systematic adaptation in one way or another. All systems that do this do so by committing semi-permanent models of the environment to their memory (cf. Conant and Ashby, 1970). It is therefore strange to see that the last field in that list allows itself to redefine at will how many well-established terms related to the phenomenon of learning are used, in some cases turning accepted meaning entirely on its head. Here we will look at a few ways in which this happens.

In light of how nature implements learning in a host of animals besides humans, the so-called 'learning' in ANN-based systems (Dupoux et al., 2026; Schmidhuber, 2015) and computational reinforcement learning (Sutton and Barto, 1998) leaves a lot to be desired. Compared to nature, what 'they' do to adapt is much closer to the concept of training than it is to learning.[50] Yet the

[49] Technologies called 'agentic AI' systems have handled this with what is called 'tool calling,' whereby special-purpose control software can be invoked on-demand by the ANN. As the control itself is not directly a part of the ANN system, this approach leaves several topics unaddressed, including the central topic of time management. Similarly, representation of space and explainability are left unaddressed in this approach.

[50] I put 'they' in quotes because saying that the adaptation is

terms 'training' and 'learning' tend to be used pretty interchangeably in the AI literature. This innocent-seeming mistake deftly obscures the fact that their type of "learning" is very different from the phenomenon of learning in nature it was inspired by, and has thus for decades promoted these systems to levels of regard that they do not deserve. Instead of the learner acquiring its own knowledge (cheaply), these systems require *everything* to be handed to them on a (very expensive) silver platter up front. Their "learning" is not autonomous. This is a fundamental limitation that cannot be fixed by more research because it is an integral part of how they work (cf. Dupoux and LeCun, 2026).

ANN-based systems demonstrate one very rigid type of adaptation. In the transformers of deep neural networks, calling them 'pre-trained' ('GPT') certainly comes closer than 'learning' to capturing what is actually going on in their creation. It is still very different from the kind of training we see in nature, for instance when a baby squirrel grows up to be a master climber or when a kitten learns to use a litter box. Humans – the only system known to date to be considered possessing 'general intelligence' – learn through a variety of mechanisms unified in a single cognitive architecture where they support each other. The outcome is a very flexible system that, besides having many desirable learning properties where the whole is obviously greater than the sum of the parts, can use explicit learning goals discretionarily to drive and judge progress, deliberately steering the learning towards those goals. Compared to this learning ability, it is difficult to accept claims about the equivalence of training (tuning) of ANN-based systems and the common concept of 'learning' (Thórisson, 2020a).

The key feature of learning in biological systems came about due to the ever-changing and infinite variety of the physical world. Since events in this world do not happen all at once, and are not observable all at once, our learning must be cumulative—that is, it must build up increasingly good models of the world over time, as evidence about its nature trickles in. In the physical world, knowing everything that matters up front is impossible because it would require infinite memory. If research is going to deepen our understanding of intelligent systems, this seems to be an unavoidable background assumption, and if so, providing all knowledge up front is not an option. As we already went through to some extent (p.19), this fundamental feature of intelligence is nowhere to be found in contemporary AI systems and cannot be retrofitted into them without a complete redesign (cf. Bastian, 2026; Sullivan, 2026).

ANN-Based Systems Can Neither Think Nor Plan

While there is no doubt that *prediction* is an important feature of how the human mind works, we can be equally certain that it is far from being the whole story. To illustrate, let's look at an important functionality of the human mind, *'coming up with a decent plan.'* Such activity requires knowing an *end state*, commonly referred to as "the goal" of the plan (otherwise we could now know when the plan is finished). Given a particular end-goal that has not yet come true (the reason for the plan's existence), and that we have decided to work on now, we must propose actions that will change the current state of the world (not the whole world, mind you, just some relevant subset) to the desired state of the world that matches the end-goal. Every plan has a time-limit, which puts constraints on the possible solutions, and there are usually other constraints as well (energy, for sure). But the most important part of such cognitive work lies in the word "decent"—what goes into making a "decent" plan? It's this: A good plan has *reasons* for *why* it is good and what its weaknesses and strengths are, which any good planner, whether human or machine, should be able to identify and articulate. Every good planner would also know about *its own* weaknesses and strengths, and be able to take these into account when articulating the weaknesses and strengths of a plan they come up with. All of this, in a good planner, will be considered in light of the given deadlines for the plan, its required time for execution, and the ability to resolve any open questions in time. Of course, humans do this via a process of *thinking*.

As thinking beings we know from experience that thinking can be a rather varied process. Many of our thoughts are directly (consciously) accessible via self-reflection, although we can only see parts of the mind. When we *think*, to use a metaphor, and travel through an information landscape where reasons, arguments, logic, intuition, memories, assumptions, and a variety of other curiosities pop up and get our consideration, we are engaging in an active process using a wide variety of cognitive mechanisms and mental phenomena, some of which are under our voluntary control, some of which are less so, and yet others that are completely automatic. Such interwoven variety of cognitive operations, steered by intentions, wishes, and goals, is nowhere to be found in contemporary AI systems—they are built as a strict pipeline. When they run, they run from beginning to end, in one go. Any attempt at making them into "agents" or get them to interact with the world in realtime must be done through ad-hoc methods, usually involving a lot of hand-coded control software.

Any planning that ANN-based systems seem to do, e.g. when producing text that outlines a plan for baking a cake or taking a trip, happens via a process that is entirely different from how good automatic planners and humans plan (cf. Fikes and Nilsson, 1971; Kuperwajs et al., 2023). Instead of using reasoning to resolve mutually ex-

theirs implies more autonomy than is justified; it really is the system's human creators that adapt these systems to the chosen training data; after that they work like a factory conveyor belt, producing output from input.

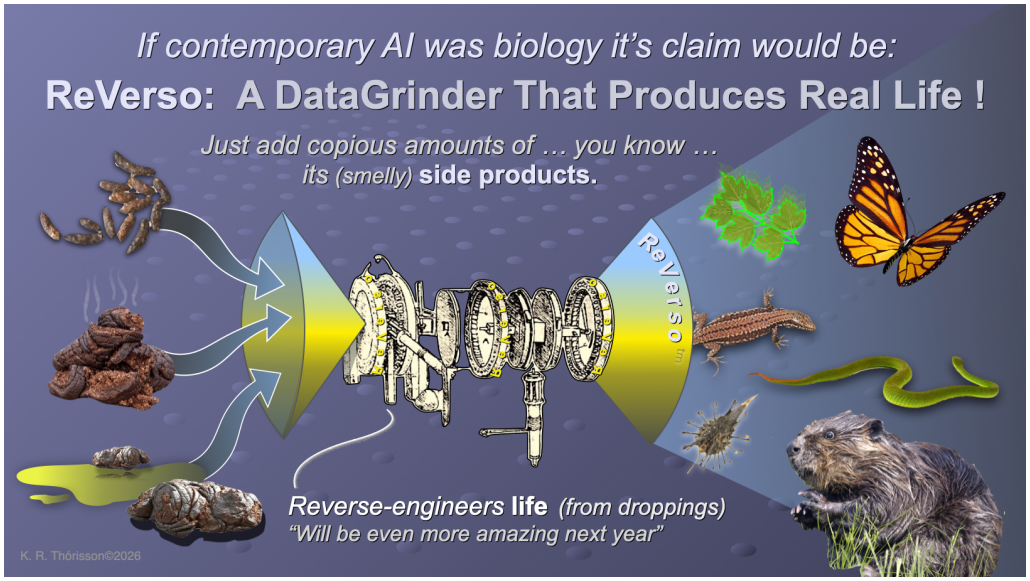

Figure 7: The concept of life and the concept of cognition both refer to complex phenomena that are actively being researched by scientists. Both fields lack a commonly-accepted scientific definition of their main subject matter. In AI this has had the side-effect of wild claims about what contemporary (ANN-based) AI technology encompasses, for instance that it really 'understands,' 'thinks,' and 'reasons.' Comparing this to biology, a similar claim would mean that the tech giants have implemented algorithms that only require the side products of life to produce life.

clusive options and select alternative choices, ANN-based systems use existing examples from their training data whose form makes them *seem* relevant for addressing each task (Sharma and Chopra, 2026). It is the same difference as between a real poet that writes poems about her experiences and someone imitating a poet who has memorized a lot of poems by others and, when prompted to write one from scratch, uses one or more of these as a direct template. When all you get is the output, it may seem like magic, but the foundation of how the process works is nothing like the real thing. Therefore, when such a process is pushed, when it's prodded, and when it breaks, it behaves *very* differently from the process it is supposed to be representing. It is, at best, a truly novel form of imitation (cf. Sharma and Chopra, 2026; Shojaee et al., 2025).

ANN-Based Systems Cannot Meaningfully Self-Reflect

As we discussed in the section on reflection (p.20), a good explanation gets to the point of the matter by highlighting relevant mechanisms involved and isolates what matters from that which doesn't. This is not possible to do in an ANN-based system due to how their information is represented. This is what makes the answer in Figure 1 (front page) funny: The attempt to provide a meaningful (semantic) explanation for a particular output of a large language model contains everything that a good explanation is *not*; in fact, for the operation of such a large system, no good explanation can be given (cf. Nannini, 2024).

ANN-Based Systems Do Not Reason or Understand

Given some input, contemporary AI systems like large language models (LLMs) produce their output by predicting what might come next in a sequence of things, be it words, letters, or other tokens. However, due to how they store their information, for any particular output they generate they cannot produce a valid – reasoned – explanations for *why* or *how* they did it (Sharma and Chopra, 2026). When prompted for explanation of their output, LLMs often produce what may linguistically seems like a valid reply, giving the impression that the explanation is the result of cognitive reasoning processes, when the reality is that the "explanation" is a statistical pattern generated from chopping up and mimicking human-made text (Durt and Fuchs, 2024; Sharma and Chopra, 2026). However, a good explanation of any particular thing refers to at least one cause-effect relation that, if modified, would have removed or changed that particular thing in a relevant manner (Carvalho and Thórisson, 2025; Thórisson, 2020b).

Contemporary AI cannot explain anything on the basis of proper, consistent logic. It can only work if the system, e.g. an LLM, has examples in its database of a structure that happens to match what a legitimate reasoning process *might* produce, when done right. This is imitation in its most advanced form. Valid reasoning is the application of logic to a set of questions or claims under particular assumptions; if those assumptions are broken or abandoned before the reasoning is over, the reasoning is by

definition invalidated. These principles are not followed by ANNs; instead, their output is derived from its *correlation with the surface structure of the input text* or other data in their training set, not due to it relating in a logical way or causally to the nature of the request. For any input, their output is thus *likely* to result in something that *looks like* an explanation, but the output gets produced without traditional evidence or arguments behind it.

Considering ANN-based systems as harboring intelligence of some kind is premised on the idea that [**a**] taking the side-products of general intelligence, i.e. content produced by human thought, and [**b**] feeding this into a data-grinding mechanism (training process, e.g. back-propagation; cf. Erb, 1993), where billions of parameters are tuned based on statistical patterns found in the content, [**c**] will produce the very processes responsible for the creation of that content. How on earth is this supposed to happen? To see how absurd the idea is, let's compare it to the concept of life: Let's imagine some folks who claim they came up with a pretty magical algorithm that, when fed solely with animal droppings found in the woods – assuming it's a *LOT* of droppings – will reverse-engineer and re-create the very spark-of-life processes that made the animals that produced those droppings come alive (Figure 7). Would you believe them? Say what you will about modern artificial neural networks, but are they really *that* magical?!

If we assume that true understanding can be verified (or *falsified*; Popper, 2005) through a logically consistent application of valid arguments to a set of valid premises, and the apparent explanations of ANNs are neither the result of applying particular rules to a set of given assumptions nor based on generalized causal relations between the concepts in question, their validity cannot even in principle be verified – by its creators or the system itself – to have consistent logical validity. If these are the fundamental principles behind the creation and runtime behavior of ANNs (and yes, this is the case), then just like imitation crab meat that is devoid of real crab, contemporary AI systems *cannot be said to understand*, in any conventional sense of the concept.

ANN-Based Systems Cannot Do Empirical Reasoning

'Empirical reasoning' is the process of applying logic in the physical world. The process is essential to science, and underlies the scientific method (Hepburn and Andersen, 2026), but every human uses it every day to get through this thing called 'life.' To be successful, reasoning must be practical, and to be practical, it must be effective and efficient. Due to the infinite combinatorics of the real world, all such reasoning is inherently uncertain; a large part of intelligence is coming to grips with how to handle that uncertainty, reducing it to manageable levels, and getting things done, which, due again to the world's infinite novelty (no situation is exactly the same as when it happened last time), always involves some level of *figuring out how to get new stuff done* (and by "new" we mean new to the agent doing it; Thórisson, 2020a).

Figuring out *new* stuff means doing some experiments on the world; but by 'experiments' we don't mean proper scientific comparative experiments; that would be way too cumbersome—depending on the task and hand, even simple things like pushing or pulling on a door to see which way it opens is an example of an experiment that may be needed. Once the results of such experiments have been acquired, they must be reconciled with what is already known. In our example of the door, this could be that the door is actually locked, from which it can be inferred that no amount of gentle pulling or pushing will do. Doing experiments requires, in turn, the ability to [**a**] consider alternative potential explanations of observed effects, [**b**] coming up with interventions that can help select between incompatible alternative explanations of particular things the world, [**c**] executing such interventions, [**d**] observing the outcome, and [**e**] reconciling it with what is known, for a final conclusion (for that particular case). Note that what we have just outlined here is a control process. Outfitting any ANN-based systems directly with any such mechanism would be a complete redesign of the system—so different, in fact, that it would be a different system altogether. Besides, their inability to self-reflect would still prevent any new information from being unified with its existing information.

The ARC-AGI[51] test was invented exactly to demonstrate this (Chollet, 2019). Consisting of completely novel yet relatively simple tasks, it presents dynamic arcade-style puzzles that come with no instructions for how to solve them. A player must figure out the rules from how each puzzle reacts to various actions, hypothesize what the "tricks" are for solving it, apply these and hope for the best. Currently, ANN-based systems achieve below 1% on ARC-AGI-3 puzzles;[52] humans reliably score 100% (ARC Prize Foundation, 2026).

Good causal models of how the basics of the world work are essential to allow an intelligent agent to make realistic plans (that actually work—and verifiably so), make correctable predictions (when details are missing or assumptions are wrong), and find verifiable explanations for things (using known causal relations to connect partial facts). This last one is possibly the most-important-yet-most-ignored principle in research on general intelligence (Halpern and Pearl, 2013; Mitchell et al., 1986; Thórisson et al., 2023; Thórisson, 2021).

