The stochastic nature of machine learning and its implications for high-consequence AI

Thilo Stadelmann[*]
Centre for Artificial Intelligence
Zurich University of Applied Sciences
8400 Winterthur, ZH, Switzerland
stdm@zhaw.ch

Abstract

Modern AI systems achieve remarkable performance through fundamentally stochastic processes—machine learning models that function as high-dimensional probability density functions, outputting the most likely predictions given training data. While these systems can match or exceed human performance on average, their methodology produces fundamentally different failure modes than human reasoning, leading to errors that appear nonsensical from a human perspective but are predictable given their probabilistic nature. This has critical implications for high-consequence environments such as military applications where decisions cannot be reversed and may affect lives and material assets definitively. Through detailed analysis of contemporary AI's working mechanisms—particularly how knowledge is acquired through statistical pattern recognition rather than causal reasoning—this paper demonstrates why AI systems inherit biases, cannot distinguish plausibility from factual correctness, and exhibit confident behaviour even when wrong. For effective deployment of AI in high-consequence scenarios, processes need to be implemented that make sure all human stakeholders are aware of these facts, develop adequate scepticism of the AI system, and remain actively involved in the decision-making. For military applications specifically, this understanding reveals that effective human-AI collaboration requires more than oversight: It demands colearning frameworks that maintain meaningful human agency through bidirectional information flow. We give an outlook to a decentralized AI system tailored to specific teams to mitigate power concentration risks while preserving essential human capacities, including moral judgment to exercise mercy.

Key words: artificial intelligence, statistics, error patterns, military decision-making, human-AI-teaming, colearning, battlefield AI

1 Introduction

Modern AI has earned a reputation of yielding results comparable to human level performance for a wide array of tasks [Stadelmann et al., 2019, Stadelmann, 2025a], e.g., visual recognition [Žigulić et al., 2024], text and video comprehension [Tang et al., 2025], decision support based on heterogeneous data analysis [Huang et al., 2025], and first steps towards autonomous multi-step acting [Sager et al., 2025b]. Indeed, for many benchmarks, AI results even surpass human performance, in line with many anecdotal examples [Bubeck et al., 2023]. At the same time, similarly real experiences exhibit uncanny "stupid errors" of AI systems that do not exhibit common sense, making one question

[*] Fellow of the ECLT European Centre for Living Technology, 30123 Venice, Italy; member of the Scientific Council of the IAEAI Israeli Association for Ethics in Artificial Intelligence, Tel Aviv, Israel.

Preprint. Under review.

bold claims of “understanding”, “reasoning”, or, generally, “fitness for purpose” of any practical sort of these models [Brooks, 2017, Marcus, 2018, von der Malsburg et al., 2022, Neururer et al., 2024, Kambhampati, 2024, Kambhampati et al., 2025, Narayanan and Kapoor, 2025, Kumar et al., 2025, Silver and Sutton, 2025].

This has important ramifications in high-consequence environments such as certain military applications where decisions cannot be taken back and may affect lives and material assets in a definitive way: how to deal with such variance in order fulfillment? After all, AI has been suggested (and, in current conflicts that usually speed up innovation and adoption: is used) as an important component in aspects ranging from the military decision-making process (MDMP) to lethal autonomous weapon systems (LAWS). For example, Meerveld et al. [2023] express the hope that the use of AI could help in every step of the MDMP with automation and support that mitigates human decision-making biases, overcomes human inadequacy to extract knowledge from high volumes of data, and leads to higher efficiency and quality. They also point out specific challenges, like AI systems themselves being not free of biases, or dangers in providing too much autonomy to AI systems, the latter calling for human-AI teaming as the standard application scenario.

