Preprint, to appear in Ximian Xu & George Coghill (Eds.), *"T&T Clark Handbook of Artificial Intelligence, Religion, and Ethics"*, T&T Clark, (to appear 2026).

AI: Meaning, Scope, and Outlook

Jan Segessenmann and Thilo Stadelmann

Abstract

This chapter introduces modern AI systems based on their applications and core working principles, with a focus on 'large language models' (LLM). Core methods are explained, illustrated, and traced back to their historical origin and development. A short survey of current definitions shows a lack of cohesion. We argue that diverging definitions reflect differences in philosophical presuppositions more than disagreement over evidence, and we propose the cautious definition of AI as the simulation of behaviour usually associated with human intelligence. We conclude that questions about AI's meaning and future ultimately point toward enduring questions about human nature, agency, and reality – questions that invite engagement with the humanities and religious traditions alongside technical expertise.

Keywords

Artificial Intelligence, Machine Learning, Agentic AI, Definition, Future, Narrative

Introduction: AI in Application

'Artificial Intelligence' (AI) has become a widely used phrase for designating a subfield of computer science, for grouping technologies with certain characteristics, and also for making narrative claims about human nature, our future, and reality as such – from stock market predictions to eschatological visions of infinite bliss.[1] Thus, before pinning down the meaning of this complex notion more closely, let us first look at some central examples of AI that illustrate what it refers to.

Perhaps the most common contemporary example associated with AI is chatbots, such as ChatGPT by OpenAI. Chatbots with the ability to generate largely consistent responses to prompts were first deployed in 2022 and have since become an integral part of everyday life. The introduction of 'large language models' (LLM) behind these successful chatbots is illustrative for the meaning of AI on multiple levels.[2] From an engineering point of view, it marks a significant step toward the long-standing problem of dealing with the fuzziness and indeterminacy of human language. At the same time, linguistic capacity has been thought of as a defining feature of human nature since

[1] Not to say that narratives are not also involved in science and in the grouping of technologies. Ultimately, human beings are narrative beings, and no sharp distinction can be drawn between narrow definitions (e.g., by scientists and engineers) and grand narratives (e.g., by transhumanist prophets). See, e.g., Ricoeur, *Time and Narrative*; Taylor, *Sources of the Self*; MacIntyre, *After Virtue*.

[2] For 'LLM', see Brown et al., "Language Models Are Few-Shot Learners." For 'chatbot', see Ouyang et al., "Training Language Models to Follow Instructions with Human Feedback."

Greek antiquity.[3] Thus, as the technological domain seemingly enters the human domain, the notion of AI also questions human identity. Furthermore, the vast range of applications, the enormous demand of resources, and the great potential for concentration of power add to the notion a political, economic, and cultural dimension.[4]

Before chatbots, AI already stirred scientific investigation, for example, with the introduction of AlphaFold, a computational model that successfully predicts how proteins fold based on their amino acid sequence.[5] Protein-folding is an immensely complex problem and has been a challenge to biology for decades, with no feasible solution in sight. And even now, little is known about *how* AlphaFold arrives at certain predictions. Nevertheless, its predictions can be used, for example, in drug design.[6] Note how AI sheds a new light not only on human identity, but also on the nature of scientific investigation. By successfully processing amounts of data beyond human comprehension, AI provides breakthroughs of utility even if knowledge stagnates. In a similar way, AI guides medical diagnosis and lends clinical decision support;[7] it recommends content and personalizes platforms for entertainment and socializing, such as YouTube and Instagram.[8] The great influence by which AI guides millions of people through the internet can hardly be overstated.

Another illustrative example of AI is autonomous driving. Here, in contrast to the previous examples, AI is thought of as a physical presence of its own.[9] Equipped with numerous sensors and actuators, AI in the form of autonomous cars increasingly fills our streets, most notably, in the US. Such 'physical AI' is considered to be the fastest-growing sub-field of AI in the coming decade.[10] Having considered some examples of AI – and there would be numerous others, from document recognition based on computer vision methodologies, to industrial applications based on prediction and control, to wayfinding on Google maps and generating interactive, immersive 3D virtual worlds –, let us shift the focus to how current AI works.[11]

---

[3] The influential Greek term — mostly associated with Aristotle but never used by him explicitly — is *zōon logikon*, which is often translated to 'rational animal', but more accurately refers to something like 'animal possessing language', as pointed out by Taylor, *The Language Animal*, 338.
[4] See Rashid and Kausik, "AI Revolutionizing Industries Worldwide."
[5] See Jumper et al., "Highly Accurate Protein Structure Prediction with AlphaFold."
[6] See Eisenstein, "Artificial Intelligence Powers Protein-Folding Predictions."
[7] See Jermain et al., "Deep Learning-Based Cell Segmentation for Rapid Optical Cytopathology of Thyroid Cancer."
[8] See Covington, Adams, and Sargin, "Deep Neural Networks for YouTube Recommendations."
[9] See Liu et al., "Aligning Cyber Space with Physical World."
[10] See Li et al., "Physical Artificial Intelligence (PAI)."
[11] See Meyer et al., "A Document Is Worth a Structured Record"; Yan et al., "Learning Actionable World Models for Industrial Process Control"; Lanning, Harrell, and Wang, "Dijkstra's Algorithm and Google Maps"; cf. Google DeepMind, "Genie," https://deepmind.google/models/genie/.

The state of the art: How modern AI works

The field of AI does not provide a unified theory. Instead it offers a vast array of different methods, comparable to a toolbox, dependent on the intended purpose at hand.[12]

Arguably, the most prominent method of AI in the last two decades has been "machine learning" (ML).[13] ML is used if a problem cannot be solved by a set of explicit rules. Consider classifying a set of images into the categories 'cat' and 'dog' (see Figure 1): We cannot find a set of rules – but give a set of examples, consisting of images and corresponding category labels that serve as desired outputs, called *targets* (or: *labels*). The ML-system can be thought of as a large mathematical function implemented by what is called an 'artificial neural network' (ANN) that maps data from the input domain (input) to the target domain (output). Before learning, the *parameters* that shape this function are random and any outputs would only match the targets by accident. The function only systematically takes shape during learning on input-target examples, such as images and category labels. Every error the current function produces provides some information about how to slightly nudge the parameters of the function towards better results in the future, that is, outputs closer to the target. Note that outputs are not given directly as categories (such as the binary of 'cat' and 'dog'), but as probabilities of categorization (such as '80% cat' and '20% dog'). Thus, once learned, such a function implements the probability of possible targets given the image: It is a probabilistic mapping of its inputs to outputs, or put yet more technically, it represents a probability density function.[14]

Importantly, ML works not only with simple visual features. One can put in all the pixels of an image and let an ANN learn a very complex relationship between input and target. Generally speaking, the more complex a relationship, the more parameters are needed for enabling a correspondingly complex function, the more example data is needed for providing sufficient information about the statistical relationship.[15]

ML also works with other input-target pairs, for example, text as input and its likely continuation as target. This gives us a LLM. Such a model maps the following relationship: The probability of the next word given the context of words in the input.[16] This context can become as large as millions of words and the respective function will

---

[12] The following explanations are presented in greater, still accessible detail in Stadelmann, "A Guide to AI"; cf. https://stdm.github.io/How-not-to-fear-AI/. An accessible introduction to AI particularly directed to scholars in the humanities can be found in Segessenmann et al., "Assessing Deep Learning."

[13] For further reading, see Jordan and Mitchell, "Machine Learning: Trends, Perspectives, and Prospects."

[14] For a deeper explanation of the nature of ML models as probabilistic functions and the implications on the results and their use, see Stadelmann, Merkt, and Barr, "The Stochastic Nature of Machine Learning and Its Implications for High-Consequence AI."

[15] The insight that ML performance scales with compute, data, and model 'capacity' drove much of AI progress since the 2010s. For the formulation of the related empirical scaling laws, see Hestness et al., "Deep Learning Scaling Is Predictable, Empirically."

[16] More precisely, instead of words, LLMs operate on the level of sub-word units called 'tokens'.

have billions of parameters to fit such data well. It needs approximately a full internet of text to be well trained. Given the current importance of LLMs, let us take a closer look at how they work.