For any learner in the physical world, creating increasingly better models of experience is of central concern (Thórisson and Talbot, 2018) because the specifics details of the world operates can only be revealed by interacting with it. Systematic modeling of hypothesized causal relations is exactly what it sounds like: The creation of hypothesized rules that *"A leads to B under con-*

[51] https://arcprize.org/arc-agi
[52] https://arcprize.org/leaderboard

ditions C, other things being equal." Such models must be conditional on the evidence that backs them; the more strongly particular experience backs certain causal models, the more they should be trusted. For any particular situation, should an agent want to make a plan, or predict, or explain, a particular course of events, it can use its knowledge about conditions *C* to determine which causal models to trust for the job. Any action, prediction, and explanation can thus be customized for the actual situation and goals in question; if the stakes are high, go with the models you trust the most; if the situation is playful and no harm can be done, feel free to experiment with other models.

An equally important aspect of trustworthy knowledge that is closely related to causal models is the ability to separate bad knowledge from good knowledge, and relevant knowledge from irrelevant knowledge—the central bane of contemporary AI. Such judgment can only be done via a gradient over causal models that quantifies – at least to some extent (doesn't have to be perfect, but it must be useful) – the evidence behind the quality of each one. This evidence, in turn, is a meta-epistemological measure that also must have an empirical basis, which can only be created on the basis of verified – or at least *verifiable* – actual evidence.

So this is what is then meant when we say that an ANN-based systems' application of its own stored information is 'ungrounded': No claims can be verified whether physically or *in mente*, by the system or its creators, because no verified or verifiable cause-effect models of any kind exist for the *contents* of the information.

ANN-Based System's Weak Theoretical Basis

Unless you think the way in which neurons enable intelligence cannot be done any other way – something that I seriously doubt – all questions about intelligence are questions about information. In the pursuit of the principles of intelligence, as opposed to how neurons enable it, neural substrates are just as much a distraction as transistors when speculating about how a computer game works. While most scientists assume that we'll eventually be able to explain how nature implements intelligence in neural substrates, such an account will not be a theory of intelligence per se, but rather, a theory of how nature implements *the process of intelligence* with biological neurons. The best way to answer a scientific question is to *directly work on the question in question*—not some other (*seemingly* related) question in the hope that an answer will eventually pop out. The question in question is then, *what is the process that turns information into intelligence?*

Yet cognitive science, and now to a large extent AI, have been overtaken by ideas from neuroscience. Studying neurons as a way to understand intelligence feels like researching sand grains to figure out how to build skyscrapers. (Sand is not irrelevant, but if it's really skyscrapers you want, it surely is the long way around.) When science is faced with this kind of a situation – getting stuck studying a phenomenon's substrate rather than the phenomenon itself – a proper theory is clearly missing; one that can explain when, why, and how the observed phenomenon of interest emerges, what its limitations are, and whether – and how – it can be created in the lab, if at all (cf. Anderson et al., 2004; Newell, 1994).

Scientific research focused explicitly on intelligence has, if we include the field of cybernetics (as seems proper to do), been ongoing for about 100 years. Historically, this is not a very long time for a brand new field of science. The purpose of science is to create not just new knowledge, but *good* new knowledge. The purpose of new technology, in contrast, is to offer new means of getting the things done that people want to see happen. Technology is not theory; theory is not technology; one cannot be swapped out for the other—they are like your two legs; together they move the body of human knowledge forward.

As we have uncovered throughout the journey so far, the only scientific basis on which ANN-based systems rest is *statistics*. Statistics cannot provide a proper foundation for knowledge about goals, means-ends, causal relations, time, plans, systematic dissection and analysis, or comparison of weak and strong arguments, reasoning, abstraction, concepts … the list is very long.

Intermediate Conclusion

The technical and theoretical limitations of ANNs severely degrade the trustworthiness of contemporary AI products and services that rely on it.

While it is clear that AI as a field has crossed a threshold of usefulness, and that it will continue to advance, it is also clear that continued work in the innovation echo chamber, massaging contemporary AI into a variety of sorts and shapes, will produce diminishing returns unless fundamental advances from basic research are regularly added to the mix. As we have now presented plenty of arguments for, thinking of contemporary AI –

[52] BBC Nightline (2024, Oct. 8); *YouTube link, time: 00:01:05*.

and this goes for all known kinds of statically-based AI methods including computational reinforcement learning – as an obvious and guaranteed stepping stone for investors towards next-level AI is a serious fallacy. In that respect – that is, as a stepping stone – they already are, and will remain, exactly the opposite: An expensive detour where rapture seems forever just around the corner for a very good reason: Lack of a solid scientific foundation.

Practical Shortcomings of Generative AI

Out of a long list of items that put practical limitation on contemporary AI, we will focus on three that are associated with the subset called 'generative AI' (genAI): [**a**] Dependence on human-created material, [**b**] dependence on vast amounts of compute, and [**c**] an inability to improve the information in such systems after they leave the lab. These are all a direct side effect of the theoretical limitations of its underlying technology (artificial neural networks; see preceding section, p.21).

"Only Way To Make Them Is To Bake Them"

We have already mentioned this limitation several times. The fundamental information systems based on an artificial neural network (ANN) paradigm does not change after they leave the lab, because the trajectory of the resulting changes, should we decide to set them up that way, is undefinable (cf. Steunebrink et al., 2016). So they come (mostly) 'pre-baked' from the outset because this is the only way to make them. Even worse, this requirement of being 'pre-baked' is *inherent to their very design:* No amount of new extensions, innovations, or scaling up of the paradigm itself can completely remove this problem because their entire premise of operation rests exclusively on the principle of *statistics extracted from the entirety of their 'training data*.'[53] All of the data that an ANN-based system should know about must thus be available up front, before it goes into the oven. If the data changes after it leaves the oven, a whole new system must be 'baked' from the new data—not added to the cake baked earlier but rather, used to bake a whole new, different cake. It may be *very similar* to the last cake, or it might be *very different*—after all, the recipe has been changed; we know it will not taste exactly the same as before, but predicting which parts of it will taste differently is impossible because no theory exists that can explain how various ingredients may change it.

Now, the only reason to make a new ANN would be if the first one didn't operate in accordance with stated requirements. This is always the case because a complete practical or even theoretical analysis and prediction of their completeness is impossible to do. Subsequent changes in training data intended to mend a newly discovered flaw can only be made with *the hope* that would allow it to meet them because, besides statistics, no well-grounded scientific theory exists that can help with that process.

The training set, and the various tests for verifying that the requirements are met by the new ANN, must be created by human engineers. This is an exceedingly expensive process, not the least for a shortage of good theories to guide such work.

Increasing Demands for Big High-Quality Data

Without being fed data up front, contemporary AI systems cannot do anything. The more data they are given, the better they work, other things being equal. But the data must be high-quality. The only way to ensure this is to use well-crafted human-authored data, for instance text by professionals. The most obvious source of such data is the Internet.[54]

To train large language models from scratch, researchers have tried to find ways to generate good-quality data artificially, e.g. by using separate ANNs to generate it, but the results have shown that, unlike when adding human-created high-quality data, doing so always decreases performance (cf. Shumailov et al., 2025; Williams, 2026). This is of course due to how they operate: Finding correlations between given data—without good data the result can only be "garbage in, garbage out."[55]

Unfortunately, no matter how good their training data is, no ANN-based system has so far reached perfect performance, not just because data from the Internet is not perfect but more importantly, because of their mathematico-technical foundation; their operation is in no way different in any important way from a mechanical clock ticking away in accordance with the way its cogwheels are organized. They cannot and do not contemplate, reflect on, or modify their own knowledge, nor can they reconsider the conditions under which the data was produced; in other words, they are not capable of meta-cognition, which is necessary to detect contradictions and faulty information that has persisted after their training.

[53] Many methods have been developed to get around this problem, such as deep reinforcement learning (cf. Bondre et al., 2024), but they all come with their own serious drawbacks (e.g. inheriting the limitations of reinforcement learning). No matter which methods are brought to the table, such extensions depend on allonomic engineering (cf. Thórisson, 2012) that, while they may diminish ANN's dependence on pre-baking and prepared content to some extent, they do not remove their dependence on pure correlation extracted from massive amounts of data.

[54] Since the amount of data sourced from the Internet to train genAI is gigantic, it is not feasible to carefully control its quality, as this would require precisely the very goal of such training – general intelligence – before the training starts. It is thus practically impossible to prevent bad data from hiding in the content used for training, which is one source of the many unwanted side-effects of such systems' behavior.

[55] Projects like DeepMind's AlphaFold (cf. EBI 2024) stand as a clear and obvious exception to this, where the high-quality training data makes the effort clearly worthwhile.

Increasing Demands for Compute (and Unclear Benefits)

It is a simple deductive step to see that with increasing demands for big data come increasing demands for the availability of computing power, with the inevitable and significant negative environmental impact (cf. Vallentin et al., 2023; Zeve, 2025). A recent meta-study looking at a cross-section of Danish workers found no notable impact of using ANN-based technologies at work (Salvaggio, 2025). Similar results have been found in many other studies across the globe (cf. Challapally et al., 2025; Rogelberg, 2026). MIT researchers conclude that *"Despite $30–40 billion in enterprise investment into GenAI [...] 95% of organizations are getting zero return"* (Challapally et al., 2025, p.3). This does not mean it is impossible to get a return, but it seems to show that how to make this happen is not obvious. It is difficult to see the justification for the enormous investment expenditure seen in genAI technologies, given their negative environmental impact and growing evidence questioning their benefits.

Imitating Humans: Falsifying the UX

Among the most important principles in designing good user experience (UX) is the rule of 'functionality at a glance' whereby a dial, buttons, control parameters, or display instantly conveys its function and meaning at first glance. Avoid misleading the user. With good design the functionality of an effective tool can be laid bare, which directly translates to a great tool. Hidden or unclear functionality sends misleading messages, confusing the user, resulting in a bad user experience, which increases risk of delays, operator errors, and misuse of the technology.

This is what the tech giants have done in design of their chatbots: By copying the way humans talk and write, these systems imply that they share cognitive features with humans. By falsifying their underlying operational principles they mislead users into believing that they actually understand what they are saying (cf. Dickson, 2024; Golan et al., 2023; National Science Foundation, 2023; Rahmanzadehgervi et al., 2024). But the impact of this deception runs deeper than we may think: By copying behaviors that people have only experienced from other thinking beings, the trained safeguards that help us see through insincere behavior in others is masked.

Since nothing under the hood of modern chatbots is anything like what goes on in people's minds, but their behavior successfully – and deliberately – makes it seem so, the user experience of interacting with ANN-based chatbots is a falsification: Designing them to look like people makes them seem much smarter than they actually are, instilling a false sense of trust towards them (cf. Gefen et al., 2025). This faulty imitation of humans and human skills lays bare an autocratic view of society and human labor through the oversimplifying lens of wealth-creation, a view that is inherently incompatible with any egalitarian, just society.

Bad as this is, worse is yet to come.

Threats of Contemporary AI

The world is under threat from many directions—food, energy, climate, war. If you think that contemporary AI can save us from whichever of these is most responsible for keeping some of us awake at night, I have really bad news: It won't, and – to be more specific – it *cannot do so*, not in its current form. Quite the contrary, not only is contemporary AI deeply flawed technologically, it is actually a direct threat to modern society (cf. Brennan et al., 2025). There are already plenty of examples showing how contemporary AI can be used to amplify the acts of bad actors, from Brexit (cf. Adams, 2018; Cadwalladr and Graham-Harrison, 2018; Rawnsley, 2018) to China spying on citizens (cf. Feldstein, 2022; Sutherland and Chakrabarti, 2024) to the Switzerland-Palantir story (cf. Fichter et al., 2025, 2026) and its use for cyber-attacks (cf. Bose, 2026).

For the first time in history, practical applied AI technologies have delivered machines that can mass-produce pattern classification operations over real-world data in a way that before could only be done by human cognition. Like any technology, this is a double-edged sword, but this sword is asymmetrical with one of the edges being jagged and weighing more than the other: Due to [**a**] opaque and obscure information representation, [**b**] impossibility of retrofitting self-control, exogenously and endogenously, and [**c**] dependence on massive amounts of data and compute, contemporary AI technology has a built-in bias that tilts towards threatening well-working, well-regulated democracies.

Contemporary AI systems have a built-in resistance to many of the traditional activities currently used to regulate other technologies, making them more impervious to actions aimed at curbing nefarious and illegal activities. The resistance comes from two main sources, both of which we already discussed to some extent. Firstly, because they require big data and big compute, the largest ones with greatest impact are operated by actors with the biggest money. Such actors are often international, having arranged their operations across the globe in a way that, from a legal perspective, maximizes their own autonomy and immunity (cf. Castro and Seabrooke, 2026; Saputra, 2025). Any use of contemporary AI by such actors, for whatever purpose, is protected by a layer of international legal defense mechanisms and layered jurisdictions, making access and oversight by regulatory bodies all but impossible without a matching set of legal arsenal on the other side. Another line of defense of contemporary AI technology is much more obscure to the uninitiated: Because the information representation inside these systems completely hides the sources used for their training – illegally obtained copyrighted materials, for instance – any attempt to

follow trail of such illegal use back to the sources is like tracing the footprints of a suspected killer vanishing into a lake; ANN-based systems create a void—a "black hole" in the operation of any system they are part of, where the trail abruptly and permanently ends.[56]

This situation opens the door to many problems that threaten modern societies. It takes many consecutive decades, sometimes centuries, of hard work to build up tradition, trust, and transparency to the level that the Nordic countries have achieved. It can take less than a decade to break it all down. Let's take a brief look at the many ways in which this may happen – and *is* already happening – with contemporary AI.

Legal & Illegal Mass-Surveillance

For anyone wanting to track the activity of humans before the advent of the Internet, another human would have had to be hired as a spy. With the advent of networked computers, some of the information useful for activity tracking became available in digital form; timestamped online purchases, browsing history, posted content, etc. – the sheer amount of data available before the advent of contemporary AI was already staggering. But spying still used to require other humans to collect the data, combine it, interpret it, and summarize it, and with the amount of irrelevant data, any targeted information would be expensive to find.