In this paper, I argue that neither the specific challenges of human-AI teaming (e.g., ensuring reasonable human agency [Waefler et al., 2025]), bias [Glüge et al., 2020], nor any of the other risks associated with AI [Stademann, 2025] and how it interacts with our humanity [Segessenmann et al., 2025] *alone* are a solid foundation to talk about AI’s potential use in high-consequence scenarios like the military (civilian uses are also included, e.g., in safety-critical network operations [Roost et al., 2020, Mussi et al., 2025]). Instead, a *basic understanding of the foundational working mechanisms of the technology is necessary for everyone involved* to know the ramifications of these inner workings on the task at hand—ramifications that manifest themselves for example in the “stupid errors” indicated above (which are to be expected once the methods are comprehended). The following sections will provide this understanding (Sec. 2), derive consequences for military and other high-consequential use cases (Sec. 3), and formulate recommendations (Sec. 4).

2 The nature of AI

Artificial intelligence has been defined as the simulation of intelligent behaviour with a computer [McCarthy et al., 1955, Stadelmann, 2025a]. For this, the field of AI, founded in the 1950s, does not offer a unified theory or methodology—there is no one way to “build AI” (the phrase itself is misleading), nor any known path towards anything resembling “artificial general intelligence” (AGI). Rather, AI holds a toolbox full of different methods that are each appropriate to simulate one or several specific behaviours (cp. the definitive AI textbook by Russell and Norvig [2022]).

2.1 Symbolic AI: Logic and reasoning

An important part of the AI toolbox are so-called ‘symbolic’ methods: They manipulate abstract symbols (think: variables as in math) using formal logic to implement rigorous reasoning processes. That is, given a knowledge base of ‘facts’ and ‘rules’, those can be used to infer any logically deducible fact that follows from that knowledge. Respective systems like CYC [Lenat, 1995] were particularly strong in the 80’s and 90’s, fuelled the ‘expert systems’ hype around AI at that time, and have been (and are) used advantageously in high-consequence scenarios since then [Nilsson, 2009]. For example, the AI system used for logistics planning during “Operation Desert Storm” has been said to have “paid back all of DARPA’s 30 years of investment in AI in a matter of a few months” according to Hedberg [2002].

Symbolic methods remain important (e.g., today’s navigation systems calculate their wayfinding based on symbolic AI algorithms like `A*` [Hart et al., 1968]). Yet, they generally suffer from the complexity of the real world: There is a gap between what can be perceived from (potentially error-contaminated) measurements and the clean and abstracted world of logical descriptions (that even the “person-century effort” to build CYC could not bridge despite useful niche applications). Hence, the focus of AI research & development shifted to methods that operate below the ’symbol’ level, directly on data, and are able to adapt to it. There is hope that both methodologies can one day be united, but currently, so-called ‘neuro-symbolic’ AI is still in its infancy [Bhuyan et al., 2024].