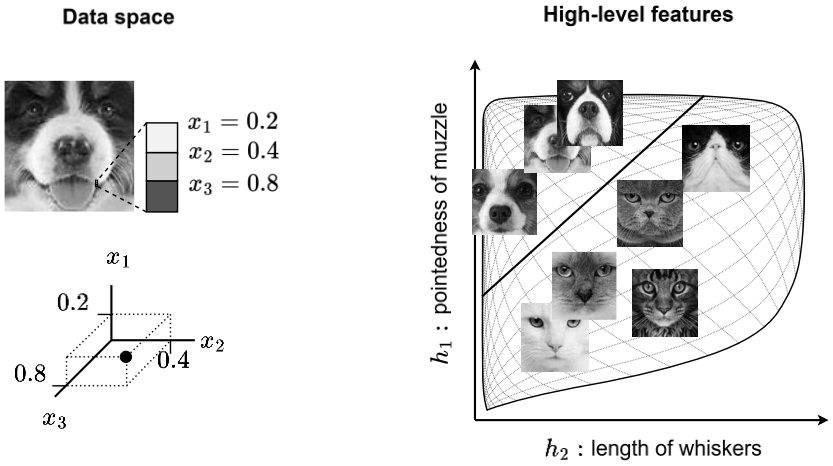

Figure 1: A greyscale image can be thought of as one single point in multidimensional data space. This is illustrated on the left on three pixels only, corresponding to a point in 3D data space. An ML-function can be thought of as transforming (e.g., squeezing, twisting, reducing, expanding) data space from the input domain to the target domain. The result is illustrated on the right, where this transformation has resulted in two high-level features (h1, h2) that allow to separate the input data along a straight decision line that separates higher probabilities for 'cat' and 'dog', respectively.

Training (building) a LLM proceeds in three basic phases (illustrated in Figure 2). First, *dataset assembly*. The data (text) is first broken into many small learning examples of input (several consecutive words) and corresponding target (the immediately following word) to learn next-word prediction. Each pair will teach the model how words tend to follow one another and by that the statistical structure of language is learned.[17]

The second phase is called *unsupervised pre-training* and consists of using the constructed examples extensively to train for correct next-word prediction. It does not teach explicit grammatical rules or symbolic representations of meaning. Rather, it teaches probabilistic patterns: which words, phrases, and structures tend to co-occur, and in what contexts – which surprisingly serves as a useful surrogate for meaning if learned 'at scale.'[18]

[17] That this simple paradigm would lead to the comprehensive language handling capabilities observed today was not obvious up to GPT-3 in 2020, and linguists habitually denied the sheer possibility of it. It is rather one of the emerging properties of large foundation models that no one actually planned for or explicitly provided structure or training for in the model. See Wei et al., "Emergent Abilities of Large Language Models."

[18] The training of models like GPT-4 reportedly cost around 100M USD, where the largest part of compute was used up by the unsupervised pre-training phase. See Cottier et al., "The Rising Costs of Training Frontier AI Models."

Pre-training is followed by the third phase, *preference alignment* by learning from human feedback. Here, human evaluators review and compare complete model outputs (not just next words), indicating which responses are more helpful, accurate, or appropriate, according to a set of rules or standards given them by the organization responsible for the training.[19] The system is then fine-tuned to align its predictions with these human preferences. Rather than explicitly endowing the model with intentions, values, or guardrails in a human sense, this phase reshapes the probability landscape the model uses when selecting words, making certain types of responses more likely and others less so.

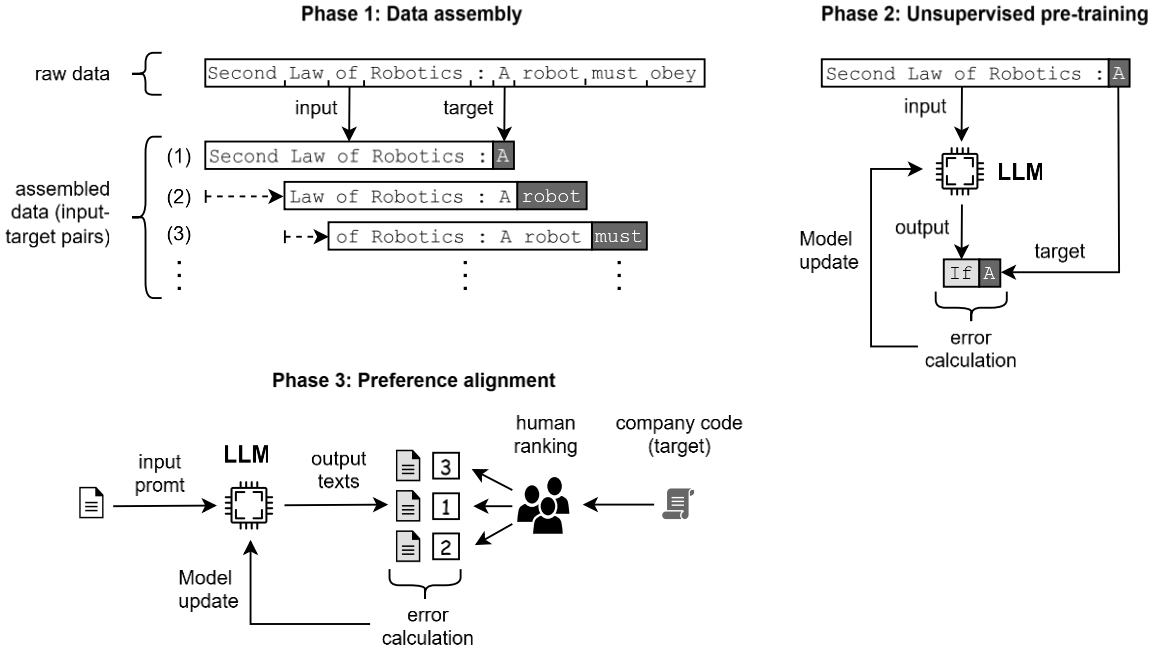

Figure 2: Illustration of the three phases of training an LLM. The text is taken from Asimov, I, Robot*, 40.*

Since scaling saw diminishing returns since 2024, model developers extended the basic LLM concept toward what are often described as 'reasoning' models.[20] The underlying mechanism, next-word prediction, remains unchanged. What differs is how the model is prompted (or prompting itself after an initial user prompt) and how much computation is invested at inference time (that is, when the system is generating an answer). Instead of producing an immediate response, the model can be guided to generate intermediate steps in a dialogue with itself – to break a problem down, to articulate sub-questions, to retrieve additional information, or even to reconsider and revise earlier steps if they appear inconsistent. This self-prompting effectively creates a richer, better-conditioned context before the final answer is produced. By allocating

[19] See Ziegler et al., "Fine-Tuning Language Models from Human Preferences."
[20] See Wei et al., "Chain-of-Thought Prompting Elicits Reasoning in Large Language Models."

more computational effort during generation, the system increases the likelihood that its next-word predictions will align with a coherent solution.

In this sense, reasoning models do not represent a fundamentally different kind of AI system. Rather, they reflect a design pattern in which iterative prompting and expanded context are used to enhance the reliability and structure of the model's outputs. This design pattern can be extended to the contemporary notion of 'agentic AI', in which foundation models like LLMs build the core of a system that has access to tools like web browsing, a programming language interpreter, or computer control to act on the user's behalf.[21] At the same time, the idea of this 'loop' around the basic next-word prediction is used to employ the models for longer-term planning, and its language handling capabilities are used for multi-agent collaboration. This way, the fundamentally limited LLM paradigm can ascend to unexpected heights of performance through an outer loop that enables such agents to follow certain human patterns of acting. However, the likelihood of error always persists such that long-term autonomy of today's agents inevitably leads to failure. [22]

Hence, a next frontier in AI is to overcome such major limitations, including hallucinations, lack of generalizability, lack of causal reasoning, lack of continual learning capabilities, the huge inefficiency to learn from training data and the difficulty of generating goal-driven outputs or behavior. LLMs are unfit for this, for example, because they draw their power from interpolating between their vast training data but lack an embodied and causal understanding of the words they use. A possible way forward is seen in 'world models': compositional internal representations of the environment that are learned from experience and exhibit more 'common sense' than LLMs since they have been trained to explicitly model cause and effect (i.e., predict the consequence of actions). Next to (and especially in combination with) physical AI, this is one of the current hot beds of AI research.[23]

Such more explicit modelling is not new to AI. In fact, it has only been largely displaced from the headlines by the success of ML-based methods from the 2000s onwards. This brings us to some notoriously recurring topics in the history of AI.

