An AI Learning Hierarchy

Peter J. Denning
Ted G. Lewis

DRAFT 9/27/24 –v6

A hierarchy of AI machines organized by their learning power shows their limits and the possibility that humans are at risk of machine subjugation well before AI utopia can come.

Peter J. Denning (pjd@nps.edu) is Distinguished Professor of Computer Science at the Naval Postgraduate School in Monterey, CA, USA, is Editor of ACM Ubiquity, and is a past president of ACM. His most recent book is *Navigating a Restless Sea: Mobilizing Innovation in Your Community* (with Todd Lyons, Waterside Productions, 2024). The author's views expressed here are not necessarily those of his employer or the U.S. federal government.

Ted G. Lewis (tedglewis@icloud.com) is the 2021 Oregon State Hall of Famer, author, and computer scientist with expertise in applied complexity theory, homeland security, infrastructure systems, and computer security. He is past director of the Center for Homeland Defense and Security at the Naval Postgraduate School. His most recent book is *Critical Infrastructure Resilience and Sustainability Reader* (2024).

Artificial Intelligence has been successful in numerous areas including speech recognition, automatic classification, language translation, Chess, Go, facial recognition, disease diagnosis, drug discovery, driverless cars, autonomous drones, and most recently linguistically competent chatbots. Yet none of these machines is the slightest bit intelligent and many of the more recent ones are untrustworthy. Businesses and governments are using AI machines in an exploding number of sensitive and critical applications without having a good grasp on when those machines can be trusted.

From its beginnings, AI as a field has been plagued with hype. Many researchers and developers were so enthusiastic about the possibilities that they overpromised what they could deliver. Disillusioned investors twice pulled back during two "AI winters". With the arrival of Large Language Models, the hype has reached new heights and has driven a huge wave of speculative investment in AI companies. Investment advisors are warning of an AI bubble. Many AI researchers have weighed in with concerns that the hype is drawing people to trust machines before

we know enough about them, and to put them into critical applications where mistakes can be costly or deadly.

In 2019 we (the authors) proposed a way to look at AI machines that is objective enough to avoid reliance on hype and anthropomorphism [1]. We found that AI machines can be grouped into classes by learning power. This way of classifying AI machines gives more insight into the trust question than the more common classifications by domains including speech, vision, natural language, games, healthcare, transportation, navigation, and so on.

One aspect of the hype that has particularly troubled us are the claims that recent advances in computing are driven by AI and that all software is a form of AI. It's the other way around: computing has made steady progress in power and reliability over the past half century and most software is not AI. Modern AI would not exist except for those advances.

Another troubling aspect is our tendency to anthropomorphize – to project our beliefs and hopes about human intelligence onto machines. This leads to unwelcome contradictions and misplaced trust in AI. For example, we believe intelligent people think fast, and yet supercomputers that run a billion times faster than humans are not intelligent. We believe that interacting communities of AI machines will be collectively smart, and yet massively parallel computers and networks are not intelligent. We believe chatbots will make new discoveries, but do not accept their outputs as intelligent.

A Hierarchy of Learning Machines

In Table 1, we offer an eight-tiered hierarchy that classifies AI machines by their learning power. A machine is more powerful at learning than another if, in a reasonable time, it can learn to perform some tasks that the other cannot. Learning power comes from structure. This definition does not rely on any notion of intelligence. No anthropomorphizing is needed to explain why one machine is more powerful at learning than another.

This definition also accommodates the two basic ways machines can learn. One is by programming: a designer expresses all the rules of operation in a database and the machine applies these rules to deduce results. The other is by self-adaptation: the machine learns from examples and experience and adjusts its internal structure according to a training algorithm. These approaches can be combined, with part of an AI machine programmed and other parts self-adapting.

This hierarchy does not rank by computational power. All the AI machines are Turing Complete. The hierarchy shows that none of the machines so far built has any intelligence at all, leading to the intriguing possibility that human intelligence is not computable.

Table 1. AI Machines Hierarchy

Level	Category of machines
0	Basic automation
1	Rule-based systems
2	Supervised learning
3	Unsupervised learning
4	Generative AI
5	Reinforcement learning AI
6	Human-machine interaction AI
7	Aspirational AI

Level 0—Basic Automation

These machines are automata that carry out or control processes with little or no human intervention. They frequently include simple feedback controls that maintain stable operation by adjusting and adapting to readings from sensors. For example, an FM radio locks on to a frequency but does not learn what frequencies it recognizes. However, basic automata cannot learn any new actions because their feedback does not change their function – they do not learn anything beyond what they were built to do. All the higher levels are forms of automation augmented with learning.