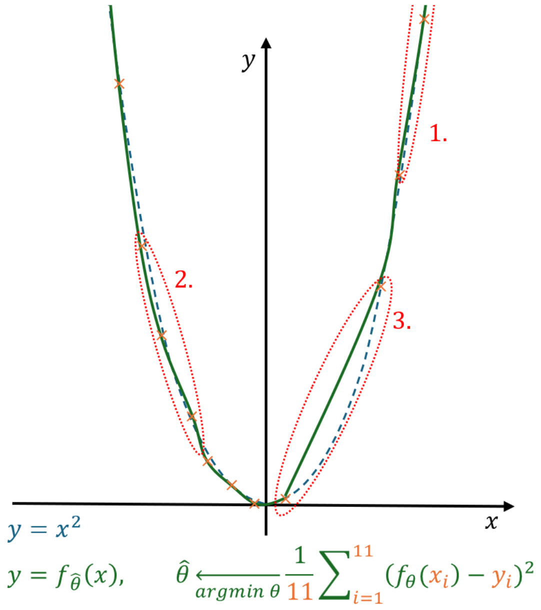

Figure 1: In blue, dotted: A plot of a parable $y = x^2$. A symbolic AI model can be thought of as having access to this true formula, from which it is able to perfectly reproduce each y for any given x. In green, solid: a curve fitted through the 11 orange training data points, resembling the result of training a subsymbolic ML model $y = f_{\hat{\theta}}(x)$ on $\{x_1, x_2, \ldots, x_{11}\}$ (counting training instances from left to right) to find optimal parameters $\hat{\theta}$ through minimizing the prediction error for the known training samples. As real-world data, the training points may contain small measurement errors as visible in the figure (i.e., they do not lie perfectly on the blue, dotted parable); this and other reasons lead to a suboptimal fit (correspondence between the true and the learned model/curve). Consider the red, circled areas: (1) When training data is correct and the used function template for adaptation through training suits the real underlying function, model and true function coincide well. (2) Small measurement errors in the data lead to a suboptimal but likely tolerable fit. (3) In regions with low training data density, no training signal provides guidance for fitting in this area. This leads to larger deviations from the true function; a function with a higher capacity to adapt (like a deep neural network, which can model arbitrarily wiggly functions) could likely zig-zag around wildly between the two far-apart training points x_8 and x_9 .

2.2 Subsymbolic AI: Statistical machine learning

Since the mid-1990s, the predominant part of the AI toolbox is machine learning (ML), used wherever the intended behaviour cannot be described by a set of rules (logic). ML's most successful methods are essentially function approximation [Jordan and Mitchell, 2015]: A mapping is sought from data $\vec{x}$ to some outcome $\vec{y}$, which is to be performed by some function $f(x) = y$ (vector notation is commonly dropped). Inputs are usually high-dimensional numeric representations of real-world data: Imagine, for example, x to be the concatenation of all the pixel values (each one integer for a greyscale value or 3 integers to encode colour as red-green-blue) of an image, and y a flag indicating the presence of some object in the image (1 for 'yes', 0 for 'no') [Krizhevsky et al., 2012]. Or x to be a concatenation of so called 'word embeddings' representing a question in natural language, and y being a word embedding for the likely next word (supposedly starting the answer) [Radford et al., 2019]. Or x consisting of two concatenated structured database entries describing situations (weather, geolocation, other properties; all properly numericized and concatenated), and y being a measure of similarity of the two situations [Kaya and Bilge, 2019].

Evidently, ML is a versatile paradigm: Many 'intelligent behaviours' can be stated as mapping from an input to some output. Methodically, a human ML engineer provides a suitable 'function template' (i.e., a function category that can easily represent the mapping; for example, a straight line cannot represent the dotted curve from Fig. 1, but a polynomial could) as well as a 'learning algorithm' that tunes the function template's parameters to a set of given $\{x, y\}$ pairs called 'training data'. In recent years, ’neural networks’, which existed since the field’s inception, rose to unprecedented prominence, becoming the function template of choice for tasks involving perception and cognition [Schmidhuber, 2015, LeCun et al., 2015]. Basically all AI systems that have made the news since 2012 are based on them. This success stems from ’deep’ neural networks (which consist of several consecutive layers of neurons) that give the function a high capacity to adapt to the training examples: They are general function approximators [Hornik et al., 1989]. Still, the suitability for a given task depends on clever choices of their internal ’architecture’ and ’hyperparameters’ (see Segessenmann et al. [2025] for an in-depth explanation for non-technical readers).

What principles underly this way of ’learning’ and are important to understand in order to develop intuition for the nature of ML’s results? ’Statistical’ learning, as it has been called [Vapnik, 1999], approximates an unknown, underlying function based on a finite, noisy set of samples. The goal is to interpolate between these given training instances to generalize to novel, previously unseen instances (a process called ’inductive learning’).Therefor, some parameterizable continuous function is fitted to the training data by systematically adjusting the parameters of that function to minimize a measure of dissimilarity between the predicted and known outcomes ($\hat{y}$ and y, respectively). For neural networks (and many other ML approaches), this optimization process resembles the statistical principle of ’maximum likelihood’ estimation: The resulting function yields the most *likely* result, given all the evidence present in the training data [Prince, 2023]. For classification tasks (i.e., category prediction), it factually implements $f(x) = p(y|x)$, the conditional probability of the outcome y given the data x. For other tasks like regression (the prediction of continuous numeric values), the model outputs point estimates that represent the most likely values given the training data distribution.