---

[21] For 'agentic AI', see Dao et al., "Agentic Design Patterns"; Sager et al., "A Comprehensive Survey of Agents for Computer Use."

[22] While it can yield better results on average than the average human response, it might be wrong in any individual case due to its stochastic nature: it does not 'know' but outputs statistical plausibility. For a detailed discussion, see Stadelmann, Merkt, and Barr, "The Stochastic Nature of Machine Learning and Its Implications for High-Consequence AI."

[23] See Ha and Schmidhuber, "Recurrent World Models Facilitate Policy Evolution"; Balestriero and LeCun, "LeJEPA: Provable and scalable self-supervised learning without the heuristics."

History: Methods and principles

The history of AI is sometimes traced back all the way to Greek mythology. However, drawing continuity, for example, between the bronze giant Talos and contemporary humanoid robots risks imposing a modern imagination to the antique mind. It is true that human beings have been viewing themselves in the image of the technology of their time – in Descartes' time, for example, these were organ pipes or automata in the Garden of Versailles.[24] However, it is only after Alan Turing, that the analogy between human intelligence and machine processes became a serious operational hypothesis.[25]

In 1936, Turing conceptualized the 'universal Turing machine', a machine that could be functionally so organized as to fulfil any imaginable logical task.[26] Later, in 1950, he linked 'computing machinery' with 'intelligence', famously proposing that if a machine behaves intelligently, there is nothing preventing us from inferring that it genuinely *is* intelligent, as we lack theoretic criteria to validate the presence of intelligence.[27] This proposition has been very influential not only for AI but also for cognitive science. Turing's ideas have later led to the realization of the first digital computer, together with John von Neumann's practical design and McCarthy's list processing language (LISP) which he based on the Logic Theorist of Allen Newell and Herbert Simon.[28] Note that the histories of AI and cognitive science are closely related to the invention of the computer and all three fields share important figures and ideas. In 1976, Newell and Simon proposed their influential 'physical symbol system' conceptualizing human thinking in terms of a system that manipulates symbols.[29] It is crucial to understand the dual role of 'physical symbols', which enables them to bridge the mental and physical realms. Physical symbols are constituted by a physical structure corresponding to, or *representing*, a meaningful content in the mental realm, such as 'apple' or 'tree'. Consequentially, symbols can be manipulated based on the principles of physics while this manipulation structurally corresponds to mental processes. In this way, thinking of a tree can be reduced to a physical process. However, what is important is not the underlying materiality such, that is, whether it involves valves, gears, or transistors, but the *functional organization*. This line of thought has gained popularity within AI research under the names of 'symbolic AI', 'expert systems', or 'Good Old-Fashioned AI'

[24] See Descartes, *Treatise of Man*.
[25] For example, Margaret Boden notes that before Turing's seminal papers (see below), "no one had yet argued that mind and/or mental processes, conceptualized as somehow distinct from matter, could be understood in machine-based terms." Boden, *Mind as Machine*, 168.
[26] Turing, "On Computable Numbers, with an Application to the Entscheidungsproblem."
[27] Turing, "Computing Machinery and Intelligence."
[28] Newell and Simon, "The Logic Theory Machine."
[29] Newell and Simon, "Computer Science as Empirical Inquiry."

(GOFAI). Simultaneously, another line of thought has emerged that approaches AI quite differently and has thus sometimes been seen as fiercely opposing symbolic AI.

This second line of thought also followed Turing but linked his ideas with work in neurophysiology and statistics.[30] In 1943, Warren McCulloch and Walter Pitts theorized an artificial neuron called 'logic threshold unit'.[31] They could show that for every logical (Boolean) function, there exists a network of units that could compute it. This caught attention as it seemed to demonstrate that the human brain can in principle be imitated by "telegraph relays" and "vacuum tubes".[32] Shortly later, Donald Hebb introduced a learning rule that enabled connection strengths between units to be adapted such that a network of units could be trained to yield a certain output based on a certain input.[33] In 1962, Frank Rosenblatt introduced the 'perceptron' which brought these theories into a form that allowed for their implementation on a computer.[34] The perceptron was very attractive because in contrast to symbolic approaches, the function it computes is not fixed. Instead, its function depends on the relative connection strengths between units, which allows for associative learning. Nevertheless, the perceptron faced many technical obstacles and was eventually abandoned by many in the 1970s (the first 'AI winter').[35] However, it has later been redeemed and became the basic principle of the ANN still dominating AI today. In fact, it was the discovery that adding many layers to an ANN (thereby making them 'deep') together with a suitable learning algorithm that resulted in the stunning success of 'deep learning' (DL) in the 2000s, which brought AI to public attention at large.[36] In the history of AI research, this line of thought is associated with 'connectionist AI', 'subsymbolic AI', or 'parallel distributed processing' (PDP).[37]

In sum, symbolic AI approaches AI in terms of logic and rules corresponding to human reason. Connectionist AI approaches AI in terms of function approximation by input–output association, which is a data-driven approach.[38] That this leveraging of data and compute always outcompeted any clever approach to engineer intelligence from first principles and biological analogy became known in the field as the "bitter lesson."[39]

---

[30] And often worked in parallel to developments made in the field of cybernetics, which parted ways with AI because of animosities between their prominent figures; cf. Fradkov and Shepeljavyi, "The History of Cybernetics and Artificial Intelligence"; and Wiener, *Cybernetics*.
[31] McCulloch and Pitts, "A Logical Calculus of the Ideas Immanent in Nervous Activity."
[32] See von Neumann, "First Draft of a Report on the EDVAC," 5.
[33] Hebb, *The Organization of Behavior*.
[34] Rosenblatt, *Principles of Neurodynamics*.
[35] See Nilsson, *The Quest for Artificial Intelligence*.
[36] Important contributions include Hinton, Osindero, and Teh, "A Fast Learning Algorithm for Deep Belief Nets"; Bengio et al., "Greedy Layer-Wise Training of Deep Networks"; Ranzato et al., "Efficient Learning of Sparse Representations with an Energy-Based Model"; Schmidhuber, "Deep Learning in Neural Networks."
[37] See Schmidhuber, "Annotated History of Modern AI and Deep Learning."
[38] See Stadelmann, Klamt, and Merkt, "Data Centrism and the Core of Data Science as a Scientific Discipline."
[39] See Sutton, "The Bitter Lesson."

Today, both rivalling approaches still have their fervent advocates, although the success of connectionist AI with DL and LLMs has put a great asymmetry to this relation. However, both still have their strengths and weaknesses. The former provides stable and reliable outcomes but often lacks methods to deal with the fuzziness of complex real-world tasks, such as simulating language. The latter often succeeds with the fuzziness but lacks stability and reliability. These opposing strengths and weaknesses are systematically *inherent* to the respective approaches. As a consequence, there have also been voices that call for 'hybrid', so-called 'neuro-symbolic' approaches to AI.[40] Others argue that connectionist AI will eventually incorporate stringent logic on abstract levels.[41]

It is interesting to note that since the beginnings of the field, AI researchers have both been concerned with pragmatically solving complex problems by computational means; and with dreams of achieving 'human intelligence' and engineering 'persons'. Both mindsets continue to play an important role today, although they are subject to much controversy. This becomes more obvious when we consider how the field has been defining AI.

Definitions: A short survey

Etymologically, 'artificial' derives from the Latin *artificialis,* meaning something like 'crafted by human skill', which contrasts with *naturalis*, that is 'by nature'. The corresponding Greek pair is *technē–physis*. For Aristotle, things according to nature (*kata physin*) have their principle of motion in themselves, while artificial things (*kata technēn*) have their principle of motion in the craftsman or artisan, that is, outside themselves.[42] 'Intelligence' comes from the Latin verb *intelligere* which translates to 'to understand' or 'to discern'. It is therefore traditionally associated with an activity of the human mind which belongs to natural beings (*kata physin*) and cannot be something artificial (*kata technēn*). In this light, the term 'artificial intelligence' contains a tension which is alive and well in today's debates about the meaning of AI.