Level 1—Rule-based Systems

These machines imitate the logic of human reasoning. They were called "rule-based programs" because they made their logical deductions by applying programmed logic rules to their inputs and intermediate results.

Board games were early targets for rule-based programs. In 1952, Arthur Samuel of IBM demonstrated a competent, self-improving checkers program. Beginning in 1957, a long line of chess research led to the IBM Deep Blue computer, which, in 1997 beat grandmaster Garry Kasparov. Computer speed is essential – the computer evaluates thousands of next moves in the same time a human can evaluate just one.

Expert systems were another early target– programs using logic rules derived from the knowledge of experts. Early examples were developed by Edward Feigenbaum at Stanford University in 1965: Dendral identified unknown organic molecules, and Mycin diagnosed infectious blood diseases. In 1980 John McDermott

of Carnegie Mellon University built XCON, which determined the best configuration of complex DEC computer systems for a given customer.

Expert systems designers soon discovered that getting experts to state their expertise as rules is an impossible task. Hubert Dreyfus, a philosopher and an early critic of expert systems, argued that much of what we call expertise is not rule based: a machine limited to rule-based operations could not be expert [2]. Not even an enormous database of common-sense facts could make these systems as smart as experts. Many expert systems are useful despite this weakness.

Level 2—Supervised Learning

These machines do not apply logic rules to inputs. Instead, they remember in their structure the proper output for each input shown it by a trainer. The artificial neural network (ANN) is the common example. The ANN trainer presents a long series of input-output examples; it adjusts the internal connection weights to minimize error between the actual and intended outputs. Although training may take days, a trained network responds within milliseconds.

An important property of ANNs is that any continuous mathematical function can be approximated arbitrarily closely by a sufficiently large ANN trained with a sufficient number of input-output pairs. This has inspired much research into ANNs to implement differential-equation models of physical systems, leading to many improvements in scientific computing.

In many applications, the data do not come from a continuous function – for example, facial recognition trained by labelled images. These ANNs have two main limitations: fragility and inscrutability. Fragility means that, when presented with a new (untrained) input that differs only slightly from a trained input, the network may respond with a wildly wrong output. Inscrutability means that it difficult or impossible to "explain" how the network reached its conclusion.

Level 3—Unsupervised Learning

These machines improve their performance by making internal modifications without the assistance of an external training agent. Classifiers are the most common examples. A classifier divides the input data into the most probable set of classes by similarity; no classes are specified in advance. An early example is the AUTOCLASS program by Peter Cheeseman that classified space telescope profiles of stars.

Level 4 – Generative AI

Machines of this level are ANNs augmented with natural-language processors. The training process presents a large corpus of text and records which words are near to each other. When presented with an input text ("prompt"), the basic ANN

produces an output word that is highly likely to be next after the input. That word is appended to the prompt and the cycle repeats, generating an output string of words that is highly probable given the original prompt. The basic model is fluent but likely to generate nonsense or fabrications that are not in the training data. The basic network is modified by a "tweaking process" that adjusts weights to reduce the chances of these unsatisfactory outputs.

Generative AI systems are often called Large Language Models because they are trained on a very large textual training set. One of the most prominent of this genre, ChatGPT-4, was trained on several hundred billion words of texts found on the internet; training took several months and consumed as much electricity as a small town. The results were astounding. LLMs can give astonishingly competent outputs. But they are so prone to generating fabrications and nonsense that Emily Bender in 2021 called them "stochastic parrots". Many people do not trust them, especially when they make recommendations for action in critical areas where mistakes are costly.

There is a controversy around whether Generative AI machines are creative. Skeptics point to many human creations that are not inferences from prior knowledge.

Level 5 – Reinforcement Learning

These machines avoid the need for massive training data. Reinforcement teaches an ANN how to achieve a goal. Two ANNs play rounds of a game with each other, keeping track of which moves were ultimately part of a win and adjusting parameters so that the machines gradually learn to select only winning moves. This is done with millions or billions of rounds, simulated on an energy-gobbling supercomputer. It can produce amazing results. DeepMind's AlphaZero became a Chess grandmaster in 4 hours and Go grandmaster in 13 days with reinforcement learning. OpenAI's ChatGPT uses reinforcement learning to make final adjustments to the weights in its core ANN so that the responses are more satisfactory to humans.