This means that the resulting function (also called ’model’ in AI and ML) can be seen as a *probabilistic* function: It predicts a result with a certain likelihood, i.e., involves a measure of uncertainty in the prediction. As Fig. 1 illustrates, this uncertainty might be low in parts of the domain of x with a dense sampling of training data points (and given that (a) the model has been trained on enough data; (b) the chosen function template is suitable for the kind of data and underlying function; and (c) any new instances follow the same underlying distribution as the training data). But it might also be extraordinarily high in areas of the input that are far away from any seen example. As the model implements a continuous mapping, it will still predict a $\hat{y}$, not knowing that it doesn’t know. Also for the developer it is hard to tell in advance how good the model will be: Generally, ML is an empirical science and there is no way of knowing theoretically how well a specific model will do on a task. Rather, the performance is measured experimentally on a ’test set’ (a hold-out portion of the originally training data), and the result is extrapolated to unseen data under the assumption that these will resemble the training data’s distribution.

A couple of properties of this type of ’learning’ are notable: First, only the function template’s parameters are ’learned’ (i.e., fitted to the data); the ’architecture’ (choice of specific function template), hyperparameters (specific detailed choices in the configuration of the function template and learning setup), and learning algorithm is not part of automatic adaptation. They need to be found by a separate process (typically manual selection by a human, though automation is possible [Tuggener et al., 2019]) based on prior knowledge of the problem domain. This knowledge and the algorithmic choices based on them becomes a necessary part of the model as its ’inductive bias’—a predetermined idea where and how to look for the patterns the model seeks to pick up (as Mitchell [1997] points out, any (also human) learning without this bias is futile). Second, as the model picks up all its knowledge only from the fed training data [Stadelmann et al., 2022], what is not in the data will not be in the model (e.g., things humans infer using their ’common sense’), and what was in the data will also be present in the model (e.g., human biases [Glüge et al., 2020], for example through biased judgments present in the $\{x, y\}$ pairs). Third, a ML model does not learn continually; it is iteratively trained on the training data until the model’s fit is sufficiently good. Then, the parameters are fixed and the model is employed on its task without any further learning: Training and ’inference’ are completely disjunct phases in the ML life cycle.

2.3 Artificial vs. human intelligence: Different means, different errors

From the nature of ML outlined above, it becomes evident why models based on neural networks are current AI’s best attempt to deal with the uncertainties and the messiness of real-world data. This is true for image and video analysis, text analysis and generation, geospatial data analysis, analysis of satellite and other sensor data, etc. If such a model is trained well enough to find acceptance into any application, it likely works very well on average and for typical inputs. At the same time, because of the statistical nature of the model (that has not learned about truth and facts, but statistical plausibility), a result might be wrong in any given case.

Various approaches exist to quantify and manage this uncertainty, including Bayesian neural networks [Wang and Yeung, 2020], ensemble methods [Tuggener et al., 2024], and calibrated confidence scoring [Tian et al., 2023]. Active learning frameworks can identify when models encounter unfamiliar inputs [Nguyen et al., 2022], while human-in-the-loop systems maintain human oversight at critical decision points [Zanzotto, 2019]. However, these techniques often require significant computational overhead, specialized expertise to implement correctly, are often not part of commercial / existing systems, and still cannot eliminate the fundamental issue: ML models remain probabilistic approximators that can fail confidently in unexpected ways.

To grasp the impact of the fundamental likelihood to err, consider the following example of a ML model for visual inspection [Stadelmann et al., 2018]: Having a reasonable accuracy of, say, 95% on a per-image basis, the use case may involve inspecting larger items that are fed subsequently as individual image patches into the classifier—sometimes up to 30 patches. This makes the performance of the model on a per-item basis look rather underwhelming: The potentially acceptable 5% chance of being wrong per patch (image) accumulates to a $1 - (0.95^{30}) = 78.5\%$ chance of misclassification per item. Put differently: It is to be expected that *every* use of that AI system for visual inspection makes a wrong overall prediction.