The term was officially coined by John McCarthy in the context of the 1956 Dartmouth conference commonly associated with the birth of the field. The conference was convened based on the premise that "any [...] feature of intelligence can in principle be so precisely described that a machine can [...] simulate it". Furthermore, the "artificial intelligence problem" was taken to consist of "making a machine behave in ways that would be called intelligent if a human were so behaving."[43] McCarthy later defined AI as "the science and engineering of making intelligent machines," while "[i]ntelligence is

[40] See, e.g., Bhuyan et al., "Neuro-Symbolic Artificial Intelligence."
[41] Although with reservations. See, e.g., Clark, "Predicting Peace."
[42] Aristotle, The Physics, II.1.
[43] McCarthy et al., "A Proposal for the Dartmouth Summer Research Project on Artificial Intelligence."

the computational part of the ability to achieve goals in the world." As such, for McCarthy, intelligence occurs in "[v]arying kinds and degrees [...] in people, many animals and some machines."[44] In his seminal book on the history of AI, John Haugeland noted that "[t]he fundamental goal of this research is not merely to mimic intelligence or produce some clever fake. Not at all. 'AI' wants only the genuine article: machines with minds, in the full and literal sense." Consequentially, "we are, at root, *computers ourselves*."[45] Later, in 2007, Shane Legg and Marcus Hutter collected definitions of 'intelligence' from notable figures in AI research.[46] Pulling many strings together, Legg and Hutter adopt the following definition: "Intelligence measures an agent's ability to achieve goals in a wide range of environments."[47] Note that this is a very different definition and a much more moderate claim compared to, for example, Haugeland's. More recently, Christopher Collins et al. have reviewed 1877 studies on AI in information systems research between 2005 and 2020, looking for definitions of AI. They identify a "lack of cohesion when defining AI."[48] Furthermore, they observe a trend in definitions to refer to the *capabilities* of AI. For example, most studies on AI derive their understanding from Stuart Russell and Peter Norvig, for whom AI – in the words of Collins et al. – "enables the machine to exhibit human intelligence, including the ability to perceive, reason, learn, and interact, etc."[49] However, for Collins et al., AI should not be taken to mean 'artificial *human* intelligence'. Instead, they propose that 'intelligence' is a more general notion, which pertains to human beings and machines in a different way. As a consequence they find a more 'robust' definition of AI provided by Rai et al. in "the ability of a machine to perform cognitive functions that we associate with human minds, such as perceiving, reasoning, learning, interacting with the environment, problem solving, decision-making, and even demonstrating creativity."[50] This is much closer to the Dartmouth conference proposal.

As this short and inexhaustive survey shows, there indeed seems to be a 'lack of cohesion' as to what AI means and most notably, how AI relates to human intelligence. Some use 'intelligence' in AI and humans *univocally*, that is, they regard AI and human intelligence as essentially one and the same thing (e.g., Haugeland). For others the relation is rather one of *analogy*, whereby AI has a share in human intelligence, or they both have a share in a yet more abstract and general notion of intelligence (e.g.,

---

[44] McCarthy, "What Is Artificial Intelligence?"
[45] Haugeland, *Artificial Intelligence: The Very Idea*, 2. Emphasis in original.
[46] To name just a few, for Marvin Minsky, intelligence means "the ability to solve hard problems." For Allen Newell and Herbert A. Simon, intelligence lends "behavior appropriate to the ends of the system and adaptive to the demands of the environment [...] within some limits of speed and complexity." For Ben Goertzel, it is simply "[a]chieving complex goals in complex environments." See Minsky, *The Society of Mind*; Newell and Simon, "Computer Science as Empirical Inquiry"; Goertzel, *The Hidden Pattern*.
[47] Legg and Hutter, "A Collection of Definitions of Intelligence."
[48] Collins et al., "Artificial Intelligence in Information Systems Research," 10.
[49] Collins et al., "Artificial Intelligence in Information Systems Research," 6; cf. Russell and Norvig, *Artificial Intelligence: A Modern Approach*, 1–5.
[50] Rai, Constantinides, and Sarker, "Next-Generation Digital Platforms."

McCarthy). Some avoid reference to human intelligence and instead simply refer to behaviour commonly associated with intelligence (e.g., Rai et al.). We could speculate that for some, this implies an *equivocal* use of the terms, whereby AI is something quite different from human intelligence. Of course, such categorization is neither exhaustive nor does it allow to draw neat boundaries.[51]

Where does this short survey of divergent views leave us? How are we to make sense of and navigate the difficult and confusing conversation about the meaning of AI? Let us first note that the practical successes of DL and LLMs has certainly contributed to a diverging of opinions in that the conversation has since been driven to a substantial degree by marketing. Nevertheless, opposing interpretations have been present long before AI became practically interesting for business purposes, indicating that the division lies deeper still.[52]

Meaning, Narrative, and the Future of AI

No one would call a pocket calculator 'intelligent'. When confronted with the question 'is AI genuinely intelligent?' we must therefore presume that 'intelligence' does not reduce to blind calculus but involves some form of context, of what calculations are *about*. However, wherein such *aboutness* or *intentionality* lies is a question that has been troubling philosophers of various traditions for centuries. Brentano, who revived the notion of 'intentionality' in the 19th century, argued that it exclusively marks the mental realm and "[n]o physical phenomenon exhibits anything like it."[53] Quine has later formulated two possible reactions to Brentano's thesis of the irreducibility of intentionality to physics: One may take it to either show the limits of physics and the need for a separate "science of intention" or to reveal the emptiness and "baselessness of intentional idioms" (Quine himself opted for the latter).[54] Philosophers have since reacted in many more nuanced ways to bridge the intentional (mental) and physical realms, for example, by positing 'physical symbols' and 'mental representations' (see above). However, the basic dichotomy persists and there is no satisfactory and potentially broadly acceptable theory in sight.[55] The depth of the problem becomes more obvious when we consider that it is not a problem about something, but about the very *aboutness* with which a problem can be about something

[51] A fruitful application of the rich theological notions of analogy, univocity, and equivocity to talk of AI can be found in Davison, "Machine Learning and Theological Traditions of Analogy."
[52] For example, already in 1965 Hubert Dreyfus argued that key features of human intelligence cannot be captured by AI. See Dreyfus, *Alchemy and Artificial Intelligence*.
[53] Brentano, *Psychology from an Empirical Standpoint*, 68.
[54] Quine, *Word and Object*, 221.
[55] According to Schlicht and Smortchkova, today's debates about the concept of 'mental representation' represents a culmination of the two mainstream philosophical traditions mostly viewed in opposition to each other, namely the analytic and the continental traditions. See Schlicht and Smortchkova, *Mentale Repräsentationen*, 28.

in the first place. Thus, there is no neutral 'unproblematic' ground from which we could begin to derive a common solution. As a consequence, whether we think of intentionality as altogether different from physics, reducible to physics (e.g., with 'physical symbols') or illusory seems to have more to do with the presumptions we begin with than with the conclusions we arrive at. Accordingly, how we define 'intelligence' and whether we think of AI as 'genuinely intelligent' derives much from the (often implicit) metaphysical presumptions and narratives we inhabit.[56] Although this is not explicitly discussed in core AI literature, it sometimes surfaces. For example Neil Lawrence, the DeepMind Professor of Machine Learning at the University of Cambridge, notes that "[a]lthough not a believer, I have sympathy with the idea that there is a spirit within us that cannot be replaced by a machine."[57]

Of course, how we think about AI in the present also largely shapes what we expect from the future. Although experts from industry and academia agree that it is hard to overestimate the future impact of AI on all aspects of society, including the economy and personal life, they do not agree on how these changes will look: The vast majority of narratives shared in public is dystopian (massive job loss, the emergence of a useless class, possible extinction, etc.),[58] while the likely majority of experts takes a less extreme, evidence-based stance (adoption will take time, the current technology is still very limited with no clear path to a next jump in capabilities).[59] This indicates that the basis on which such predictions on the future of AI are made consists not of generalisable evidence but mainly on philosophical presumptions and attitudes toward the world, human beings, and technology.[60] Here, we take a short and oversimplified look at two particular and prominent opposite positions that could be taken to loosely correspond with presuming the priority of the physical and the intentional (mental) realms, respectively.