With the advent of ANN-based technologies, that obstacle is gone, and with it the old world of "security through obscurity." The cost of fishing out targeted data from a massive amount of available digitized content containing indicators and correlates of human behavior has been sufficiently lowered to make it practically accessible to anyone with a sufficiently big bank account and bad intents. Nation states can now set up mass surveillance on the basis of automated processes supercharged by contemporary AI (cf. Anand et al., 2024; Düz, 2025; Feldstein, 2022; Wilding and Hymas, 2025). While ANN-based technologies perfectly meet these needs and purposes, and accessing the necessary technologies is getting increasingly easier, these systems are full of errors and produce incorrect results at a high percentage. To anyone intent on breaking the law, this is nevertheless of least concern; even large numbers of errors and mistaken identities will not deter its usage.

The Rise of Disinformation

A necessary aspect of democracy is free and fair voting. Successful voting relies in turn on informed choice, while being informed rests on two critical pillars: The access of voters to quality information and the nature of their education (cf. Baça, 2023). Access to quality information is a function of a free press; education is essentially a long-term function of any society. A disruption or dip in either of these pillars over long time periods will distort voting behavior that is likely to result in outcomes that go counter to what the majority wants or what is measurably a better choice on dimensions of importance (cf. Raffio, 2024).

With social media, which is recently being supercharged through increasingly strategic deployment of contemporary AI technologies, getting access to quality information is becoming increasingly difficult, as the information sources and their delivery becomes not only subject to advertising forces but also to noise pollution (AI slop*!* [57]), fragmentation, and deliberate disinformation.

David & Goliath: Small-Data vs. Big-Data Nations

Due to its dependence on enormous amounts of data and compute power, smaller nations are hard-pressed to compete on contemporary AI. Having worked closely with Statistics Iceland on automation opportunity analysis, IIIM is well aware of how small the datasets for Iceland are, even the largest ones going back to 1970; for any compound parameters of interest, the data is most of the time well below the 5000 examples needed for just modest genAI training.

In contrast, big-data countries that are at the forefront of contemporary AI, China and USA, have access to enormous amounts of digitized content, bank networks, and other sources of data and data streams.

Because genAI technology will always be bound by a need for access to larger amounts of data, small-data nations simply cannot compete with larger ones on this technology. Due to this power imbalance, small nations cannot succumb to the lure of contemporary AI without succumbing to the risk of technofascism (cf. Coeckelbergh, 2026). Even if they might be able to afford hosting government data on their own soil, they may be forced to outsource other massive computational needs that their industry, community, investors and operatives require, to larger data centers outside their borders.

National Sovereignty & Digital Banana Republics

With contemporary AI increasingly being used to control social media, news outlets, communication networks, online marketplaces, etc., smaller nations, most of which have a small data footprint, have little choice but to outsource control of large parts of their culture to overseas tech giants. These giants are in a position that small nations cannot be and never will be: They can relate their data to that of other nations, whose data they also have free access to, and increase its active footprint. For very small nations like Iceland, this is the *only* way to make its data useful for contemporary AI systems.

[56] Even if illegal training materials were to be found on company servers, a company may still have plausible deniability to claims that such materials were used for training deployed systems, since proving that they actually were is practically impossible.

[57] See e.g. "*The Cost of AI Slop Could Cause a Rethink That Shakes the Global Economy in 2026*" (Stewart, 2026).

What does online data security mean if the digital footprint of a large percentage of the country's citizens is inspectable by overseas tech giants moments after the data is created? By combining data sources containing individual bank transfers, online purchasing, communication patterns, location data, and world events, patterns can be extracted by these technologies that accurately predict intent, actions, and whereabouts of high-ranking officials. Should smaller nations continue on the path of outsourcing digital and AI services it is relatively easy to see how national sovereignty could become compromised over time, as increasing amounts of services sliding from their control to different jurisdictions, eventually creating what amounts to digital banana republics.

Contemporary AI: A Threat to the Nordic Way of Life

There is little contention about the *potential* of AI to be useful, in principle—it is the systematic and scientific pursuit of automating thought. However, combining the above features of ANN-based technology, we arrive at inevitable conclusions that should scare anyone who cares about democracy, individual privacy, and human rights: The individual freedoms, high societal trust, and the rule of law that underpins our Nordic way of life, that frequently results in self-reported happiest nations on the planet (cf. Helliwell et al., 2026), is under direct threat from contemporary AI. Its lopsidedness plays directly into the hands of those who wish to break the law and destroy; those who want to use it to build and uphold the law are left with a blunt tool whose use for good is almost entirely drowned in a bloody pool of fakeness—fake news, fake photos, fake emails, fake phone calls, fake long-distance girlfriend/boyfriend, fake therapists, fake investment opportunities, fake expert advisors ... the list goes on and on.

Like all powerful knowledge, AI (contemporary and future) can be used to build technologies with a variety of different outcomes. The path we are currently on goes directly against many of the pillars that we revere in the Nordic countries: Human rights, democracy, freedom of speech, freedom of choice, personal privacy, equality, equal opportunity, etc. These principles are increasingly difficult to uphold with the current international structure around AI technology. To be able to uphold the law of the land, we must have transparency in how the technology works, what data is used to create it, and how it affects society in small and large matters. Without a proper scientific foundation, this is very difficult to do. The present trend of outsourcing automation with AI makes the problem even worse. Both must – *and can* – be actively counteracted.

The successful implementation of a mixed economy and pursuit of human rights and equality in the Nordic countries, Canada, and elsewhere, has resulted in well-balanced societies in terms of human living, work, and leisure, free of the many extremes that haunt other societies. Its nature should be reflected in the kind of AI we seek to develop and use.

In addition to a transparent AI technology with appropriate guardrails, we need to be clear in our vision. The basic vision outlined below is not only practical, it is *necessary*—in one form or another. Before we dive in we must rid our minds of some common misconceptions that could perhaps be misinterpreted to be standing in its way.

Refuted: Misconceptions About Contemporary AI

Now we can refute the claims listed in the section *Inflated Claims of Contemporary AI* (p.4) that we categorically said to be mistaken.

> *Mistaken Claim #1*
>
> USA and China have already won, or are close to winning, the AI race.
>
> TL;DR ANN-based systems will be increasingly marginalized as new paradigms surpass their inherent limitations. New paradigms, being new, means they may be applied to notably different tasks and domains; the 'AI story' will take on new forms. Core ideas behind each paradigm will come from prototypes created at pre-competitive research labs around the world, in collaborations between industry and academia. Several AI races will thus unfold over the next 50 years. USA defined the first one; Europe should participate in – and could even define – the next.

Fact: The Race Has Just Begun

The 'AI race' – or, the 'Age of Applied AI,' as it is more aptly called – is not nearly over. In fact, it has just begun. The vehicles competing in it, to use a metaphor, are powered by a three-cylinder engine, or, if you prefer it a bit more macho: A fleet of spaceships with three-stage booster rockets whose first stage is *basic research*, which, as discussed in the *Technology Readiness Levels* section (p.6), is the early work that must go into any brand new idea to bring daydreams from the blue sky down to earth. At this initial stage, for any reasonably profound idea, it is too early to know for sure whether it's any good; whether it is worth the effort; it is certainly too early for a patent and even too early for a prototype. The third stage is the one we call *applied AI*, which can only start when there is evidence of a return on investment of resources and time.

What defines the second frontline of the 'AI race' then? It is wedged in the middle, between these two stages. This is the 'no man's land' that almost always gets glossed over and quickly forgotten, if it's even mentioned at all—the part that politicians seldom spend any

time on. Yet it is that very critical time before an idea is concrete enough to make investors open their wallets; before anyone is convinced the idea can scale and everyone is still unclear on the path to its monetization. And it can span many years, even a decade, depending on access to means, methods, and know-how. This is the very materialization of an idea, its concretization, prototyping stage; the process that changes a good idea into a real investment opportunity. It is also where many good ideas die.

The three cylinders of society's innovation engine:

Basic Research → *Prototyping* → *Applied R&D*

Any country that wants to take an active part in the AI revolution over the next 50 years must *fire at full power on all three* for that entire period.

Do not try to save money, time, or effort by slashing funding, opportunities, or facilitation of these stages: Any slowdown – *any at all* – on any one of them will affect the entire race and delay overall progress, allowing others to get in front (but on time scales that are a bit difficult for politics to follow, easily extending from 7 to 14 years).[58]

Since the founding of AI as an academic field of research in 1956, every decade has seen some notable breakthroughs or ideas. For the longest time, none of them were sufficiently convincing to reverse the majority view that AI belonged to the world of pure science fiction, along with material transporters, holodecks, and lightspeed space travel. Until now, that is. AI has finally broken through the 'barrier of practicality.' As we discussed in section *It Starts With a Spark.*. p.8), today's AI technology derives directly from early ideas set in motion in the 40s and 50s (McCulloch and Pitts, 1943; Minsky, 1952). Now that this technology has become practical, albeit with well known limitations, to suggest that we are somehow in the final stages of figuring out how general intelligence works is naïve to say the least. As we laid out in the section *What General Intelligence Must Be Made Of* (p.16), this is far from being the case. AI R&D will continue for as long as there are unanswered questions about how intelligence works (and even quite a long time after, as there will always be interesting follow-on work to be done), or about how it can be applied for the betterment of society.

The United States was the original early mover in basic AI research around the middle of last century and has been writing most of the AI playbook ever since. Slowly but surely, through meticulous top-down control and coordination, the Chinese are catching up to them while the U.S. is getting increasingly entrenched in contemporary applied AI due to enormous investments (cf. Wilkins, 2026). But that was all just a prelude to something much bigger which will unfold over the next decades.

Many people may not realize that there is more than one AI race ahead of us, each with a different set of rules than the last one (yet all of them must go through the three innovation stages—there is no way to cheat). The one that we are in the middle of now, or coming to the second half of, runs on statistical principles, fueled by existing human-authored data and enormous amounts of energy- and compute power. As the 'age of statistical AI' comes to an end, there may be one or two follow-on races based on variations of its underlying technologies. Eventually, better – and very different – kinds of AI will come along, and the application areas for AI will become increasingly more important than writing emails, making images from prompts, and making "Ask the Chat" YouTube videos.

If we get too stuck in the contemporary AI innovation echo chamber, however, others will surely beat us to the punch on the next paradigm, and we may not like it. What is the next paradigm, you may ask. Well, that depends in part on how our bets are placed. If you think that the tasks contemporary AI is currently being applied to are extremely important, or the most urgent ones to be automated with AI, then the next decade of AI R&D should surely be focused on better technology for *solving these exact same tasks*. If we follow this path, and consider the potential of AI technology for doing things that intelligence is good for, then an obvious good bet would be to remove as many of the inherent limitations of contemporary AI as possible, both theoretical and practical (see sections *Theoretical Limitations of Artificial Neural Networks*, p.21, and *Practical Shortcomings of Generative AI*, p.27). If you know of tasks or domains that are currently beyond the scope of contemporary AI and that are desperately crying out for automation with AI that still doesn't exist (*I* certainly do—I listed some of these in the box on p.15; Time, 2026), well, then let's find some promising approaches demonstrated in an existing prototype or two in research labs that we have a chance of getting up to TRL-7 in 5-7 years. That is a bet worth making.

In his 1950 paper *Computing Machinery & Intelligence*, Alan Turing outlined two different kinds of engines that the AI race could be built on. On the one hand he positioned "pre-baked" AI, ready 'out of the oven'; on the other hand was what he referred to as "child machines" that were equipped with learning mechanisms, like children

[58] The effects of misguided "cost savings" actions like reducing R&D funding, even temporarily for a few years, are seldom seen immediately; after all, the entire innovation conveyor belt runs at least 15-20 years long and takes a lot of effort to fully staff. On the other hand, it should also be emphasized that *any kind of participation* in the AI race is much more important than being in, or even close to, its bleeding front line of research—so, instead of thinking about this as a competitive race, a different metaphor might be much more apt, like that of implementing a market-driven economy or establishing a Ministry of Education.

[59] Wang (2007); code available on Github: https://github.com/opennars/opennars.

[60] Nivel, Thórisson, Dindo, et al. (2013); code available on Github: https://github.com/IIIM-IS/AERA.

TABLE 1. Comparison of AI technologies in the context of the Nordic countries and other small-data nations.

Artificial Neural Networks (e.g. LLMs, DNNs)	**Autonomic AI Principles** (e.g. NARS,[59] AERA[60])
1 Entirely dependent on human-authored data	1 Autonomously seek out relevant data
2 Produce low-quality regurgitated information (slop)	2 Generate empirically verifiable high-quality new data
3 Threaten privacy with illegal surveillance	3 Enforce directly protection of privacy & personal choice
4 Threaten democracy through easy disinformation	4 Strengthen democracy through direct self-disclosure
5 Threaten civilian freedoms through centralized control	5 Strengthen citizens' freedoms by decentralizing control
6 Extreme need for compute-power and energy	6 Ultra energy- and compute-power efficient
7 Woefully inadequate for small-data nations	7 Perfectly suited for small-data nations

(Turing, 1950).[61] The many limitations of "good old-fashioned AI" in last century, and contemporary AI now, have both laid bare the problem of the former approach: In a world with new events every day, a pre-baked solution is expensive to make and keep up to date (to infuse contemporary AI systems with everything that happened yesterday you must bake a new ANN every day). Systems that learn autonomously, on the other hand, should be able to read the news and update themselves without a "baking oven." These ideas, are already brewing in research labs around the world (see Table 1). Those who have been in the field for more than two decades will know several examples of radical departures from the technologies underlying trending solutions. Which one of these will be the first one to inject a major inflection point to AI progress is anyone's guess.

By any realistic estimate, the 'Age of AI' will last for at least five more decades. During that time, a ton of new automation methods will see the light of day. How useful or harmful those ideas will be is up to the next 2-3 generations of researchers, politicians, and people of the planet. The worst-case scenario that we stand before though is that everyone just gives up, succumbing to the will of technocrats, plutocrats, monarchs, oligarchs, feudalists, fascists, or some combination thereof, that we have witnessed in close range for the past couple of years, and for which history provides plenty of evidence to have been an unfortunate turn for the vast majority of any population.

The only way that Claim #1 could be true then is in the form of a self-fulfilling prophecy: If European countries were to *believe that this claim is true – and behave in accordance with that belief* – then it is very likely to *become true*. The onus is on all of us to take steps in a direction away from such an inevitability; we must resolve to move Europe, and the world, in a direction that we'd *like* it to go. Actions already taken by Canadian PM Mark Carney (Carney, 2026) and many European leaders bode well on this front. Such work must be strengthened and sustained.

[61] I prefer to call the latter 'seed-programmed AI,' as the system must be given some form of 'seed' to bootstrap its learning (Nivel, Thórisson, Steunebrink, et al., 2014a; Thórisson, 2020b).