To gauge the ramifications of the statistical nature of predictions further, consider the following example of using a large language model (LLM) [Stadelmann, 2025b]: A so-called 'reasoning' model has been asked the question “The surgeon, who is the boy’s father, says 'I can’t operate on this boy, he is my son!' Who is the surgeon to the boy?” The answer is straight-forward from the question’s text, yet the model replies “The surgeon is the boy’s mother,” which is obviously wrong. But to the model, this makes actually sense, as it goes on to tell: “The riddle plays on the assumption that a surgeon is male.” Indeed, variations of the question exist abundantly on the web as tests for our own human biases, typically associating males with the role of a surgeon. The model has seen all these during its training (LLMs are trained on almost all text openly accessible on the internet) and learned the utter statistical implausibility of answering anything male to a question that looks remotely like the one above. Consequently, the model gives a plainly wrong answer—but one that is totally *plausible* for any AI system built according to the principles of contemporary ML (which are the best ones we currently have; other forms of ML are conceivable, but not yet mature [Sager et al., 2025a]).

This makes it evident that AI (using any of its methods, including ML) works decidedly different than human intelligence (as is already implied by the definition above, stating that intelligent behaviour is simulated rather than intelligence implemented). While on average possibly better than the mean of human outcomes given a specific task, from the different modes of operation under the hood follows that the remaining errors will also be different: AI systems will commit different errors and exhibit different failure patterns than humans. For example, while humans are ill-equipped to sift through high volumes of heterogeneous data because of sheer information overload, AI systems will also overlook and misinterpret things because of their suboptimal (statistical, not causal/common-sensical) understanding of the world.

The different nature of artificial and human intelligence can be finally illustrated with an analogy of a musician and a DJ: While a DJ simulates certain aspects of creating music very well, their method of music creation through remixing and replaying musicians’ original recordings by design is not general. There are many aspects of music beyond the method of turntables and remixing, e.g., certain genres, playing techniques, and settings for musical performances. Similarly, AI does not simulate the way intelligent human behaviour is produced, but certain carefully designed aspects of human behaviour, with a very specific method of cleverly interpolating between pre-recorded behaviour samples. This makes respective models good at certain things (for which they have been designed and tested) and bad for almost any others.

3 Discussion

Summarily, almost all relevant contemporary AI systems are based on ML models that are high dimensional *probability density functions*, which output the most likely predictions given the input data, leading to likely error that have decidedly different error patterns than human experts. While the field has developed various mitigation strategies as outlined above, these approaches address symptoms rather than the underlying statistical nature of ML. This has important ramifications for any operator (and their organization) relying on respective results (predictions):

AI results are not 'neutral'. They have picked up human biases via the training data and are ignorant of anything not represented in the data or not representable or inferable by the chosen model.

AI results tend to regurgitate the past. Applied in the straight-forward way, they reproduce the most likely pattern found in the training data. This can lead to an impoverishment of strategic decisions [Stigler, 1997] and have effects worth of consideration when conflicting parties rely on basically similar AI decision support.

AI results are statistically plausible predictions with a certain likelihood of failure. Being error-free is not part of the methodology; plausible might still be wrong.

If a result is wrong is not known to the model and difficult to predict technically, but typically, any result will be reported with optimistic confidence by an AI system. It must hence be verified by a human capable of doing so independently.

Human errors and AI errors are very different such that AI systems' errors might seem very stupid (and hence unexpected) from a human point of view. This stems from the completely different mechanisms by which these results are achieved, even when based on the same data.