---

[56] See Matter, *Verleiblichte Geschichten*.
[57] Lawrence, *The Atomic Human; Understanding Ourselves in the Age of AI,* 13.
[58] See, e.g., Future of Life Institute, "Pause Giant AI Experiments: An Open Letter,"; Yudkowsky, "Will Superintelligent AI End the World?"; "AI 2027," https://ai-2027.com/; Citrini Research, "2028 GIC," https://www.citriniresearch.com/p/2028gic; and Harari, *Homo Deus*.
[59] See, e.g. Dash, "The Majority AI View"; Stadelmann, "Debate: Evidence-Based AI Risk Assessment for Public Policy"; Narayanan and Kapoor, "AI as Normal Technology"; Brooks, "The Seven Deadly Sins of AI Predictions"; Marcus, "Deep Learning: A Critical Appraisal"; von der Malsburg, Stadelmann, and Grewe, "A Theory of Natural Intelligence"; Kambhampati, "Can Large Language Models Reason and Plan?"; Kambhampati et al., "Stop Anthropomorphizing Intermediate Tokens as Reasoning/Thinking Traces!"; Kumar et al., "Questioning Representational Optimism in Deep Learning"; and Silver and Sutton, "Welcome to the Era of Experience."
[60] Most scenarios of course make some claims about potential steps of technological progress. What we mean here: These steps may appear likely based on an underlying worldview; they are however not explainable based on a foreseeable progression of technological inventions *per se* but presuppose innovation leaps like the sudden appearance of 'artificial general intelligence' (AGI), for which researchers currently know no viable approach.

First, we look into what we call the *futurist* position (which can be loosely associated with the reduction of intentionality to physics).[61] Gebru and Torres analysed the respective worldviews and coined the acronym *TESCREAL* to refer to the combination of Transhumanism, Extropianism, Singularitarism, Cosmism, Rationalism, Effective Altruism, and Longtermism.[62] They show how respective philosophies are widespread in the tech industry, and how they profoundly shape the global narratives on AI. The TESCREAL narrative for AI goes roughly like this: Intentional human beings are nothing but information processors – just on biological, decaying hardware (according to Rationalism)[63]. This makes them akin to machines, just inferior, because of the fragile hardware. If humans could gain intelligence and intentionality, then nothing prevents machines from reaching the same soon, called 'artificial general intelligence' (AGI). AGI will progress further to superintelligence (according to Singularitarism), so humans should upgrade themselves to become more like machines (Transhumanism). In TESCREALism, the future belongs to AI, and humans become a marginal note.[64] Even if these views are not explicitly held in their comprehensive form, belief in a sudden, inevitable AGI makes most futurists proponents of negative future outlooks.[65]

Second, we consider the *normalist* position holding to AI as 'normal technology' (which can loosely be associated with the irreducibility of intentionality, although without ultimately committing to it).[66] The normalist position focuses on the current state of the technology which – based on established and more robust patterns of technological adoption – it considers as both useful and astonishing, yet fundamentally limited.[67] It

---

[61] Futurists can be either *Doomers* (seeing a likely path to human extinction) or *Accelerationists* (supporting speedy AI progress even if there could be temporary bad effects like massive job loss). Both fundamentally believe in an imminent AGI (or: superintelligence) based on recursive self-improvement of the current technology.

[62] The acronym 'TESCREAL' is formed from the first letters of the listed ideological schools of thought. See Gebru and Torres, "The TESCREAL Bundle.", and the glossary of terms below. It is sometimes perceived as containing traits of an ideological rally call.

[63] Note that 'rationalism' as understood within TESCREAL deviates in important ways from its understanding in the history of philosophy.

[64] According to Lawrence, this is due to flawed thinking: "[W]e have to be careful to separate the notion of intelligence as a property from that of an intelligence as an entity. […] Both Bostrom's definition of superintelligence and the notion of the technological singularity are flawed. They misrepresent intelligence as a unidimensional quality and this doesn't reflect the diversity of intelligences we experience". See Lawrence, "The Atomic Human", 25-26.

[65] Respective views are comprehensively discussed in community forums like https://www.lesswrong.com/. One argument for negative outlooks besides existential threats from misaligned superintelligences is that individual economic incentives will create a race to the bottom, leading to disruption and collapse. This concept is discussed under the term 'Moloch', based on Alexander, "Meditations on Moloch."

[66] Narayanan and Kapoor, "AI as Normal Technology."

[67] Consider the recent case of 'Moltbook', a social network or exchange forum for prototypical autonomous AI agents in which these LLM-based bots exchange uncanny texts on a 'rise of the bots' etc. Futurists, e.g., Elon Musk on X, commented on signs of the advent of the Singularity. Normalists, however, note that (a) the underlying LLMs were trained on virtually all text on the internet, including vast amounts of sci-fi pulp fiction (in which AI takes over the world, gains consciousness, etc.) and (b) are on Moltbook, now sampling each other within a context that (via pre-prompts, etc.) ascribes 'autonomy' to

does not necessarily deny the eventual advent of AGI-like technology sometime in the future, but it doesn't see it as very sudden or disruptive.[68] Instead, the progress to AGI must check numerous safety interventions and will be slowed by the necessary steps for adoption in complex real-world scenarios, where human beings stay in charge. And even if highly autonomous systems eventually lead to enormous power concentrations, it is not that 'humanity' loses control over 'AI'. Instead, the power concentrates in those humans that control AI over those humans that do not. [69] Here, it is almost exclusively human beings that appear as intentional actors who are *about* things, using AI as a tool at their hands.[70]

Having shortly glanced at two ways of interpreting AI that prioritize the physical and the intentional (mental) realms respectively, let us briefly point to some developments in cognitive science and the philosophy of technology that look promising in taking us beyond Brentano's dichotomy. Since the 1990s, *enactivism* has been shifting the focus away from the dichotomy between physical objects on the one side and intentional subjects on the other, to the co-constitution of organism and environment. This implies, for example, that if 'apple' is uttered by an LLM lacking any kind of embodied interaction with an environment, such as in metabolism, this uttering should be conceptualized not within a human context of embodiment, but within a context of statistical next-word prediction. In this new light, although the intentional and the physical realms are co-dependent, human beings and AI are very different from each other.[71] Furthermore, studies in *postphenomenology* suggest that human beings incorporate the use of technologies, whereby technologies enter the co-constitution of organism and environment. More precisely, technologies play a *mediating* role and should thus neither be conceptualized from an instrumentalist view as neutral tools, nor from a substantivist view as controlling agents themselves. Rather, "[t]echnologies-in-use help shape the relation between the users and their environment."[72] In this light, although AI does not have agency in the way living human beings do, they nevertheless have a kind of agency as they transform our social, cultural, and experiential environments quite drastically, shaping and 'nudging' us in return.[73]

---

them. They remark that in this context, what can be read is to be expected, not as a sign of machine consciousness, but fully explained by the statistical plausibility of language in context.

[68] Where it is considered, it doesn't change the respective scenarios.

[69] E.g., "power and loss of control are essentially tautological". Narayanan and Kapoor, "AI as Normal Technology."

[70] Viewing technology as 'neutral' tools at the hands of human beings has been called the 'Value Neutrality Thesis'. See, e.g., Pitt, "'Guns Don't Kill, People Kill'."

[71] For further reading, we suggest: Varela, Thompson, and Rosch, *The Embodied Mind*; Thompson, *Mind in Life*; and Fuchs, *Ecology of the Brain*. See also Segessenmann et al., "Assessing Deep Learning."

[72] Verbeek, "Beyond Interaction," 263.

[73] For further reading, we suggest: Ihde, *Postphenomenology and Technoscience*; Verbeek, *What Things Do*. Similar conclusions are drawn in the field of Science and Technology Studies; see Sismondo, *An Introduction to Science and Technology Studies*; Felt et al., *The Handbook of Science and Technology Studies*.

Conclusion: What is AI?