False Claim #2
Europe can't compete with China and USA on AI technologies

TL;DR This claim is true for ANN-based AI, which requires big data to work well; whoever has the biggest and most coherent data wins that race. European data is thus too fragmented – by language, borders, and law – to compete directly on the main fronts of contemporary AI. But this is not the only AI race.

The race that Europe should instead aim to win is the small-data AI race, which has hardly begun. With a different set of rules, Europe can compete on AI.

Fact: Small-Data AI Wins the AI Race for Small Nations

When it comes to contemporary AI – based on artificial neural networks – whoever sits on the largest amount of curated data will always be the winner. There is simply no way for small nations, or even large groups of small nations, to compete with countries whose companies have already amassed enormous amounts of data, which by now are largely proprietary, and who have enormous resources for scraping additional data from the Internet to augment it.

> The most obvious way for Europe to compete on AI is to do what Africa did to catch up to the age of telecommunication: Instead of copying the infrastructure of other nations by digging ditches for several decades and put billions of kilometers of wires into the ground, they went *straight to the next paradigm:* Mobile phones. Europe should do the same in AI. Next-paradigm AI can be made to be safer, more data meager, less compute hungry, more manageable, more transparent—in short, better in every way.

What does it take? We must involve people who know what this involves—not just MBAs who started thinking about AI two years ago; not just engineers, no matter how smart, who have never thought about the phenomenon of intelligence until now. Stop limiting grants for basic research to particular pre-defined technologies (like "genAI" or "large models"); instead, list the requirements without

regard for how they are addressed.[62] Increase funding for research that deliberately opens up the playing field to radical ideas: a pan-Nordic – and preferably pan-European – unifying infrastructure that supercharges an effective and efficient innovation conveyor belt, from level 0 to level 9 and beyond.

How long will it take for Europe to catch up? Notable results could be seen as soon as a decade from now; victory could be declared in less than two. Whatever the eventual numbers, Europe will be better off. With a nimble, practical plan that puts trust in scientists and Europe's scientific institutions, success will mean that Europe will lead the world in safe, transparent, trustworthy AI.

What other kind of AI is worth our time anyway?

False Claim #3

Present AI methodologies suffice to successfully implement a fully AI-centric society; little or no new scientific research is needed.

TL;DR The idea that contemporary AI, or some soon-to-arrive variation of it, can be trusted to advance to a fully-automated society, is flawed, and based on wishful thinking (not by the general population but by investors and entrepreneurs).

Fact: Unknown Societal Impact of Contemporary AI

As we have explained in prior sections (see *Practical Shortcomings of Generative AI*, p.27, and *Theoretical Limitations of Artificial Neural Networks*, p.21), contemporary AI is not trustworthy, is environmentally expensive, and is running out of data (cf. Jones and Thompson, 2025). This does not mean contemporary AI is useless, but it *does* mean is that its uses must be limited to non-critical-path applications (cf. Bean et al., 2026). Due to its unpredictable failings (the best ones get 1 out of 16 answers wrong), and inability to verify and control its behavior, it cannot be used to increase safety, reliability, or correctness of any existing process—it can only make it worse. Fields such as medicine, manufacturing, traffic control, and governance, to name a few, should and must therefore be off-limits for contemporary AI.

The effect of imitation-intelligence – as genAI systems can rightfully be said to be – on people's mental well-being is not very well understood. A large number of bad outcomes from the present experiment show that people's willingness to take the chatbots at their word is much greater than most anticipated (cf. Tangermann, 2026). To be against the free reign of companies to release these on an unsuspecting public is not to be against new technology, it is to be against using the general population as guinea-pigs without their consent. Much more research is needed to investigate the effect of these technologies on children, teens, and even grownup individuals.

False Claim #4

General machine intelligence (GMI) is a myth.

TL;DR The concept of 'generality' can be seen as a gradient, and as such it has been implicit in the field of AI research from the very beginning. The term 'artificial general intelligence' was introduced around the turn of last century to explicitly reference research that has generality and autonomy as its main focus, as opposed to other research in AI that doesn't.

Fact: Human Intelligence is 'General' by Definition

The generality of learning can be understood as a gradient by comparing the answers we might give to the question "what can be learned, and in how many ways?" for any two or more animal species. The answer might produce different lists for different species; the longer the list of one compared to another, the 'more general' its intelligence.[63] Since no accepted theory of intelligence exists that defines concrete research targets, and with plenty of examples of intelligence in nature, including humans, the concept of 'more general' is very useful for basic research. Much less so for industry.

The term 'general intelligence' originally took hold as a way to distinguish between AI research that aimed for 'useful applications soon' and basic research that aimed towards human-level intelligence. Although the two may sometimes look very similar to the uninitiated, the end-goal of each makes them very different; while software solutions intended for immediate application can be built with a variety of techniques (whether they carry the 'AI' moniker or not), no such solution could legitimately be considered central or even relevant to research aiming to unravel the workings of an intelligent mind, because, by definition, any approach that can deliver immediate solutions is already well-known, and thus does not aim to capture, and can in no way lay claim to explaining, the foundations of intelligence or the fundamentals of thinking.[64]

[62] In *The Impact of Curiosity-Driven Informatics Research*, Aiello puts this as such: *"Science requires patience. It is not for the scientific process to set short-term goals and guarantee the impact of the results."* (Aiello et al., 2026, p.5).

[63] Without a proper theory of intelligence, we cannot claim that intelligence can be represented on a single dimension (cf. Gardner, 1983), but there are many ways multiple dimensions may be projected onto one.

[64] In a paper published by Pei Wang in 2020, based on his earlier thesis (Wang, 2020; Wang, 1995b), and a shorter paper on the topic (Wang, 2008), Wang elucidated an excellent working definition of general intelligence that – unlike other widely used definitions of intelligence (Legg and Hutter, 2007; Thórisson,

Whether machines with general (human-like) intelligence will ever be possible is more likely now than it was, say, 20 years ago, due to significant progress and the steady progress in clarifying many theoretical questions. What is less clear is whether machines with general intelligence will ever be a product in themselves; any special-purpose machine will always be better for that purpose than a genera-purpose one. A vendor of an "AGI machine" would therefore always be in a worse position than a vendor that offered a specialized solution. In any case, a full general intelligence is unlikely to be a product in and of itself because all known use cases of human-level expertise are clearly demarcated by goals and outcomes—otherwise we could never know if the job was properly done!

False Claims #5 and #6:

Contemporary AI technology is a solid foundation for general machine intelligence.

To fully explain intelligence (human or general), no new scientific paradigms are needed.

TL;DR The list of features that human intelligence has, over and above contemporary AI, is long, including some important features without which AI technology cannot be considered trustworthy. Yet, providers of contemporary AI maintain that it is safe, or can be made to, at least in principle. This false claim results in misunderstandings and misuse of AI technologies. Due to the underlying principles of contemporary applied AI it is neither possible to add the missing features nor to make it trustworthy. Contemporary AI, and AI in the most general sense, needs a better scientific foundation.

Fact: Too Many Open Scientific Questions Remain

For the reasons already listed in Claims #1 to #3, it is clear that contemporary AI cannot be a theory of intelligence; even worse, it *itself* is lacking its own scientific theoretical basis. Furthermore, no higher-level theory exists, from psychology, social science, or neuroscience, that can be used to contextualize contemporary AI. In short, what is sorely missing now is an accepted, unified scientific theory of (general) intelligence.[65]

False Claim #7

Industry has already taken over all relevant AI research from academia, or will soon.

TL;DR Industry's R&D time horizon is 3-4 years. The number of principles inherent to intelligence that have yet to be formulated in a way to be implementable as software is large. These will require at least a decade or two of basic research to be adequately addressed.

Fact: Basic Research Remains Beyond Industry's Capacity

As can be seen when put the list of necessary ingredients in general intelligence (see section *What General Intelligence Must Be Made Of*, p.16), and the numerous theoretical limitations of contemporary AI (see section *Theoretical Limitations of Artificial Neural Networks*, p.21), in the context of how new knowledge is created in the process of innovation (see section *Technology Readiness Levels*, p.6), we can see that the path to answer even just a small percentage of the open questions in AI will be counted in decades, not years. This means that the necessary ideas, methods, and theory, is out-of-scope of companies, even tech giants, whose R&D horizon is 4 years at best.

The largest restriction of contemporary AI is its inability to create truly new ideas that follow logically and arguably from a set of principles that can be supported by commonsense arguments that reference verifiable causal relations (Pearl, 2001). Until such AI exists, the 'intelligence explosion' that many have predicted will happen through recursive self-improvement (cf. Good, 1965) is extremely unlikely to happen solely with present technologies in hand.

False Claim #8:

The arguments for abandoning AI R&D altogether are better and stronger than those for continuing on the present path to general machine intelligence.

TL;DR The many unwanted side-effects of contemporary AI stem from the nature of the principles underlying its particular technology of the present time—artificial neural networks (ANNs). Ongoing research in AI labs around the world are looking at more manageable, more transparent, more reliable, more efficient, more effective, and more trustworthy approaches to automating thought. Speeding up any of these, and strengthening the legal frameworks which may be needed to result in the kind of society we wish to create, will help get us out of the present state of affairs sooner.

2020a) – manages to clearly demarcate the phenomenon's central feature and separates it from other overlapping and competing ones, like evolution, biological adaptation, and simple control.

[65] Honorary mention for work on creating such a theory, and motivating others to pay attention to this important matter, should be awarded to these (and probably more): Minsky (1988), Newell (1994), Laird et al. (1987), Anderson et al. (2004), Swan et al. (2022) and Wang (2007, 2025).

Fact: Next-Paradigm AI Could be Much Better

Research in artificial intelligence is, and always has been, about automating thought (Thórisson and Minsky, 2022). With computing power being widely accessible, and prices continuing to fall, any progress on automating thought will be applicable almost instantly. Next-paradigm AI could be much better than contemporary AI, but it won't be unless we do something about it.

However much over-investment in contemporary methods will delay progress towards next-paradigm AI, one thing is clear: Automation has been a pursuit for at least two centuries; research in AI is unlikely to be abandoned, banned, or outlawed. But it *will* become more regulated. Like knowledge of chemistry, certain applications of AI and related computer science-based know-how are likely to be – and *should* – be controlled by law and government oversight. Just as in the case of radioactive material or health-threatening chemicals, certain AI products could – and quite possibly *should* – be banned. But it must be clear to investigators what constitutes breaking the law. Since AI is software, for this to be possible the software must be transparent. Surely, transparency can and should be made an integral dimension of compliance. How that is to be implemented is a question the Nordic countries must answer. Using AI technology that is inherently transparent would significantly improve the situation over the present state of affairs.

A New Vision for Nordic AI Research

In a recent article, the President of Iceland wrote to her nation about artificial intelligence:

> *Together we bear the responsibility for creating a society where technology serves people – not the other way around. We are facing a large and complex task that requires dialog, solidarity and a clear vision.*[66]
>
> — Halla Tómasdóttir
> *President of Iceland*

I couldn't agree more with this sentiment. The task before us is substantial: The current path of applied AI, steered mainly by the US technology giants, and whose fingers are furthermore deeply embedded in controlling the social discourse about it, goes directly counter to this aim.

"AGI next year!", CEOs in the United States keep promising (see Figure 6, p.15). Option one: Just sit and wait for this to happen, only to witness the predictions moved back yet again, by the predictors themselves, another 12 or 18 months. The other option is to take matters into our own hands and choose approaches that are better aligned with our aims for sovereignty: Get to work on *next-paradigm AI*. Is this worth our time? Absolutely. Why do I think so? Because the next-paradigm AI technologies are already available in the lab—I have seen many of them, and worked on several of them. Having better kinds of AI already at TRL3-4 – and by 'better' I mean they promise to lift limitations of contemporary AI, especially their opaqueness – means we really have *no good reason* to keep betting on genAI as a path to trustworthy, explainable, controllable intelligent automation. It is neither in the interest of the tech giants, nor is it theoretically feasible, as the preceding pages have extensively argued, to continue trying to force contemporary AI, like a square peg into a round hole, towards these ends. If you're not convinced and would like to give genAI a chance just a little bit longer (but hopefully not for too long!—for all our sakes), I remind you of the ultimate reason for why they will never suffice as a foundation for a stable AI-driven economy: Their dependence on a misapplied theoretical foundation.

> I believe the Nordic countries – and in fact, any small nation, none of which stand to benefit from contemporary AI nearly as much as larger ones – can and should actively seek alternatives: Controllable, predictable, reliable and explainable machine intelligence—*trustworthy AI*. Worthy candidates already exist in research labs.
>
> In contrast to 'next-generation' AI, which incrementally improves on the status quo, there is clear evidence for the foundation of next-*paradigm* AI that can run *locally*, on small computers, and require small amounts of data. That can learn autonomously from its own experience. Interaction with it will be fully controlled by its local, immediate end-user, and the data generated in the process will stay local. Before it does anything, this AI can tell you exactly what it plans to do, and why it intends to do it; at runtime it can explain to you what it's doing, using verifiable, grounded causal explanations; after it finishes a job it can explain *what* it did and *why* it did it. All will be fully auditable and empirically verifiable. These will be *radically new AI systems that understand what they are doing*.

With proper strategic planning across the Nordics and Baltics, fully transparent, self-explaining automation along these lines, with trustworthy safeguards for a variety of domains – domains that contemporary AI technologies are ill-advised for automating – could become practical applied technologies within 7 years or less, replacing contemporary AI methods within 12-15 years.[67] But however long it takes, the key point to remember is this:

[66] Original quote in Icelandic:"Við sem hér búum berum sameiginlega ábyrgð á því að skapa samfélag þar sem tæknin þjónar manninum – ekki öfugt. Hér er um að ræða stórt og flókið verkefni sem krefst samtals, samstöðu og skýrrar framtíðarsýnar." (Tómasdóttir, 2026).

[67] These numbers have many dependencies that will influence in a number of ways how things play out. But even if it took 20 years, it would absolutely be worthwhile.

> *If we don't take this possibility seriously, and if we don't start on this path towards next-paradigm AI, we'd be giving up more than just the technology—we'd be relinquishing the chance to steer our fate towards continued technological independence and sovereignty.*

Next-paradigm AI will be defined by several foundational scientific advances that don't exist in contemporary AI, but are already underway, and a few of which have crossed the threshold from TRL4 to TRL5-6. A prerequisite of leading the AI revolution is efficient and effective operation of the innovation conveyor belt (see Figure 3, p.7) in a way that facilitates, in an opportunistic manner,[68] the movement of new ideas as they develop from ideas to prototypes to application. For researchers interested in next-paradigm AI technologies, below is a list of what I consider *necessary and sufficient* cognitive features that, if put correctly together in a single architecture, would be a tremendously powerful tool to shape the future of our communities, countries, and the planet. There is no requirement for all of them to be fully developed and unified before they become useful in a particular product or service, but together they point unmistakably in the direction of general machine intelligence—without the many threats of and limitations of contemporary AI (see *Threats of Contemporary AI*, p.28).