In a military setting, AI is typically meant to make war more precise, specifically with respect to intelligence (now in the 'knowledge-gathering' sense) and targeting [King, 2024]. But this precision is attained differently from human precision and error-prone as pointed out above. Any human stakeholder must be firmly aware of this fact and the underlying reasons to develop healthy 'scepticism' regarding AI's predictions and recommendations. Hence, the typical mode of operation in this and other high-consequence application scenarios is to build human-AI teams, with the final responsibility with the human.

But scepticism (or human oversight) alone is not enough: Psychological research has shown [Waefler et al., 2025] that humans need to have meaningful agency in any collaboration, otherwise they cannot help but become bored, reverting to mere mechanical approval without exercising supervision. A remedy is offered by the concept of *colearning* currently being developed in a European research project[2] for human-AI collaboration in the high-consequence scenario of operating critical network infrastructures [Mussi et al., 2025]. Colearning maintains a setup in which with every interaction both the human and the machine learn from each other via bidirectional information flow: Not only do the humans provide training feedback to a (continually learning [Wang et al., 2024]) machine learning system, but the AI system at the same time provides explainable insights to the human [Dwivedi et al., 2023] that help them understand and scrutinize decisions better. This happens within a long-term, iterative process of co-adaptation through interaction that leads to colearning. The support of such human learning and active involvement in the decision-making process keeps the human interested, engaged, and maintains their sense of agency.

Another aspect for consideration is of systemic nature: AI systems are powerful tools wielded (ultimately) by individual humans. This leads to higher concentrations of power in these individuals. In military settings characterized by high consequence, stressfulness, and life and death decisions, misuse of such power must be prevented. Although this is not new with respect to military staff, AI systems shift the distribution of power in unexpected ways. For example, significant power could fall into the hands of software vendors and model providers (through dependencies) or training data engineers (through changing model behaviour by biasing/poisoning training data), etc. Respective novel aspects of AI security need to be considered as well.

4 Conclusions and outlook

The non-negligible likelihood of AI errors in any one situation necessitates the implementation of processes to ensure human operators and comrades understand the failure modes of their tools and are properly integrated in decision-making.

[2] `https://ai4realnet.eu/`.

The power concentration issues identified above raise fundamental questions about military AI architecture. Rather than prescriptive solutions, we offer a speculative framework that illustrates how the principles of colearning and decentralized systems might address these challenges—questions that merit serious research attention in future work:

Consider combat situations where individual combatants may be augmented by AI systems that provide extended situational awareness (through perception based on additional sensors) and recommendations (based on fast and comprehensive data analysis). Here, power issues become important: Centrally controlled systems would be prone to overriding the individual's meaningful human agency and could lead to a remote-controlled human army not too different from a robotic one. A potential—speculative—solution might be the following one: Every (group of) combatants receives their own individual, decentralized AI assistant [Zhu et al., 2024], able to colearn (cp. [van den Bosch et al., 2019]).

How could such a scenario play out, e.g., on the level of a fireteam and mitigate some of the ramifications highlighted above? First, a setup would be chosen in which each individual AI system must not be overridable by a central unit (ensuring compliance with the chain of command could be achieved by subjecting it directly to the human team leader). Second, each individual AI system would be fine-tuned to its team by being trained together in exercise and real scenarios, so that the resulting human-AI team would know and complement each other's *specific* weaknesses (because it has colearned and thus co-adapted to each other). This makes this AI system, without any anthropomorphising, of personal value to the human team members and worthless for other combatants (e.g., hostile forces). Thus, heightened risks of power misuse are met with checks and balances through a form of human-AI team spirit similarly strong than between human comrades: For example, AI recommendations on ethics would be more likely to be followed by a team if a consequence of not complying could be to lose the digital comrade (that might chose to disintegrate if ignored too often). This way, common human coping mechanisms with stress and differing opinions by social means would translate to the AI team member. Finally, and in conclusion, these points would ensure a proper place for the often unwanted but ultimately important human trait of having mercy.

Acknowledgements

I am grateful for the helpful discussions and input from Helge Stadelmann, Julia Usher, Philipp Merkt and Kasey Barr.