All this being said, we are still to ask, what is AI? Following the above developments, what we mean by 'intelligence' can only ever be drawn from our embodied experience as intelligent organisms, that is, as human beings. Only in a later step do we apply it to other animals or things by way of analogy.[74] Furthermore, mounting evidence for the vast differences between living, experiencing organisms and technologies such as AI indicates that AI is not the simple *reproduction* of human intelligence. However, AI should also not be understood within a naïve instrumental conception of human–technology relations. Instead, AI shapes our view of ourselves and the world by way of mediation and (only) in this sense, has its own form of agency.[75] In this light, we propose to follow the Dartmouth conference and Rai et al. and define AI as: *the simulation of behaviour usually associated with human intelligence*. Admittedly, this definition is rather vague and leaves many questions unanswered. What do we commonly mean by human intelligence? How is this meaning influenced in turn by developments in AI? (Note the circularity.) However, the current developments in AI research and cognitive science do not substantiate any more precise or stronger claims.

How to carry on the conversation? We have already emphasized that opposing interpretations of AI often implicate conflicting metaphysical presumptions (that are seldom explicitly discussed). This further implicates that a 'neutral' assessment and comparison of opposing interpretations is inherently difficult, since this would involve the assessment and comparison of the very background on which we assess and compare things. After all, the presumptions that enable us to inhabit a meaningful world and act sensibly in it, are to some degree incommensurable. However, this is not to say that they are fixed or cannot be made transparent to some extent. To the contrary, we think that bringing the conversations about what AI is forward desperately requires us to attend to them.[76] The precedence that AI sets forces us to think about

---

[74] Thus, in a sense, anthropomorphism is inevitable (not to confuse with anthropocentrism); see Spaemann, "Wirklichkeit als Anthropomorphismus." However, this does not mean that we should not be cautious of speaking of AI in human terms; see, e.g., Bleckmann and Segessenmann, "'I Grasped That the Computer Calculates Everything and Does Not Think'"; Dürr, Segessenmann, and Steinmann, "Meaning, Form and the Limits of Natural Language Processing."

[75] In the technical literature on AI, it is typically understood that terms borrowed from the human sphere like 'agency' (but also 'learning', 'reasoning' etc.) are first and foremost used by analogy and do not necessarily constitute identity (because, in a very real sense, better terms are missing and so these got established as technical terms with an analogous meaning). See Stadelmann, Merkt, and Barr, "The Stochastic Nature of Machine Learning and Its Implications for High-Consequence AI," for a continued discussion and references.

[76] This is also necessary to develop a greater diversity of possible scenarios for a societal future with AI and qualify them with any degree of likelihood: for this, the underlying assumptions need to be made clear. See for example Stadelmann, "AI in 2035 - A hope-filled vision for a humane future with AI". If respective scenarios are positive in the sense of contributing to human flourishing also depends on the

what we are as human beings, and how we relate to technology and culture. We have already hinted at enactivism and postphenomenology as two recent developments that point the discussion in a promising direction. However, these are age-old questions and in the last analysis, what we regard as really 'real' is ultimately an ethical and religious question. Thus, hopefully, AI leads us also into a conversation with religious traditions that have been centring around such questions for millennia.

Glossary

AI	Artificial Intelligence. The simulation of behaviour usually associated with human intelligence, on a computer.
LLM	Large Language Model. An AI model that maps the probability of the next word given the context of words in the input and thus has learned to "deal with language".
ML	Machine Learning. AI method to establish the relationship between arbitrary input and output by systematically estimating the underlying function's parameters from given training data. (Note that the resulting AI model will be a probabilistic model, that approximates the true relationship only with a certain probability.)
AGI	Artificial General Intelligence. An AI system able to perform all commercially relevant cognitive work at or above the human level of performance. (Note that since the development of AI shifts what we associate with 'human intelligence' the meaning of AGI will always shift with it.)
ANN	Artificial Neural Network. An AI model used for approximating arbitrary functions by combining many very simple units, roughly inspired by neurophysiology.
TESCREAL	Transhumansism, Extropianism, Singularitarism, Cosmism, Rationalism, Effective Altriusm, and Longtermism. The acronym is meant to refer to (warn about) the mash-up of respective philosophical ideas regarding future technological developments, with the implication to emphasize the wellbeing of hypothetical myriads of future machine-beings in space over actual suffering of people today on earth.

technological design of mainstream AI systems. See Stadelmann et al., "Pro-Human AI Design: Concept, Methodology, and Pre-liminary Results".

Bibliography

Alexander, Scott. “Meditations on Moloch.” *Slate Star Codex*. 30 July 2014. https://slatestarcodex.com/2014/07/30/meditations-on-moloch/

Aristotle. *The Physics*. Vol 1. Translated by Philip H. Wicksteed and Francis M. Cornford. Loeb Classical Library. Cambridge, MA: Harvard University Press, 1929 (revised and reprinted 1957).

Asimov, Isaac. *I, Robot*. New York: Gnome Press, 1950.

Balestriero, Randall, and Yann LeCun. “LeJEPA: Provable and Scalable Self-Supervised Learning without the Heuristics.” arXiv preprint arXiv:2511.08544 (2025).

Bengio, Yoshua, Pascal Lamblin, Dan Popovici, and Hugo Larochelle. “Greedy Layer-Wise Training of Deep Networks.” In *Advances in Neural Information Processing Systems*, vol. 19. Cambridge, MA: MIT Press, 2006. https://proceedings.neurips.cc/paper_files/paper/2006/file/5da713a690c067105aeb2fae32403405-Paper.pdf

Bhuyan, Bikram Pratim, Amar Ramdane-Cherif, Ravi Tomar, and T. P. Singh. “Neuro-Symbolic Artificial Intelligence: A Survey.” *Neural Computing and Applications* 36, no. 21 (2024): 12809–12844.

Bleckmann, Paula, and Jan Segessenmann. “‘I Grasped That the Computer Calculates Everything and Does Not Think’: ICT Education Can Challenge or Cement Computer-Anthropologies.” *Philosophy, Theology and the Sciences* 11, no. 2 (2024): 227–251. https://doi.org/10.1628/ptsc-2024-0017

Boden, Margaret. *Mind as Machine: A History of Cognitive Science*. Oxford: Oxford University Press, 2006.

Brentano, Franz. *Psychology from an Empirical Standpoint*. Edited by Oskar Kraus and Linda L. McAlister. New York: Routledge, 1995 [1874].

Brooks, Rodney. “The Seven Deadly Sins of AI Predictions.” *MIT Technology Review*. 6 October 2017. https://www.technologyreview.com/2017/10/06/241837/the-seven-deadly-sins-of-ai-predictions/

Brown, Tom et al. “Language Models Are Few-Shot Learners.” *Advances in Neural Information Processing Systems* 33 (2020): 1877–1901.

Clark, Andy. “Predicting Peace: The End of the Representation Wars — A Reply to Michael Madary.” In *Open MIND*. Edited by Thomas Metzinger and Jennifer M. Windt. Frankfurt am Main: MIND Group, 2015. https://doi.org/10.15502/9783958570979

Collins, Christopher, Denis Dennehy, Kieran Conboy, and Patrick Mikalef. “Artificial Intelligence in Information Systems Research: A Systematic Literature Review and Research Agenda.” *International Journal of Information Management* 60 (2021): 102383.

Cottier, Ben et al. “The Rising Costs of Training Frontier AI Models.” arXiv preprint arXiv:2405.21015 (2024).

Covington, Paul, Jay Adams, and Emre Sargin. “Deep Neural Networks for YouTube Recommendations.” In *Proceedings of the 10th ACM Conference on Recommender Systems*. 191–198. New York: ACM, 2016.

Dao, Minh-Dung et al. “Agentic Design Patterns: A System-Theoretic Framework.” arXiv preprint arXiv:2601.19752 (2026).

Dash, Anil. “The Majority AI View.” 17 October 2025. https://www.anildash.com/2025/10/17/the-majority-ai-view/

Davison, Andrew. “Machine Learning and Theological Traditions of Analogy.” *Modern Theology* 37, no. 2 (2021): 254–274. https://doi.org/10.1111/moth.12682

Descartes, René. *Treatise of Man*. Translated by Thomas S. Hall. Cambridge, MA: Harvard University Press, 1972.