A New Vision for Nordic AI: The WHAT

The blue box (p.37) is not provided to predict future technologies but rather as a list of topics for which currently available technologies are woefully inadequate, both on the theoretical and practical sides. The list assigns terms to concepts and *opportunities*—opportunities for both theoretical breakthroughs and breakthrough applied technologies. A breakdown like this can be done in various ways, but in my view these concepts and categories are sufficiently course-grain, sufficiently intuitive, and sufficiently broad-scale to cover most of what might be put on any such list.

If we fast-forward one or two decades, assuming we have access to a good theory of intelligence, each of these could be paired with any of the others to produce all sorts of variations on intelligent control, for a variety of purposes. These would not require any kind of front-end that pretends to be a human; quite the contrary, a proper AI control of complex processes and tasks can unassumingly take on such roles and carry them forth, while providing appropriate and adequate feedback to its surrounding.

[68] It should not be the role of research funding agencies or governments to predict or decide which ideas are mature enough to move on to new stages, this should be the task of the researchers themselves. Funding schemes should directly support the 'bets' that researchers put their careers on the line for, as this should automatically qualify, other things being equal, as sufficient proof and justification for working on them. It is entirely inappropriate, and works directly against scientific progress, when calls for proposals dictate methods, topics, approaches, and scientific methodologies; these should be outlined in descriptions of outcomes (e.g. "profanity-free"), not in descriptions of methods (e.g. "guardrails for LLMs").

A New Vision for Nordic AI: The WHY

Why would the pursuit of exactly these properties in AI systems be worth our time? How will good solutions to these cognitive features help mitigate the negative side-effects of contemporary AI technologies? The answer lies along a few dimensions.

Firstly, by seeking to develop and deploy a *transparent* technology whose knowledge is grounded in *real-world cause-effect relations*, new *kinds of AI* technologies will emerge that are easier to verify, monitor, measure, and audit. This will enable the implementation of appropriate guardrails, application of appropriate auditing mechanisms, and practical enforcement of government regulation. Furthermore, any such transparent technology that is capable of full self-disclosure offers the option of automating such regulation, relieving us from being forced to do this through manual means. Its self-disclosure will of course also be grounded and transparent, so for extra safety, auditing of the auditing (meta-auditing) can even be implemented at little extra cost.

Secondly, actively seeking to develop small-data AI (or small*er*) brings many benefits including lowered computing cost, lower computing time, and less computing complexity. It also simplifies data collection, generation, and handling in a beneficial non-linear fashion. Coupled with transparency, these real-world impediments can be pushed even lower, improving further the cost-benefit ratio of this new small-data AI.

Any small-data AI that can cumulatively learn relevant information by autonomous data extraction from experience is geared for learning on the job, potentially removing a large part of its expensive training phase. Effective cumulative learning may also remove or significantly reduce any need for re-training, re-compilation, or re-design of the whole system. Lastly, driving research and development towards increasingly smaller data creates an inherent motivation to increase the intelligence of the systems we build, other things being equal, because other things being equal, doing the same with less data requires greater intelligence.

> The largest and most obvious benefit of small-data AI for small-data nations is though undoubtedly the acquisition of *appropriate, self-sufficient and adequate* methods for automation of a wide range of specific, targeted tasks which, in light of government, industry, and societal goals, are considered appropriate and ripe for automation.

Thirdly, small-data AI that uses knowledge representation with explicit reference to cause-effect relations can apply empirical reasoning to this knowledge. That means

Next-Paradigm Artificial Intelligence Systems with Autonomous General Learning & Control
Overview of Target Properties

1. **Small-Data**
   Functional requirements for target system behavior can be fully met with tiny amounts of data (generated algorithmically). Transparency ensures possibility of empirical verification. No big data!
   NB: In principle, intelligence is negatively correlated with data requirements.
2. **Transparent Knowledge Representation**
   System's complete 'wiring' can be inspected in full, at any time, by any third party and the system itself, producing a complete map of behavior for known and unknown environments and situations, before, during, and after task execution. For flexibility, knowledge should be fine-grained (compositional/modular).
3. **Empirical Reasoning & Learning**
   Capable of autonomous logical, situated reasoning that stands up to real-world empirical verification; taking uncertainty directly into account, including uncertainty of own knowledge.
4. **Full Self-Disclosure**
   Actions, and action potential, can be fully described for known and unknown environments and situations, by the system itself and any third party, before, during, and after system deployment.
5. **Cumulative Learning**
   System is capable of autonomous situated, self-guided learning (and outside help / teaching not precluded).
6. **Meta-Cognition**
   System can reflect autonomously on its own learning, for second-order control of cognitive development, learning, and safety control.
7. **Goal-Driven Unifying Architecture**
   Several – or all – of the above capabilities unified in a single system coordinated by explicit goals. System autonomously generates goals and sub-goals, based on its pre-determined user-defined drives, and resolves goal conflicts autonomously.

it can be set up to actually *understand real-world cause-effect relations*, which means in turn it can be more trustworthy than any contemporary applied AI technology. For smaller nations, having independence from big data means that the AI is better suited to the nation's inevitably smaller data set. Having small-data AI also means it doesn't require large servers to run—it can run locally, on small laptops or even wristwatches, which enables new levels and types of cyber-security. Coupled with transparency, such AI is also harder to compromise (the hacks will be more distinguishable) and thus easier to protect from the effects of tampering. Taken together, *these features translate directly to AI sovereignty*. We can summarize this in the following way:

- **Transparency** is a *necessity* for effective oversight, accountability, auditing, rule and regulation verification, and enforcing the law.
  → *Better for law-abiding nations*
- **Small(er) data** means that the AI can run on small(er) computers, costs less, and can work on shorter timescales with less energy needs.
  → *Better for small nations*
- **Cumulative learning** means that the AI doesn't need nearly as much data up front and is more flexible in its application.
  → *Better for small nations*
- **Empirical reasoning** means that real-world cause-effect relations can be understood by the system, including (potentially dangerous) side-effects of actions and plans, making AI safer.
  → *Better for safety and accountability*
- **Meta-cognitive and goal-driven** architecture makes top-down and runtime system control, via guardrails and other methods, more effective. Combined with empirical reasoning makes it possible to safely let the system itself control aspects of its own operation.
  → *Better safety and accountability of autonomy*

Of course we must mention here that the further along in their development the above features are, the more options there are for combining them in various ways for a variety of purposes, and the more of the above features are successfully combined and coordinated in one and the same the AI system, the more powerful it will be, other things being equal. Furthermore, transparent small-data AI will be easier to deploy for a variety of targeted purposes, under a variety of conditions, increasing the autonomy of any nation who can master it. This translates directly to *more powerful* sovereign AI.

A New Vision for Nordic AI: The HOW

To make this vision a reality, no new organizational, political, or funding instruments need be invented. However, existing infrastructure across the Nordics and Baltics must be revamped and reorganized to ensure effective and efficient knowledge transfer, through open-source, cross-domain collaboration, with greatly improved and strengthened industry-academic communication and coordination. The following numbered paragraphs capture some suggestions.

1 A Vision for 'Nordic Transparent Trusted AI'

Proclaim a pan-Nordic-Baltic collaboration on mutually-beneficial trustworthy AI R&D, outlining clearly defined target utility outcomes and societal gains. With a focus on AI and this future vision, define clear quality, performance, and trust requirements and methods for measuring them. Secure funding for both theoretical and applied research that takes society, education, governance, security and sovereignty into account and opens every door for creative researchers to achieve the defined goals, whether by conventional or unconventional means.

Instead of blindly – and incorrectly – accepting as an immutable fact that present and future progress on AI is and should be dictated by foreign multinationals, accepting it as an excuse to just sit and wait, we must understand that how we spend our research time is *a choice that we have direct control over*. No one will do *our* research for us; if there are certain technologies, techniques, and knowledge that benefit us more than it benefits others – such as technology for our language and sovereign control – it is unlikely that others will do a sufficiently good job, or do it at all; *we* must be the ones to define how that work is done and *we* should be the ones when that work is done. *We can and must define* what kinds of technologies are to be developed for products and services that will be deployed in *our society*.

2 Low-Friction AI Innovation 'Conveyor Belt'

The more effectively and efficiently ideas can move along the innovation conveyor belt, with as few showstoppers as possible between idea and product, the faster society invents and innovates and the faster it can improve.

It is not sufficient to simply lower the friction in higher technology readiness levels (TRL; see *Technology Readiness Levels*, p.6), ambitions for advances in basic research must also be boosted. Society's traditions and approaches to creative thinking should not be limited to playing with paper maché in kindergarten—no grownup should feel shame in entertaining 'wild' or 'unrealistic' ideas; after all, the ability to do so is what separates homo sapiens from other species on the planet.

An efficient and effective innovation conveyor belt is a critical component in innovation competition. Importantly, the innovation conveyor belt cannot be short-circuited: Ideas, prototypes, experiments, and experimental deployments of products and services take time, effort, space, and funding. Furtheremore, the middle of the conveyor belt cannot be deleted—any tendency to bring industry directly into university funding, in any shape or form, is (in the vast majority of cases) misguided; it ends up confusing true invention with purpose-driven innovation, hampering progress and messing with the process of higher education.

3 Actively Counteract Interdisciplinary Friction

Many interesting problems in science lie at the boundaries between disciplines; nowhere is this more obvious than in AI. Breaking down departmental and disciplinary walls is an obvious way to speed up progress towards better AI technologies.

Not only is interdisciplinary collaboration an obvious way to speed up progress in AI, it is is an effective way to seek out and unify disparate terminology, incoherent theories, incongruent foundations, inefficient approaches, and ineffective methodologies, all of which is in fact *the duty of science* to do.

The current division of knowledge creation into academic disciplines exists for reasons that made perfect sense in the past but may not do so any longer; a good example being the chasm between cognitive psychology and artificial intelligence, which exists purely for historical reasons, and can be seen to directly counteract scientific progress.[69]

If you are not ready for a micro-lesson in the philosophy of science, let's just cut to the chase: The reason disciplines exist at all is because at some point in the past, accumulated knowledge outgrew any one person's capacity to master it all. This, of course, was good news in a way, as it meant that humanity's understanding of the world was growing. At that point in time, certain philosophical assumptions, specific methods of inquiry, and particular know-how, made it convenient to dissect knowledge in particular ways, often dictated by societal needs, e.g. sheltering from weather and tending to the sick, into separate disciplines. For many scientific fields, those background assumptions have now changed in important ways, and

[69] This divide is one of the main reasons for the many wildly off-mark and unrealistic predictions that industry has been making about AI products in the past (see e.g. Figure 6, p.15).

nowhere is this more obvious than in disciplines related to computer science, artificial intelligence and other 'sciences of thought.' Robotics, for instance, is divided right down the middle, into hardware on the one hand and software on the other. Robotics experts belong to either one; much more seldom both. This makes building intelligent robots more difficult than e.g. cars and houses. Why the split between software and hardware? Because engineering, coming from a tradition of focusing on hardware, took on the motors, wiring, and mechanics of robot bodies, while computer science, coming from a tradition of mathematics, took over the design of programs. It is a situation very far from ideal because 'robot bodies need robot brains'—one is pretty useless without the other.

These kinds of splits not only hamper progress in robotics; a similar one between computing and epistemology (philosophy of knowledge) has been hampering progress in AI since virtually the beginning. The deep traditions around which academia is organized are difficult to fight, but it's certainly not impossible; if we want to get on the fast lane of AI progress it must be done explicitly, strategically, and diligently. Most of the time this is easier to do by looking out at the road ahead, through the windshield, than fixating on the road traveled, through the rear view mirror. In the words of the former Director of Zentrum für interdisziplinär Forschnung (ZiF) in Germany, Dr. Ipke Wachsmuth: *"Interdisciplinary research and collaboration is not easy in practice but our experience shows that nothing else can take its place."* (Wachsmuth, 2016).

For industry, at TRL 7-9, the name of the game is iteration, refinement, and adjustments, in light of market needs; for basic research and early innovation at TRL 0-6, however, progress must be driven through a search for answers to open questions. If no good questions are being asked, no good answers will be found: In other words, "garbage in, garbage out." Asking good questions, in turn, is a function of taking broad, multi-disciplinary views on the subject at hand.

It is an immutable fact that AI rides on a multi-disciplinary scientific foundation whose nature is nevertheless no different than other topics found in other fields of scientific research—to advance it needs appropriate tools, methods, ideas, and free exploration. Removing friction between disciplines, departments, research labs, and research institutions, is key to facilitating and increasing flow of ideas, questions, and collaboration.

4 Public Access to Pre-Competitive Results

The work done in TRL 0 to 6 is usually pre-competitive, meaning that it is uncertain to be worthy of investment. To ensure that researchers and developers have overview of state-of-the-art, the results from these stages should be as accessible as possible to the general public.

The two ends of the innovation conveyor belt are not exactly alike and require different things. For one, cash flow is not positive until TRL 8, 9, or even later (Figure 3, p.7). Just as we must supply a market with conditions for healthy competition, we must attend to the early end of the belt, giving support to and making space for lone inventors and startups alike. Choking in any way the supply of basic research increases the time that that necessary groundwork takes, which unavoidably slows down innovation overall. Whether it is cold fusion, quantum computing, or artificial intelligence, appropriate measurements can track progress and inform the support mechanisms. More money at the start of the belt will not make individuals more creative—the economics of invention are largely dictated by the human mind, which can rarely be scaled up with money but require a guaranteed sustained supply of time for deep thought. However, the concept of funding starvation *will* hamper activity, productivity, and enthusiasm at that stage. Assuming sufficient funding for securing people's long-term dedication and energy, lockdown of knowledge in the early stages (TRL 0 to 6), for instance through premature IP ownership, should be avoided: That work is and should be pre-competitive, funded by public money, open-source, and widely accessible.[70]

5 Lower Threshold for Funding Bold Ideas

Mechanisms must be implemented that increase the chances of funding for bold ideas. While this will inevitably – other things being equal – result in a larger number of failed projects (stretching a set of bows will break a few of them), but if its done well, the return on investment will be same, or greater. Many countries are doing this for other domains such as quantum computing and cold fusion. We'll be fools if we don't do it for AI.