Level 6—Human-Machine Interaction

It is generally agreed that humans and machines blending together are more powerful than either working alone. Humans are particularly good with judgments and machines with computations. Achieving good blends is a very difficult problem in design.

One approach to this was popularized by Marvin Minsky in his book *Society of Mind* [5]. The idea is that thousands or millions of agents, each trained to be good at a narrow human skill, cooperate together and collectively generate results better than any human (or individual machine). This idea permeates many proposals for achieving Artificial General Intelligence (AGI).

Another approach, pioneered in the 1960s by Doug Engelbart, was based on the idea of amplifying human intelligence by augmenting humans with machines. In his day, the machines were external devices using tools such as windows, mice, and hyperlinks. Today the augmentation tools are much more sophisticated and include smartphones, virtual reality glasses, and simulations. After IBM Deep Blue beat him in 1997, Garry Kasparov invented Advanced Chess, where a "player" is a team consisting of a human augmented by a computer. It was soon found that the teams of competent players and good chess programs were able to defeat the best machines. According to futurist Ray Kurzweil, in the next decade or two augmentations may include nanobots introduced into the human bloodstream that interface with external computers and provide organ repair and enhancements like photographic memory [4].

These examples show that human-machine teaming is a rich area and can often be achieved with simple interfaces that do not rely on AI tools.

Level 7—Aspirational AI

This level includes a variety of speculative machines that represent the dreams of many AI researchers. The most ambitious feature machines that think, reason, understand, and are self-aware, conscious, self-reflective, compassionate, and sentient. No such machines have ever been built and no one knows whether they can be built [3].

AI Progress Models

The AI hierarchy can be seen as a progress model. As machines gain in learning power, they approach AGI.

In *The Last AI* (2024), S M Sohn lays out a progress model depicted as a pyramid of increasing automation from AI (Table 2) [7]. He envisions that automation will make basic necessities abundant and cheap, leading eventually to 0-person organizations (no humans involved in running things) and AI utopia. While some consider this model to be preposterous, we take it seriously – as a very plausible path to a society of human subjugation by unintelligent machines.

Table 2. Sohn's AI Adoption Hierarchy

Level	Category of machines "in charge of"
1	Human business roles (AI copilot, AI assistant)
2	Machine business roles (Ai agent, AI butler)
3	Business (AI CEO, AI company)
4	Government (AI president, AI bureaucracy, AI congress)

The process seen by Sohn is already well underway at all four levels – CoPilot and LLMs at Level 1, business workflow automation at Level 2, automated purchasing and customer service at Level 3, and automated bureaucracies and political deepfakes at Level 4. These systems are already distrusted because of their rigidity, fragility, lack of care, lack of compassion, and intolerance of human errors. We are drifting toward a new singularity– the subjugation of humans to networks of low-intelligence, uncaring machines– well before Kurzweil's Singularity merges humans with machines in 2045.

Inspired by Sohn, the OpenAI company promoted its own progress hierarchy, its roadmap to safe and beneficial AGI (Table 3) [6]. It is a business plan for the new singularity! Our collective eagerness to push toward AGI may accelerate our prospect of being sucked into the quicksand of machine orchestrated stupidity.

Table 3. OpenAI's Adoption Hierarchy

Level	Category of machines "in charge of"
1	Chatbots (AI with conversational language)
2	Reasoners (human-level problem solving)
3	Agents (systems that can take actions)
4	Innovators (AI that aids in innovation)
5	Organizations (AI doing the work of organizations)

Bibliography

1. Denning, P.J. and T.G. Lewis. 2019. Intelligence might not be computable. *American Scientist 107* (Nov-Dec), 346-349.
2. Dreyfus, H. 1972. *What machines (still) cannot do.* MIT Press. [Updated in 1978 and 1992.]
3. Koch, Christof. 2019. *The Feeling of Life Itself: Why Consciousness is Widespread But Can't Be Computed.* MIT Press.
4. Kurzweil, R. 2024. *The Singularity is Nearer: When We Merge with AI.* Viking.
5. Minsky, M. 1986. *The Society of Mind.* Simon and Schuster.
6. Sohn, S. M. 2024. Comments on OpenAI's Adoption Model. https://medium.com/@The_Last_AI/openais-new-5-stages-of-ai-development-agi-and-the-ai-adoption-pyramid-454c3e773e2d
7. Sohn, S. M. 2024. *The Last AI: Of Humans Climbing the AI Pyramid*. SM Research Institute.