Dürr, Oliver, Jan Segessenmann, and Jan Juhani Steinmann. “Meaning, Form and the Limits of Natural Language Processing.” *Philosophy, Theology and the Sciences* 10, no. 1 (2023): 42–72. https://doi.org/10.1628/ptsc-2023-0005

Dreyfus, Hubert L. *Alchemy and Artificial Intelligence*. Santa Monica, CA: RAND Corporation, 1965. https://www.rand.org/pubs/papers/P3244.html

Eisenstein, Michael. “Artificial Intelligence Powers Protein-Folding Predictions.” *Nature* 599, no. 7886 (2021): 706–708. https://doi.org/10.1038/d41586-021-03499-y

Felt, Ulrike et al. *The Handbook of Science and Technology Studies*. 4th ed. Cambridge, MA: MIT Press, 2017.

Fradkov, Alexander L., and Alexander I. Shepeljavyi. “The History of Cybernetics and Artificial Intelligence: A View from Saint Petersburg.” *Cybernetics and Physics* 11, no. 4 (2022): 253–263.

Fuchs, Thomas. *Ecology of the Brain: The Phenomenology and Biology of the Embodied Mind*. Oxford: Oxford University Press, 2018.

Future of Life Institute. “Pause Giant AI Experiments: An Open Letter.” 22 March 2023. https://futureoflife.org/open-letter/pause-giant-ai-experiments/

Gebru, Timnit, and Émile P. Torres. “The TESCREAL Bundle: Eugenics and the Promise of Utopia through Artificial General Intelligence.” *First Monday* 29, no. 4 (2024). https://doi.org/10.5210/fm.v29i4.13636

Goertzel, Ben. *The Hidden Pattern*. Boca Raton, FL: Brown Walker Press, 2006.

Google DeepMind. “Genie.” Accessed 10 March 2026. https://deepmind.google/models/genie/.

Ha, David, and Jürgen Schmidhuber. “Recurrent World Models Facilitate Policy Evolution.” *Advances in Neural Information Processing Systems* 31 (2018): 2451–2463.

Harari, Yuval Noah. *Homo Deus: A Brief History of Tomorrow*. London: Harvill Secker, 2016.

Haugeland, John. *Artificial Intelligence: The Very Idea*. Cambridge, MA: MIT Press, 1985.

Hebb, Donald O. *The Organization of Behavior*. New York: Wiley and Sons, 1949.

Hestness, Joel et al. “Deep Learning Scaling Is Predictable, Empirically.” arXiv preprint arXiv:1712.00409 (2017).

Hinton, Geoffrey E., Simon Osindero, and Yee-Whye Teh. “A Fast Learning Algorithm for Deep Belief Nets.” *Neural Computation* 18, no. 7 (2006): 1527–1554. https://doi.org/10.1162/neco.2006.18.7.1527

Ihde, Don. *Postphenomenology and Technoscience: The Peking University Lectures*. Albany, NY: State University of New York Press, 2009.

Jermain, Peter R. et al. “Deep Learning-Based Cell Segmentation for Rapid Optical Cytopathology of Thyroid Cancer.” *Scientific Reports* 14, no. 1 (2024): 16389.

Jordan, Michael I., and Thomas M. Mitchell. “Machine Learning: Trends, Perspectives, and Prospects.” *Science* 349, no. 6245 (2015): 255–260.

Jumper, John et al. “Highly Accurate Protein Structure Prediction with AlphaFold.” *Nature* 596, no. 7873 (2021): 583–589. https://doi.org/10.1038/s41586-021-03819-2

Kambhampati, Subbarao. “Can Large Language Models Reason and Plan?” *Annals of the New York Academy of Sciences* 1534, no. 1 (2024): 15–18.

Kambhampati, Subbarao et al. “Stop Anthropomorphizing Intermediate Tokens as Reasoning/Thinking Traces!” arXiv preprint arXiv:2504.09762 (2025).

Kumar, Akarsh, Jeff Clune, Joel Lehman, and Kenneth O. Stanley. “Questioning Representational Optimism in Deep Learning: The Fractured Entangled Representation Hypothesis.” arXiv preprint arXiv:2505.11581 (2025).

Lanning, Daniel R., Gregory K. Harrell, and Jin Wang. “Dijkstra's Algorithm and Google Maps.” In *Proceedings of the 2014 ACM Southeast Conference*. 1–3. New York: ACM, 2014.

Lawrence, Neil D. *The Atomic Human; Understanding Ourselves in the Age of AI*. Allen Lane, 2024.

Legg, Shane, and Marcus Hutter. “A Collection of Definitions of Intelligence.” *Frontiers in Artificial Intelligence and Applications* 157 (2007): 17–24.

Li, Yingbo et al. “Physical Artificial Intelligence (PAI): The Next-Generation Artificial Intelligence.” *Frontiers of Information Technology & Electronic Engineering* 24, no. 8 (2023): 1231–1238.

Liu, Yang et al. “Aligning Cyber Space with Physical World: A Comprehensive Survey on Embodied AI.” *IEEE/ASME Transactions on Mechatronics* 30 (2025): 7253–7274. https://doi.org/10.1109/TMECH.2025.3574943

MacIntyre, Alasdair. *After Virtue: A Study in Moral Theory*. 3rd ed. Notre Dame, IN: University of Notre Dame Press, 2007 [1981].

Marcus, Gary. “Deep Learning: A Critical Appraisal.” arXiv preprint arXiv:1801.00631 (2018).

Matter, Nicolas. *Verleiblichte Geschichten: Studien zur Schnittstelle von Leiblichkeit und Narrativität als Prolegomena zu einer ekklesialen Pädagogik*. (=Jerusalemer Theologisches Forum 48). Münster: Aschendorff, 2025.

McCarthy, John, Marvin L. Minsky, Nathaniel Rochester, and Claude E. Shannon. “A Proposal for the Dartmouth Summer Research Project on Artificial Intelligence, August 31, 1955.” *AI Magazine* 27, no. 4 (2006): 12–14.

McCarthy, John. “What Is Artificial Intelligence?” 2004. http://www-formal.stanford.edu/jmc/whatisai/whatisai.html

McCulloch, Warren, and Walter Pitts. “A Logical Calculus of the Ideas Immanent in Nervous Activity.” *Bulletin of Mathematical Biophysics* 5 (1943): 115–133.

Meyer, Benjamin et al. “A Document Is Worth a Structured Record: Principled Inductive Bias Design for Document Recognition.” arXiv preprint arXiv:2507.08458 (2025).

Minsky, Marvin. *The Society of Mind*. New York: Simon and Schuster, 1986.

Narayanan, Arvind, and Sayash Kapoor. “AI as Normal Technology.” *Knight First Amendment Institute* 25-09 (2025). https://perma.cc/HVN8-QGQY

Newell, Allen, and Herbert Simon. “The Logic Theory Machine: A Complex Information Processing System.” *IRE Transactions on Information Theory* 2, no. 3 (1956): 61–79.

Newell, Allen, and Herbert A. Simon. “Computer Science as Empirical Inquiry: Symbols and Search.” *Communications of the ACM* 19, no. 3 (1976): 113–126.

Nilsson, Nils J. *The Quest for Artificial Intelligence*. Cambridge: Cambridge University Press, 2009.

Ouyang, Long et al. “Training Language Models to Follow Instructions with Human Feedback.” *Advances in Neural Information Processing Systems* 35 (2022): 27730–27744.

Pitt, Joseph C. “‘Guns Don't Kill, People Kill’: Values in and/or Around Technologies.” In *The Moral Status of Technical Artefacts*. Edited by Peter Kroes and Peter-Paul Verbeek. Philosophy of Engineering and Technology 17. 89–101. Dordrecht: Springer, 2014.

Quine, Willard V. W. *Word and Object*. Cambridge, MA: MIT Press, 1960.

Rai, Arun, Panos Constantinides, and Saonee Sarker. “Next-Generation Digital Platforms: Towards Human-AI Hybrids.” *MIS Quarterly* 43 (2019): iii-ix.