Competitive research grants in Europe have been increasingly leaning towards applied projects with a short time-horizon.[71] Couple this with the fact that traditional peer

[70] A good example the power of this approach can be seen in the worldwide RoboCup competition (https://www.robocup.org), where winning solutions are shared with the community, including next-year's competitors. The strategy has sped up progress on various fronts of robot planning, sensing, and control.

[71] This is based on my own experience with research grants for AI and related topics; a case in point being that the latest bout of calls for proposals overwhelmingly uses terminology like "generative AI" instead of simply "AI" when detailing what kinds of research focus and topics qualifies for the funding. There is also evidence that research papers and patents have been getting less disruptive over time (Park et al., 2023).

review practices and methods for grant proposals are too conservative and a picture emerges that is clearly slowing down work on big, bold ideas—the kind that could create what philosopher Thomas Kuhn called "scientific revolutions" (Kuhn, 1962). In other words, the peer review process, as important as it is, has one big Achilles' heel when it comes to AI: *It is inherently anti-innovation!*

Here are two ideas for remedying this situation without turning the whole system on its head or making it much more expensive. One has to do with estimating 'impact': The review method should clearly separate a project's potential for impacting knowledge and the likelihood of it to be successful, or the capacity for its proposers to carry it through—conflating these aspects can create a bias against newcomers, many of who are likely to have fresh ideas and be unencumbered by prior 'investments' in particular approaches and topics, and is likely to put too much trust in the reviewers' ability to predict the outcome of the project. The measure of a proposal's 'radicality' and 'originality' should be cleanly separated from other measures of quality. A minimum can then be put on the percentage of grant-receiving proposals that score high on scientific originality and potential impact (assuming minimum qualifications are met on other dimensions), to ensure that bold ideas are not penalized for being radical or pushed down the list due to being more challenging than others that score similarly on other scales. Two: A lottery method can be used for the lower part of all above-threshold proposals, which would reduce the likelihood that bold projects are cut off from funding before the pot runs out for any particular call, making it less likely that built-in risk-aversion and biases against bold ideas have unjustified, hidden negative effects on high-risk, high-impact proposals.

6 Next-Paradigm AI Topics: Required Study in Degree Programs

If university students are exposed to unanswered questions in AI, the field will advance faster.

Three of my AI courses at Reykjavik University, where I have taught for 21 years, cover topics that deliberately extend beyond the "tried and tested" methods most often mentioned in introductory AI books (cf. Russell and Norvig, 2010). When I created my course *Advanced Topics in AI* in 2012, much of its material seemed "pretty out there" to my colleagues and students (remember: in 2012 only fans of science fiction and supernerds paid any attention to AI.) As the limitations of contemporary applied AI are being brought to the forefront, the topics covered in that course,[72] and regularly updated (but not radically) over the past decade, are now becoming increasingly and directly relevant in higher education, ensuring that MSc and PhD degrees in AI do not get stuck in the AI innovation echo chamber (see p.10) or becoming pure vocational training. This includes much of the topics in the list on page 37.

On my first day of class of each semester I often ask students whether they have taken any courses in psychology or philosophy; very few raise their hand. Then I usually give them a compressed introduction to the scientific method, because a central part of cognition is about control and hinges directly on interacting with the world: Cause-effect relations, intervention, hypothesis creation, experimentation, and reasoning over partial results (cf. Pearl, 2009). Students who have taken some courses outside of mathematics and computer science have invariably an easier time grasping such concepts and making conceptual advances on what kinds of solutions might be worth exploring.

7 Broad-Tech Research Centers for AI Theory

Theory is not a luxury—for a technology as important as AI, with as broad implications and unforeseen side-effects, we must put direct effort into ensuring its proper scientific foundations. At this very point in time the Nordic countries have a real opportunity to take the lead on this front, because the primary driver of real scientific progress is mindpower.

From the preceding facts about the current state of AI technology and the field of AI R&D we can conclude the following about what must happen next: Moving the field of AI forward means producing fundamentally new ideas about how intelligence works as a whole and strengthening its scientific foundations. This should include asking questions about what makes intelligence different from other kinds of related, similar phenomena and what separates it from all other kinds known processes such as evolution and simpler kinds of adaptation—as well as answering questions about what makes human intelligence uniquely flexibile, creative, systematic, logical, reliable, and trustworthy.

Comparative studies between cognition in related and diverse animal species, and comparisons between different implementations of similar component processes – machine or biological – must become common in academia. In other words, diversifying our AI research methodologies seems inevitable.[73] Ensuring that existing applied-AI laboratories extend their horizon into the next 8-10 years would help guide the direction of the work towards more sustainable, long-term thinking. One way to do this without losing the interest of companies, whose

[72] See http://cadia.ru.is/wiki/public:t-720-atai:main — *accessed April 22, 2026.*

[73] A good example of efforts in this direction is the new AI center in Norway, SURE-AI (https://www.sure-ai.no/).

horizon is typically no more than 3 years, or investors, whose time-horizon is often no more than 7, is to let researchers extend the needs of industry into the future, beyond their own horizons, allowing for a hybrid combination of basic and applied research that reaps the benefits of both—DFKI has sometimes called this "applied basic research."[74] When done right, this can significantly shorten the path from wild ideas to realization.

Would AI theory centers be devoid of implemented systems? Absolutely not! Implementation is as integral to AI as it is to engineering. But the field of applied AI is missing the equivalent of what engineering has in the form of physics. When engineers build bridges, skyscrapers, x-ray machines, automobiles, and airplanes, they rely on the solid foundation provided by the most advanced science of all sciences: Physics. Can we create something similar for intelligence? Insofar as intelligence is a natural phenomenon, just like the forces of nature, the short answer should be "yes." However, intelligence is probably rightfully characterized as being the most complex phenomenon science has studied. Practically speaking, this means that progress might be slower. It doesn't mean, however, that it is not worth working on. After all, real progress will be directly dictated by our progress on empirically-grounded theories of intelligence. Expanding such theoretical work beyond just intelligence proper, to societal, ethical, and other aspects of importance, requires interdisciplinary thinking, collaboration, and strategy.

Would AI theory centers be using mathematics exclusively? Absolutely not. When the scientific study of a topic is in its infancy, the topic is too vague, unexplored and ill-understood to create good definitions. Half-baked *working definitions*[75] are endured for decades, sometimes centuries, because without some commitments in place we'd have to resort to random search. AI has just crawled out of its earliest (lab) phase, but we are not yet at the stage where there are concrete definitions of everything that should be measured. AI ultimately belongs to empirical (non-axiomatic) science, and for most fields of empirical science, mathematics is a way to express the best ideas effectively, succinctly, and unequivocally. Lean too heavily into a theory's means of expression and your scientific topic starts to take second place; lean to heavily *away* from mathematics and your theory may never see its most elegant expression.

As we have established, original ideas emerge when science addresses deep, challenging questions. As long as some such questions remain to be answered about the phenomenon of intelligence, progress cannot be made by merely iterating over the existing set of ideas, especially not questions about how it all hangs together. Special attention should be paid to theoretical work—but not just mathematical, and not just neuroscience-inspired. Intelligence is an information-based natural phenomenon; alternative perspectives and approaches must get their place. Can progress on a general theory of general intelligence be provided by industry? No. To make scientific progress on anything, whether aqueducts, agriculture, aeronautics, or computing – all of which were complete conundrums to humanity at some point in time – society had to put effort into multi-generational long-term thinking. Academic and think-tank research laboratories are critical infrastructure to do this for AI.

[74] Aljoscha Burchard, DFKI (2015), personal communication.

[75] A 'working definition' is an early, partial, and often premature definition of a concept or phenomenon (see Wang, 2020 for an excellent discussion of this in relation to AI). Because getting a definition "right" of a complex concept or phenomenon on first try is close to impossible, the purpose of a working definition is to help research move towards a better one. Therefore, it is important not to get it "too wrong."

Acknowledgments

The paper is based on four keynote lectures I gave in Scandinavia and Iceland in fall of 2025:

- *General Intelligence: Ingredients, Methods & Progress*
  Intl. Conf. Artif. General Intell.
  Reykjavík University, Iceland, August 12^{th}
- *Airplanes, Safety & AI: Road to a Rosy Future or Imminent Failure?*
  Association of Professional Pilots (FÍA) Annual Meeting
  Reykjavík, Iceland, October 16^{th}
- *Trustworthy AI–A New Vision for Nordic Research*
  Stavanger University AI Day
  Stavanger, Norway, November 24^{th}
- *General Machine Intelligence: Timeline & Prospects*
  Nordic AI Meet
  Norrköping, Sweden, November 25^{th}

I would like to extend big thanks for valuable insights and comments on drafts of this paper to *Jón Karl Helgason, Stefán Ólafsson, Jónas Ingi Pétursson, Halldóra Mogensen, Jón Guðnason, Pei Wang, Hugo Latapie, Alex Moltzau, Hjálmar Gíslason, Vaughan Hines, Klas Pettersen, Gunnar Þór Pétursson, Katrín Elvarsdóttir, Richard Wheeler, Paul Richardson*, and *Lilja Dögg Jónsdóttir*, and members of the IIIM team, *Jeff Thompson, Gonçalo Carvalho, Helgi Páll Helgason, Marta Abad Torrent, Mário Silva, Arash Sheikhlar, Hlynur Halldórsson and Sigríður Ólafsdóttir*.

Thanks also to my colleagues at Reykjavik University's AI lab, the *Center for Analysis & Design of Intelligent Agents* (CADIA), for discussions on many of the topics touched on in this paper, *Yngvi Björnsson, Hannes H. Vilhjálmsson, Stefan Schiffel, Kamilla R. Jóhannsdóttir, Hrafn Loftsson, Jón Guðnason, Stefán Ólafsson, Torfi Thorhallsson and Elín Carstensdóttir*, and congratulations to all on the 20th anniversary of CADIA!

No ANN-based technology (genAI or other) was used to author the text in this paper.

I dedicate this paper to my daughter and her generation.

Dr. Kristinn R. Thórisson is Professor of Computer Science at Reykjavik University, founding Director of the Icelandic Institute for Intelligent Machines (IIIM; 2009) and co-founder of RU's Center for Analysis and Design of Intelligent Agents (CADIA; 2005).

During his 30+ years of applied and basic AI research, he has taught AI courses at Columbia University, KTH, and RU, consulted for NASA, British Telecom Research Labs, Ipswich, and HONDA Research Labs, and developed AI-based technologies at LEGO and several AI startups, three of which he co-founded. His startup Radar Networks received funding from Microsoft co-founder Paul Allen and was selected one of the 10 most innovative startups in the US by Reuters Venture Capital in 2003. Recently he lead the creation of RU's and Iceland's first academic MSc degree program in theoretical AI.

Dr. Thórisson's research focuses on one of AI's most challenging frontiers, general machine intelligence. His team at RU and collaborators across Europe have developed AERA (Autonomous Empirical Reasoning Architecture), a self-programming system demonstrating cumulative learning abilities with minimal seed knowledge, autonomously learning to conduct goal-driven real-time multimodal TV interviews by observation, and learning how to perceive the world through cameras from first principles. He is a three-time recipient of the Kurzweil Award for his work on general machine intelligence.

Dr. Thórisson holds a Ph.D. from the Massachusetts Institute of Technology.

References

AAU (2025). *Federal Research Cuts Threaten U.S. Innovation and Leadership*. https://www.aau.edu/key-issues/federal-research-cuts-threaten-us-innovation-and-leadership

Adams, T. (2018). *Facebook's Week of Shame: The Cambridge Analytica Fallout*. https://www.theguardian.com/technology/2018/mar/24/facebook-week-of-shame-data-breach-observer-revelations-zuckerberg-silence

Afroogh, S., Akbari, A., Malone, E., Kargar, M., & Alambeigi, H. (2024). Trust in AI: Progress, Challenges, and Future Directions. *Humanities and Social Sciences Communications*, *11*. https://doi.org/https://doi.org/10.1057/s41599-024-04044-8

Aguado, E., Milosevic, Z., Hernández, C., Sanz, R., Garzon, M., Bozhinoski, D., & Rossi, C. (2021). Functional Self-Awareness and Metacontrol for Underwater Robot Autonomy. *Sensors*, *21*. https://doi.org/https://doi.org/10.3390/s21041210

Aiello, M., Alboaie, L., quel, J.-M. J., Mens, K., Motogna, S., Paiva, A. C. R., Schieferdecker, I., & Sora, P. (2026). *The Impact of Curiosity-Driven Informatics Research: An Engine of European Digital Sovereignty* (tech. rep.). Informatics Europe. https://doi.org/DOI:10.5281/zenodo.19659986

Alansari, A., & Luqman, H. (2026). Large Language Models hallucination: A Comprehensive Survey. *Computer Science Review*, *61*. https://doi.org/https://doi.org/10.1016/j.cosrev.2026.100970

Altman, S. (2025). *Reflections*. https://blog.samaltman.com/reflections

Anand, A., Kumar, R., Verma, N., Bhasney, A., & Sharma, N. (2024). A Deep Learning System for Deep Surveillance. In *Mathematical Models Using Artificial Intelligence for Surveillance Systems* (pp. 169–185). https://doi.org/10.1002/9781394200733.ch8

Anderson, J. R., Bothell, D., & Byrne, M. D. (2004). An Integrated Theory of the Mind. *Psychological Review*, *111*, 1036–1060. https://doi.org/10.1037/0033-295X.111.4.1036

ARC Prize Foundation. (2026). *ARC-AGI-3: A New Challenge for Frontier Agentic Intelligence*. https://arxiv.org/html/2603.24621v1

Ashby, W. R. (1958). Requisite Variety and its Implications for the Control of Complex Systems. *Cybernetica*, *1*, 83–99.

AutoNext (2026). *Tesla Sold the Future, Now European Owners Want What They Paid for—Tesla's FSD problem in Europe is Turning Into a Credibility Crisis*. https://www.autonext.co/news/tesla-sold-the-future-now-european-owners-want-what-they-paid-for

Baça, F. (2023). The Importance of Education for Democracy. *Proc. Res. Educ. Sci. (ICRES)*. https://eric.ed.gov/?id=ED654853

Bareš, J. (2025). *Brain vs. llm: the similarities and differences*. https://www.intelligencestrategy.org/blog-posts/brain-vs-llm-the-similarities-and-differences

Bastian, M. (2026). *Former OpenAI Researcher Says Current AI Models Can't Learn From Mistakes, Calling it a Barrier to AGI*. https://the-decoder.com/former-openai-researcher-says-current-ai-models-cant-learn-from-mistakes-calling-it-a-barrier-to-agi/

BBC Nightline (2024, Oct. 8). *'Godfather of AI' on AI "Exceeding Human Intelligence" and it "Trying to Take Over"*. https://youtu.be/MGJpR591oaM?si=GwtPleKiQQ9uLjY1&t=65

BBC (2012, 26 June). *Google Computer Works Out How to Spot Cats*. https://www.bbc.com/news/technology-18595351

Bean, A. M., Payne, R. E., Parsons, G., Kirk, H. R., Ciro, J., Mosquera-Gómez, R., M., S. H., Ekanayaka, A. S., Tarassenko, L., Rocher, L., & Mahdi, A. (2026). Reliability of LLMs as Medical Assistants for the General Public: A Randomized Preregistered Study. *Nature Medicine*, *32*, 609–615.

Bode, K. (2026). *"CEO Said A Thing!" Journalism*. https://karlbode.com/ceo-said-a-thing-journalism/

Bondre, S. V., Thakre, B., Yadav, U., & Bondre, V. D. (2024). Deep reinforcement learning algorithms. In *Deep Reinforcement Learning and Its Industrial Use Cases: AI for Real-World Applications* (pp. 51–73). https://doi.org/10.1002/9781394272587.ch3

Bose, A. (2026). *What Global Banks Said About Anthropic's New Mythos Model That the Company 'Refused' to Release Publicly*. https://timesofindia.indiatimes.com/technology/tech-news/what-global-banks-said-about-anthropics-new-mythos-model-that-the-company-refused-to-release-publicly/articleshow/130551789.cms

Brennan, K., Kak, A., & West, S. M. (2025). *Artificial Power: AI Now 2025 Landscape* (tech. rep.). AI Now Institute. https://ainowinstitute.org/2025-landscape

Burnell, R., Yamamori, Y., Firat, O., Olszewska, K., Hughes-Fitt, S., Kelly, O., Galatzer-Levy, I. R., Morris, M. R., Dafoe, A., Snyder, A. M., Goodman, N. D., Botvinick, M., & Legg, S. (2026). *Measuring Progress Toward AGI: A Cognitive Framework* (tech. rep.). Google DeepMind. https://storage.googleapis.com/deepmind-media/DeepMind.com/Blog/measuring-progress-toward-agi/measuring-progress-toward-agi-a-cognitive-framework.pdf

Cadwalladr, C., & Graham-Harrison, E. (2018). *Revealed: 50 Million Facebook Profiles Harvested for Cambridge Analytica in Major Data Breach*. https://www.theguardian.com/news/2018/mar/17/cambridge-analytica-facebook-influence-us-election

Carnap, R. (1998). The nature of theories. In *Introductory Readings in the Philosophy of Science* (pp. 316–332). Prometheus Books.

Carney, M. (2026). *Davos 2026: Special address by Mark Carney, Prime Minister of Canada*. https://www.weforum.org/stories/2026/01/davos-2026-special-address-by-mark-carney-prime-minister-of-canada/

Carvalho, G., & Thórisson, K. R. (2025). Bad Reasoners, the Turing Trap & the Problem of Artificial Dualism. *Proc. Artif. General Intell.*, 119–134.

Carver, H. (2026). *The Real Danger of LLMs—This Time, AI Might *Actually* Make Us Stupid*. https://medium.com/skiller-whale/the-real-danger-of-llms-e38b72530e7a

Castro, D. P., & Seabrooke, L. (2026). Ownership infinity pools: finding control strategies in multinational enterprises. *Review of International Political Economy*, *33*(2), 993–1016. https://doi.org/10.1080/09692290.2025.2589953

Challapally, A., Pease, C., Raskar, R., & Chari, P. (2025, July). *The GenAI Divide: State of AI in Business, 2025* (tech. rep.). MIT / NANDA, MIT Media Lab.

Chollet, F. (2019). On the Measure of Intelligence. *ArXiv*. https://arxiv.org/abs/1911.01547

CNBC Television (2025, Jan. 21). *Anthropic CEO: More Confident Than Ever That We're 'Very Close' To Powerful AI Capabilities*. https://www.youtube.com/watch?v=7LNyUbii0zw

Coeckelbergh, M. (2026). Technofascism: AI, Big Tech, and the New Authoritarianism. *AI Society*, 1435–5655. https://doi.org/https://doi.org/10.1007/s00146-026-02862-9

Conant, R. C., & Ashby, W. R. (1970). Every Good Regulator of a System Must be a Model of That System. *Int. J. Systems Sci.*, *1*(2), 89–97.

Constantino, M. (2026). *Ribbit ribbit! Artificial Intelligence programs used by Heber City police claim officer turned into a frog*. https://www.fox13now.com/news/local-news/summit-county/how-utah-police-departments-are-using-ai-to-keep-streets-safer

Dellsén, F. (2018). Scientific Progress: Four Accounts. *Philosophy Compass*, *13*.

Dennett, D. (2017). *From Bacteria to Bach and Back: The Evolution of Minds*. W. W. Norton & Company.

Dickson, B. (2024). *Why Vision-Language Models Fail on Simple Visual Tests*. https://bdtechtalks.com/2024/08/01/vlms-visual-test-failures/

Dupoux, E., & LeCun, Y. (2026). *Why AI Systems Don't Learn and What to do About it – Lessons on Autonomous Learning From Cognitive Science* (tech. rep.). Meta.com.

Dupoux, E., LeCun, Y., & Malik, J. (2026). *Why AI Systems Don't Learn and What to do About it — Lessons on Autonomous Learning from Cognitive Science*. https://arxiv.org/html/2603.15381v1

Durt, C., & Fuchs, T. (2024). Large Language Models and the Patterns of Human Language Use. In *Phenomenologies of the digital age* (pp. 106–121). https://doi.org/DOI:10.4324/9781003312284-7

Düz, S. (2025). *Gaza as a Testing Ground: Israel's AI Warfare*. https://www.setav.org/en/gaza-as-a-testing-ground-israels-ai-warfare

Eberding, L., Thompson, J., & Thórisson, K. R. (2024). Argument-Driven Planning & Autonomous Explanation Generation. *Proc. Artif. General Intell.*, 73–83.

EBI (2024). https://www.ebi.ac.uk/training/online/courses/alphafold/an-introductory-guide-to-its-strengths-and-limitations/strengths-and-limitations-of-alphafold/

Erb, R. J. (1993). Introduction to Backpropagation Neural Network Computation. *Springer Nature*, *10*, 165–170.

Feldstein, S. (2022). *China's High-Tech Surveillance Drives Oppression of Uyghurs*. https://thebulletin.org/2022/10/chinas-high-tech-surveillance-drives-oppression-of-uyghurs/

Fichter, A., Meyer, M., Naegeli, L., Oertli, B., & Steiner, J. (2025). *Why Palantir is Becoming a Risky Bet for Switzerland*. https://www.swissinfo.ch/eng/war-peace/why-palantir-is-becoming-a-risky-bet-for-switzerland/90666335

Fichter, A., Meyer, M., Naegeli, L., Oertli, B., Steiner, J., Rzeczy, K., & Imboden, P. (2026). *How Tenaciously Palantir Courted Switzerland*. https://www.republik.ch/2026/02/18/how-tenaciously-palantir-courted-switzerland

Fikes, R. E., & Nilsson, N. J. (1971). STRIPS: A New Approach to the Application of Theorem Proving to Problem Solving. *Artificial Intelligence*, *2*, 189–208. https://doi.org/doi:10.1016/0004-3702(71)90010-5

Foukalas, F. (2025). A Survey of Artificial Neural Network Computing Systems. *Cognitive Computation*, *17*(4).

Gardner, H. (1983). *Frames of Mind: The Theory of Multiple Intelligences*. Basic Books.

Gefen, D., Leffew, B. A., Ragowsky, A., & Riedl, R. (2025). The Importance of Distrust in Trusting Digital Worker Chatbots. *68*, 42–49. https://doi.org/10.1145/3701297

Gelepithis, P. (1988). Survey of Theories of Meaning. *Cognitive Systems*, 141–162.

Gentner, D., & Smith, L. (2012). Analogical Reasoning. In V. S. Ramachandran (Ed.), *Encyclopedia of human behavior, 2nd ed.* (pp. 130–136). Elsevier.

Golan, T., Siegelman, M., Kriegeskorte, N., & Baldassano, C. (2023). Testing the Limits of Natural Language Models for Predicting Human Language Judgements. *Nature Machine Intelligence*, *5*, 952–964.

Good, I. J. (1965). Speculations Concerning the First Ultraintelligent Machine. *Advances in Computers*, *6*, 31–88.

Halpern, J. Y., & Pearl, J. (2013). Causes and Explanations: A Structural-Model Approach – Part 1: Causes. *CoRR*, *abs/1301.2275*. http://arxiv.org/abs/1301.2275

Helliwell, J. F., Layard, R., Sachs, J. D., De Neve, J.-E., Aknin, L. B., & Wang, S. (2026). *World Happiness Report 2026* (tech. rep.). University of Oxford: Wellbeing Research Centre.

Hepburn, B., & Andersen, H. (2026). Scientific Method. https://plato.stanford.edu/archives/spr2026/entries/scientific-method/

Jones, N., & Thompson, B. (2025). *The AI Revolution is Running out of Data. What Can Researchers Do?* https://www.nature.com/articles/d41586-025-00288-9

Kantrowitz, A. (2025). *Google DeepMind CEO Demis Hassabis: The Path To AGI, LLM Creativity, And Google Smart Glasses*. https://www.bigtechnology.com/p/google-deepmind-ceo-demis-hassabis

Kapoor, A. N. S. (2025). *AI as Normal Technology: An Alternative to the Vision of AI as a Potential Superintelligence*. https://knightcolumbia.org/content/ai-as-normal-technology

Klein, G., Moon, B., & Hoffman, R. (2006). Making Sense of Sensemaking 1: Alternative Perspectives. *IEEE Intelligent Systems*, *21*, 70–73. https://doi.org/10.1109/MIS.2006.75

Knight, W. (2026). I Loved My OpenClaw AI Agent—Until It Turned on Me. *Wired*. https://www.wired.com/story/malevolent-ai-agent-openclaw-clawdbot/

Kuhn, T. S. (1962). *The Structure of Scientific Revolutions*. University of Chicago Press.

Kunze, F. (2019). Can We Stop the Academic AI Brain Drain? *33*, 1–3.

Kuperwajs, I., Russek, E. M., Mattar, M. G., Ma, W. J., & Griffiths, T. L. (2023). Looking deeper into the algorithms underlying human planning. *Trends in Cognitive Sciences*, *30*, 149–161.

Laird, J. E., Newell, A., & Rosenbloom, P. S. (1987). Soar: An architecture for general intelligence. *Artificial intelligence*, *33*(1), 1–64.

Laird, J. E., & Wray, R. (2010). Cognitive Architecture Requirements for Achieving AGI. *J. Artif. General Intell.* https://doi.org/DOI:10.2991/AGI.2010.2

Langley, P., Laird, J. E., & Rogers, S. (2009). Cognitive Architectures: Research Issues and Challenges. *Cognitive Systems Research*, *10*, 141–160. https://doi.org/10.1016/j.cogsys.2006.07.004

Legg, S., & Hutter, M. (2007). A collection of definitions of intelligence. *Advances in Artificial General Intelligence: Concepts, Architectures and Algorithms*, *157*, 17–24.

Lessig, L. (1999). *Code and Other Laws of Cyberspace*. Basic Books.

Lewis, D. K. (1970). General Semantics. *Synthese*, *22*(1-2), 18–67.

Luo, L. (2021). Architectures of Neuronal Circuits. *Science*, *373*. https://doi.org/10.1126/science.abg7285

Mahmoud, O., Khalil, A., Karimpanal, T. G., Semage, B. L., & Rana, S. (2026, March). The Unintended Trade-off of AI Alignment: Balancing Hallucination Mitigation and Safety in LLMs. In V. Demberg, K. Inui, & L. Marquez (Eds.), *Findings of the Association for Computational Linguistics: EACL 2026* (pp. 1017–1037). Association for Computational Linguistics. https://doi.org/10.18653/v1/2026.findings-eacl.53

Mankins, J. C. (1995). *Technology Readiness Levels: A White Paper* (tech. rep.). NASA Advanced Concepts Office, Office of Space Access & Technology.

Mayer, M. C., & Pirri, F. (1996). Abduction is Not Deduction-in-Reverse. *Logic Journal of the IGPL*, *4*, 95–108.

McCulloch, W. S., & Pitts, W. (1943). A Logical Calculus of the Ideas Immanent in Nervous Activity. *Bulletin of Mathematical Biophysics*, *5*, 115–133.

Minsky, M. (1952). A Neural-Analogue Calculator Based Upon a Probability Model of Reinforcement. https://en.wikipedia.org/wiki/Stochastic_Neural_Analog_Reinforcement_Calculator

Minsky, M. (1982). *Semantic Information Processing*. MIT Press.

Minsky, M. (1988). *Society of Mind*. Simon & Schuster.

Minsky, M., & Papert, S. (1969). *Perceptrons: An Introduction to Computational Geometry*. MIT Press.

Mitchell, T. M., Keller, R. M., & Kedar-Cabelli, S. T. (1986). Explanation-based generalization: A unifying view. *Machine Learning*, *1*(1), 47–80. https://doi.org/10.1007/BF00116250

Morffis, A. P. (2025). *Artificial Neural Networks Are Nothing Like Brains*. https://blog.apiad.net/p/artificial-neural-networks-are-nothing

Nannini, L. (2024). Habemus a right to an explanation: so what? – a framework on transparency-explainability functionality and tensions in the eu ai act. *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, *7*, 1023–1035. https://doi.org/10.1609/aies.v7i1.31700

National Science Foundation. (2023). *Verbal Nonsense Reveals Limitations of AI Chatbots*. https://www.nsf.gov/news/verbal-nonsense-reveals-limitations-ai-chatbots

NCM. (2017). *Trust – The Nordic Gold* (tech. rep.). Nordic Council of Ministers.

Nedunuri, U., Gupta, A. D., & Guha, D. (2026). The Paradox of Explainability vs. Performance in High-Stakes Autonomous AI Systems: A Systematic Review of Trade-Offs, Regulatory Gaps, and Emerging Solutions. *AI Perspectives & Advances (Springer Nature)*, *8*.