Ranzato, Marc'Aurelio, Christopher Poultney, Sumit Chopra, and Yann LeCun. “Efficient Learning of Sparse Representations with an Energy-Based Model.” In *Advances in Neural Information Processing Systems*, vol. 19. Cambridge, MA: MIT Press, 2006. https://proceedings.neurips.cc/paper_files/paper/2006/file/87f4d79e36d68c3031ccf6c55e9bbd39-Paper.pdf

Rashid, Adib Bin, and MD Ashfakul Karim Kausik. “AI Revolutionizing Industries Worldwide: A Comprehensive Overview of Its Diverse Applications.” *Hybrid Advances* 7 (2024): 100277.

Ricoeur, Paul. *Time and Narrative*. 3 vols. Translated by Kathleen McLaughlin and David Pellauer. Chicago: University of Chicago Press, 1984–1988.

Rosenblatt, Frank. *Principles of Neurodynamics: Perceptrons and the Theory of Brain Mechanisms*. Washington DC: Spartan Books, 1962.

Russell, Stuart, and Peter Norvig. *Artificial Intelligence: A Modern Approach*. 4th ed. Hoboken, NJ: Prentice Hall, 2020.

Sager, Patrick J. et al. “A Comprehensive Survey of Agents for Computer Use: Foundations, Challenges, and Future Directions.” *Journal of Artificial Intelligence Research* (accepted for publication, 2026).

Schlicht, Tobias, and Joulia Smortchkova. “Einleitung.” In *Mentale Repräsentationen*. Edited by Tobias Schlicht and Joulia Smortchkova. Berlin: Suhrkamp, 2018. 7–42.

Schmidhuber, Jürgen. “Deep Learning in Neural Networks: An Overview.” *Neural Networks* 61 (2015): 85–117.

Schmidhuber, Jürgen. “Annotated History of Modern AI and Deep Learning.” arXiv preprint arXiv:2212.11279 (2022).

Segessenmann, Jan, Thilo Stadelmann, Andrew Davison, and Oliver Dürr. “Assessing Deep Learning: A Work Program for the Humanities in the Age of Artificial Intelligence.” *AI and Ethics* 5 (2025): 169–200. https://doi.org/10.1007/s43681-023-00408-z

Silver, David, and Richard S. Sutton. “Welcome to the Era of Experience.” In *Designing an Intelligence*. Edited by George Konidaris. Cambridge, MA: MIT Press (forthcoming).

Sismondo, Sergio. *An Introduction to Science and Technology Studies*. Oxford: Wiley-Blackwell, 2010.

Spaemann, Robert. “Wirklichkeit als Anthropomorphismus.” In *Was heisst ‘wirklich’? Unsere Erkenntnis zwischen Wahrnehmung und Wissenschaft*. Waakirchen-Schaftlach, 2000. 13–34.

Stadelmann, Thilo. “A Guide to AI: Understanding the Technology, Applying It Successfully, and Shaping a Positive Future.” *Global Resilience White Papers*, no. 2. 28 January 2025. https://stdm.github.io/downloads/papers/GRW_2025.pdf.

Stadelmann, Thilo. “AI in 2035 - A hope-filled vision for a humane future with AI.” *AIssays blog*, October 2025. https://stdm.github.io/AI-in-2035/.

Stadelmann, Thilo. “Debate: Evidence-Based AI Risk Assessment for Public Policy.” *Public Money & Management* 46, no. 1 (2026): 5–7.

Stadelmann, Thilo, Christoph Heitz, Rebekka von Wartburg-Kottler, and Andrea Luca Schärer. “Pro-Human AI Design: Concept, Methodology, and Pre-liminary Results”. *Proceedings of uDay XXIV Workshop on “#responsible AI: Europas Weg zum Erfolg?”*, FHV Vorarlberg University of Applied Sciences, May 21, 2026.

Stadelmann, Thilo, Philipp H. Merkt, and Kasey Barr. “The Stochastic Nature of Machine Learning and Its Implications for High-Consequence AI.” *AI and Ethics* 6, 195, Springer, March 15, 2026.

Stadelmann, Thilo, Tino Klamt, and Philipp H. Merkt. “Data Centrism and the Core of Data Science as a Scientific Discipline.” *Archives of Data Science, Series A* 8, no. 2 (2022): 1–16.

Sutton, Richard. “The Bitter Lesson.” *Incomplete Ideas* (blog). 2019. http://www.incompleteideas.net/IncIdeas/BitterLesson.html

Taylor, Charles. *Sources of the Self: The Making of the Modern Identity*. Cambridge, MA: Harvard University Press, 1989.

Taylor, Charles. *The Language Animal: The Full Shape of the Human Linguistic Capacity*. Cambridge, MA: Belknap Press of Harvard University Press, 2016.

Thompson, Evan. *Mind in Life: Biology, Phenomenology, and the Sciences of Mind*. Cambridge, MA: Harvard University Press, 2007.

Turing, Alan M. “On Computable Numbers, with an Application to the Entscheidungsproblem.” *Proceedings of the London Mathematical Society*, ser. 2, 42 (1936–1937): 230–265.

Turing, Alan M. "Computing Machinery and Intelligence." *Mind* 59, no. 236 (1950): 433–460.

Varela, Francisco J., Evan Thompson, and Eleanor Rosch. *The Embodied Mind: Cognitive Science and Human Experience*. Cambridge, MA: MIT Press, 1991.

Verbeek, Peter-Paul. “Beyond Interaction: A Short Introduction to Mediation Theory.” *Interactions* 22, no. 3 (2015): 26–31. https://doi.org/10.1145/2751314

Verbeek, Peter-Paul. *What Things Do: Philosophical Reflections on Technology, Agency, and Design*. Translated by Robert P. Crease. University Park, PA: Penn State University Press, 2005.

von der Malsburg, Christoph, Thilo Stadelmann, and Benjamin F. Grewe. “A Theory of Natural Intelligence.” arXiv preprint arXiv:2205.00002 (2022).

Von Neumann, John. “First Draft of a Report on the EDVAC.” *IEEE Annals of the History of Computing* 15, no. 4 (1993 [1945]): 27–75.

Wei, Jason et al. “Emergent Abilities of Large Language Models.” *Transactions on Machine Learning Research* (2022). https://doi.org/10.48550/arXiv.2206.07682

Wei, Jason et al. “Chain-of-Thought Prompting Elicits Reasoning in Large Language Models.” *Advances in Neural Information Processing Systems* 35 (2022): 24824–24837.

Wiener, Norbert. *Cybernetics: Or Control and Communication in the Animal and the Machine*. New York: Wiley, 1948.

Yan, Peng et al. “Learning Actionable World Models for Industrial Process Control.” In *Proceedings of the 12th IEEE Swiss Conference on Data Science (SDS'25)*. 111–118. Zurich, 2025. https://doi.org/10.1109/SDS66131.2025.00022.

Yudkowsky, Eliezer. “Will Superintelligent AI End the World?” TED Talk. 18 April 2023. https://www.ted.com/talks/eliezer_yudkowsky_will_superintelligent_ai_end_the_world

Ziegler, Daniel M. et al. “Fine-Tuning Language Models from Human Preferences.” arXiv preprint arXiv:1909.08593 (2019).

Jan Segessenmann is currently pursuing a PhD at the URPP Digital Religion(s) of the University of Zurich studying anthropological assumptions in AI research and cognitive science. He holds a BSc in Microengineering from the Bern University of Applied Sciences and an MSc in Biomedical Engineering from the University of Bern. While conducting research in computational neuroscience and data science in Bern, he concurrently pursued theological studies at the University of Bern and the University of Fribourg. Since 2023, his research has focused on the intersection of theological anthropology, the philosophy of artificial intelligence, and cognitive science. [jan.segessenmann@uzh.ch, https://orcid.org/0000-0001-7754-2578]

Thilo Stadelmann is professor of AI and machine learning at the ZHAW School of Engineering in Winterthur, Switzerland, and founding director of the ZHAW Centre for Artificial Intelligence. He studied computer science in Giessen and Marburg and received his Doctor of Science degree from Marburg University, Germany, in 2010, where he worked on multimedia analysis. He held engineering and leadership roles in the automotive industry before his appointment at the ZHAW. His research group focuses on robust representation learning for pattern recognition applications. His current research interests include the societal implications of AI, leading to pro-human AI design approaches. [stdm@zhaw.ch, https://orcid.org/0000-0002-3784-0420]