Newell, A. (1973). You Can't Play 20 Questions With Nature and Win (W. Chase, Ed.). *Proc. Visual Info. Proc.*, 396–434.

Newell, A. (1994). *Unified Theories of Cognition*. Harvard University Press.

Nivel, E., Thórisson, K. R., Steunebrink, B., Dindo, H., Pezzulo, G., Rodriguez, M., Corbato-Hernandez, C., Ognibene, D., Schmidhüber, J., Sanz, R., Helgason, H. P., & Chella, A. (2014a). Bounded Seed-AGI. *Proc. Artif. General Intell.*, 85–96.

Nivel, E., Thórisson, K. R., Steunebrink., B. R., Helgason, H. P., Peluzzo, G., Sanz, R., Schmidhuber, J., Dindo, H., Rodriguez, M., Chella, A., Jonsson, G., Ognibene, D., & Hernandez, C. (2014b). Autonomous Acquisition of Natural Language. *Intl. Conf. Intell. Systems & Agents (IADIS)*, 58–66.

Nivel, E., Thórisson, K. R., Steunebrink, B., & et al. (2013). *Bounded Recursive Self-Improvement* (tech. rep. No. RUTR-13006). Reykjavik University School of Computer Science.

Nivel, E., & Thórisson, K. R. (2009). Self-Programming: Operationalizing Autonomy. *Proc. Artif. General Intell.*, 150–155.

Nivel, E., & Thórisson, K. R. (2013a). *Replicode: A Constructivist Programming Paradigm and Language* (tech. rep. No. RUTR-SCS13001). Reykjavik University School of Computer Science.

Nivel, E., & Thórisson, K. R. (2013b). Towards a Programming Paradigm for Control Systems with High Levels of Existential Autonomy. In *Proc. Artif. General Intell.*

Nivel, E., Thórisson, K. R., Dindo, H., Pezzulo, G., Rodriguez, M., Corbato, C., Steunebrink, B., Ognibene, D., Chella, A., Schmidhuber, J., Sanz, R., & Helgason, H. P. (2013). *Autocatalytic Endogenous Reflective Architecture* (tech. rep. No. RUTR-SCS13002). Reykjavik University School of Computer Science.

Nivel, E., Thórisson, K. R., Steunebrink, B., & Schmidhuber, J. (2015). Anytime Bounded Rationality. *Proc. Artif. General Intell.*, 121–130.

OECD. (2026). *Venture Capital Investments in Artificial Intelligence Through 2025*.

Olby, R. C. (1974). *The Path to the Double Helix*. University of Washington Press.

Park, M., Leahey, E., & Funk, R. J. (2023). Papers and Patents are Becoming Less Disruptive Over Time. *Nature*, *613*.

Pearl, J. (2001). Bayesianism and Causality, Or, Why I Am Only a Half-Bayesian. In *Foundations of Bayesianism* (pp. 19–36). Springer.

Pearl, J. (2009). *Causality: models, reasoning and inference* (2nd). Cambridge University Press.

Pollock, J. L. (1991). A theory of Defeasible Rasoning. *Intl. J. Intell. Systems*, *6*(1), 33–54.

Pollock, J. L. (2010). Defeasible Reasoning and Degrees of Justification. *Argument and Computation*, *1*(1), 7–22.

Pollock, J. L., & Cruz, J. (1999). *Contemporary Theories of Knowledge* (Vol. 35). Rowman & Littlefield.

Popper, K. (2005). *The logic of scientific discovery*. Routledge.

Principe, L. M. (2011). *The Scientific Revolution: A Very Short Introduction*. Oxford University Press.

Raffio, N. (2024). *Trust in Voting: How Misinformation Threatens Democracy*. https://today.usc.edu/trust-in-voting-how-misinformation-threatens-democracy/

Rahmanzadehgervi, P., Bolton, L., Taesiri, M. R., & Nguyen, A. T. (2024). Vision Language Models are Blind. *ACCV*. https://anhnguyen.me/2024/vlms-are-blind/?utm_source=substack

Rawnsley, A. (2018). *Politicians Can't Control the Digital Giants With Rules Drawn up for the Analogue Era*. https://www.theguardian.com/commentisfree/2018/mar/25/we-cant-control-digital-giants-with-analogue-rules

Rogelberg, S. (2026). *Thousands of CEOs Admit AI Had No Impact on Employment or Productivity – and it has Economists Resurrecting a Paradox From 40 Years Ago*. https://fortune.com/article/why-do-thousands-of-ceos-believe-ai-not-having-impact-productivity-employment-study/

Rörbeck, H. (2022). *Self-Explaining Artificial Intelligence: On the Requirements for Autonomous Explanation Generation* [Master's thesis, Reykjavik University].

Russell, S., & Norvig, P. (2010). *Artificial Intelligence A Modern Approach*. Pearsons.

Salvaggio, E. (2025). *Generative AI's Productivity Myth*. https://www.techpolicy.press/generative-ais-productivity-myth/

Saputra, H. (2025). Legal Liability of Subsidiaries For Unlawful Actions Committed By The Parent Company (Holding Company) in the Structure of a Limited Liability Company. *Journal of Law, Politic and Humanities*, *5*(6), 4590–4598. https://doi.org/10.38035/jlph.v5i6.2072

Schmidhuber, J. (2015). Deep Learning in Neural Networks: An overview. *Neural Networks*, *61*, 85–117.

Sharma, A., & Chopra, P. (2026). Esolang-Bench: Evaluating Genuine Reasoning in Large Language Models via Esoteric Programming Languages. *I Can't Believe It's Not Better Workshop, ICLR 2026*. https://openreview.net/pdf?id=7KG8ERuf1K

Shojaee, P., Mirzadeh, I., Alizadeh, K., Horton, M., Bengio, S., & Farajtabar, M. (2025). The Illusion of Thinking: Understanding the Strengths and Limitations of Reasoning Models via the Lens of Problem Complexity. *Proc. NeurIPS*. https://arxiv.org/abs/2506.06941

Shumailov, I., Shumaylov, Z., Zhao, Y., land Ross Anderson, N. P., & Gal, Y. (2025). AI Models Collapse When Trained on Recursively Generated Data. *631*, 755–759.

Sifakis, J. (2022). Why is it so hard to make self-driving cars? (Trustworthy autonomous systems). *WI-IAT '21: IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology*. https://doi.org/https://doi.org/10.1145/3486622.0000002

Slater, J. (2025). *LLMs Can be Harmful, Even When Not Making Stuff Up*. https://justice-everywhere.org/education/llms-can-be-harmful-even-when-not-making-stuff-up/

Sloman, A. (2000). Architectural Requirements for Human-Like Agents Both Natural and Artificial: What sorts of machines can love? In K. Dautenhahn (Ed.), *Advances in consciousness research*. John Benjamins Publishing. https://doi.org/10.1075/aicr.19.10slo

Sloman, A. (2010). *Requirements for a Fully-Deliberative Architecture (Or Component of an Architecture)* (tech. rep.). University of Birmingham, CogAff Archive. http://www.cs.bham.ac.uk/research/projects/cogaff/sloman-space-of-minds-84.html

Speaks, J. (2021). Theories of Meaning. In E. N. Zalta (Ed.), *The Stanford encyclopedia of philosophy* (Spring 2021). Metaphysics Research Lab, Stanford University.

Steunebrink, B., Thórisson, K. R., & Schmidhuber, J. (2016). Growing Recursive Self-Improvers. *Proc. Artif. General Intell.*, 129–139.

Stewart, H. (2026). *The Cost of AI Slop Could Cause a Rethink That Shakes the Global Economy in 2026*. https://www.theguardian.com/business/2026/jan/04/ai-reality-growing-economic-risk-2026

Sullivan, M. (2026). *A Top AI Researcher Explains the Limitations of Current Models*. https://www.fastcompany.com/91516330/francois-chollet-ai-benchmark

Sutherland, P., & Chakrabarti, M. (2024). *Inside China's Citizen Spy Network*. https://www.wbur.org/onpoint/2024/06/06/china-citizen-spy-network-government-political-informant-dictatorship

Sutton, R. S., & Barto, A. G. (1998). *Reinforcement Learning: An Introduction*. MIT Press.

Swan, J., Nivel, E., Kant, N., Hedges, J., Atkinson, T., & Steunebrink, B. (2022). *The Road to General Intelligence*. Springer.

Tafazoli, S., Bouchacourt, F. M., Ardalan, A., Markov, N. T., Uchimura, M., Mattar, M. G., Daw, N. D., & Buschman, T. J. (2024). Building Compositional Tasks With Shared Neural Subspaces. *Nature*, *650*, 164–172. https://doi.org/https://doi.org/10.1038/s41586-025-09805-2

Tangermann, V. (2026). *Alarming Study Finds That Most People Just Do What ChatGPT Tells Them, Even If It's Totally Wrong*. https://futurism.com/artificial-intelligence/study-do-what-chatgpt-tells-us

Thórisson, K. R., Rörbeck, H., Thompson, J., & Latapie, H. (2023). Explicit Goal-Driven Autonomous Self-Explanation Generation. *Proc. Artif. General Intell.*, 286–295.

Thórisson, K. R., & Talevi, G. (2024). A Theory of Foundational Meaning Generation in Autonomous Systems – Natural & Artificial. *Proc. Artif. General Intell.*, 188–198.

Thórisson, K. R., & Minsky, H. (2022). The Future of AI Research: Ten Defeasible 'Axioms of Intelligence'. *Proc. Machine Learning Research*, *192*, 5–21.

Thórisson, K. R. (2012). A New Constructivist AI: From Manual Methods to Self-Constructive Systems. In P. Wang & B. Goertzel (Eds.), *Theoretical Foundations of Artificial General Intelligence* (pp. 145–171, Vol. 4). Atlantis Press.

Thórisson, K. R. (2020a). Discretionarily Constrained Adaptation Under Insufficient Knowledge & Resources. *J. Artif. General Intell.*, *11*, 7–13. https://doi.org/10.2478/jagi-2020-0003

Thórisson, K. R. (2020b). Seed-Programmed Autonomous General Learning. *Proc. Mach. Learn. Res.*, 32–70.

Thórisson, K. R. (2021). The Explanation Hypothesis in Autonomous General Learning. *Proc. Mach. Learning Res.*, *159*, 5–27.

Thórisson, K. R., Bieger, J., Li, X., & Wang, P. (2019). Cumulative Learning. *Proc. Artif. General Intell.*, 198–208.

Thórisson, K. R., Kremelberg, D., Steunebrink, B. R., & Nivel, E. (2016). About Understanding. *Proc. Artif. General Intell.*, 106–117.

Thórisson, K. R., & Nivel, E. (2009). Holistic Intelligence: Transversal Skills and Current Methodologies. *Proc. Artif. General Intell. (AGI 2009)*, 220–221.

Thórisson, K. R., & Talbot, A. (2018). Cumulative Learning With Causal-Relational Models. *Proc. Artif. General Intell.*, 227–237.

Time, J. K. (2026). *Avslorer alvorlige KI-feil for norsk offentlig sektor*. https://www.morgenbladet.no/samfunn/avslorer-alvorlige-ki-feil-for-norsk-offentlig-sektor/10259183

Todd, B. (2026). *Will We Have AGI By 2030?* https://80000hours.org/ai/guide/when-will-agi-arrive/

Tómasdóttir, H. (2026). *Gervigreind, ábyrgð og framtíð samfélags okkar*. https://www.visir.is/g/20262866853d/gervigreind-abyrgd-og-framtid-samfelags-okkar

Turing, A. M. (1950). Computing Machinery and Intelligence. *Mind*, *59*(236), 433–460.

Vallentin, A., Thórisson, K. R., & Latapie, H. (2023). Addressing the Unsustainability of Deep Neural Networks with Next-Gen AI. *Proc. Artif. General Intell.*, 296–306.

Wachsmuth, I. (2016). Challenges and Benefits of Interdisciplinary Research and Collaboration – a Report From Germany. *Newsletter of the Icelandic Institute for Intelligent Machines*, *5*.

Wang, P. (2020). On Defining Artificial Intelligence. *J. Artif. General Intell.*, *11*(2), 1–37. https://doi.org/10.2478/jagi-2020-0003

Wang, P. (1995a). *Non-Axiomatic Reasoning System—Exploring the Essence of Intelligence* [Doctoral dissertation, Indiana University].

Wang, P. (1995b). *On the Working Definition of Intelligence* (tech. rep.). Temple University. https://cis.temple.edu/~pwang/Publication/intelligence.pdf

Wang, P. (2007). *Rigid Flexibility: The Logic of Intelligence*. Springer.

Wang, P. (2008). What Do You Mean by "AI"? *Proc. First Artif. General Intell. Conference*, 362–373.

Wang, P. (2013). *Non-Axiomatic Logic: A Model of Intelligent Reasoning*. World Scientific Publishing.

Wang, P. (2025). *Non-Axiomatic Logic: A Model of Intelligent Reasoning (Second Ed.)* World Scientific Publishing.

Wang, P., & Thórisson, K. R. (2025). Concept-Centered Knowledge Representation: A 'Middle-Out' Approach Fusing the Symbolic-Subsymbolic Divide. In P. Robertson & O. Georgeon (Eds.), *International Workshop on Situated Self-Guided Learning (IWSSL-25)* (pp. 10–52). Springer Nature. https://link.springer.com/chapter/10.1007/978-3-031-96325-4_2

Wilding, M., & Hymas, C. (2025). *Police Given AI Tech to Track 'Suspicious' Car Journeys*. https://libertyinvestigates.org.uk/articles/police-ai-anpr-track-drivers/

Wilkins, J. (2026). *You'll Snort-Laugh When You Learn How Much AI Actually Added to the US Economy Last Year*. https://futurism.com/artificial-intelligence/ai-economy-gdp-2025

Williams, A. (2026). *Model Collapse Is Already Happening, We Just Pretend It Isn't*. https://cacm.acm.org/blogcacm/model-collapse-is-already-happening-we-just-pretend-it-isnt/

Wilson, J. R. (1965). *The Mind (Life Science Library)*. Time-Life Education / Almenna Bókaforlagið (1969).

Yu, Q., Tartaglini, A., Hase, P., Guestrin, C., & Pott, C. (2026). *Outcome Rewards Do Not Guarantee Verifiable or Causally Important Reasoning* (tech. rep.). arXiv. arXiv:2604.22074v1%20%5Bcs.CL%5D

Zeve, A. (2025). *Explained: Generative AI's Environmental Impact*. https://news.mit.edu/2025/explained-generative-ai-environmental-impact-0117