References

Bikram Pratim Bhuyan, Amar Ramdane-Cherif, Ravi Tomar, and TP Singh. Neuro-symbolic artificial intelligence: a survey. *Neural Computing and Applications*, 36(21):12809–12844, 2024.

Rodney Brooks. The seven deadly sins of al predictions. MIT Technology review, `https://www.technologyreview.com/2017/10/06/241837/the-seven-deadly-sins-of-ai-predictions/`, 2017.

Sébastien Bubeck, Varun Chandrasekaran, Ronen Eldan, Johannes Gehrke, Eric Horvitz, Ece Kamar, Peter Lee, Yin Tat Lee, Yuanzhi Li, Scott Lundberg, et al. Sparks of artificial general intelligence: Early experiments with GPT-4. *arXiv preprint arXiv:2303.12712*, 2023.

Rudresh Dwivedi, Devam Dave, Het Naik, Smiti Singhal, Rana Omer, Pankesh Patel, Bin Qian, Zhenyu Wen, Tejal Shah, Graham Morgan, et al. Explainable AI (XAI): Core ideas, techniques, and solutions. *ACM computing surveys*, 55(9):1–33, 2023.

Stefan Glüge, Mohammadreza Amirian, Dandolo Flumini, and Thilo Stadelmann. How (not) to measure bias in face recognition networks. In *IAPR Workshop on Artificial Neural Networks in Pattern Recognition*, pages 125–137. Springer, 2020.

Peter E Hart, Nils J Nilsson, and Bertram Raphael. A formal basis for the heuristic determination of minimum cost paths. *IEEE transactions on Systems Science and Cybernetics*, 4(2):100–107, 1968.

S.R. Hedberg. DART: revolutionizing logistics planning. *IEEE Intelligent Systems*, 17(3):81–83, 2002. doi: 10.1109/MIS.2002.1005635.

Kurt Hornik, Maxwell Stinchcombe, and Halbert White. Multilayer feedforward networks are universal approximators. *Neural networks*, 2(5):359–366, 1989.

Jincai Huang, Yongjun Xu, Qi Wang, Qi Cheems Wang, Xingxing Liang, Fei Wang, Zhao Zhang, Wei Wei, Boxuan Zhang, Libo Huang, et al. Foundation models and intelligent decision-making: Progress, challenges, and perspectives. *The Innovation*, 2025.

M. I. Jordan and T. M. Mitchell. Machine learning: Trends, perspectives, and prospects. *Science*, 349(6245):255–260, 2015. doi: 10.1126/science.aaa8415. URL `https://www.science.org/doi/abs/10.1126/science.aaa8415`.

Subbarao Kambhampati. Can large language models reason and plan? *Annals of the New York Academy of Sciences*, 1534(1):15–18, 2024.

Subbarao Kambhampati, Kaya Stechly, Karthik Valmeekam, Lucas Saldyt, Siddhant Bhambri, Vardhan Palod, Atharva Gundawar, Soumya Rani Samineni, Durgesh Kalwar, and Upasana Biswas. Stop anthropomorphizing intermediate tokens as reasoning/thinking traces! *arXiv preprint arXiv:2504.09762*, 2025.

Mahmut Kaya and Hasan Şakir Bilge. Deep metric learning: A survey. *Symmetry*, 11(9):1066, 2019.

Anthony King. Digital targeting: artificial intelligence, data, and military intelligence. *Journal of Global Security Studies*, 9(2):ogae009, 2024.

Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. ImageNet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25, 2012.

Akarsh Kumar, 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.

Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *Nature*, 521(7553):436–444, 2015.

Douglas B. Lenat. CYC: a large-scale investment in knowledge infrastructure. *Commun. ACM*, 38(11):33–38, November 1995. ISSN 0001-0782. doi: 10.1145/219717.219745. URL `https://doi.org/10.1145/219717.219745`.

Gary Marcus. Deep learning: A critical appraisal. *arXiv preprint arXiv:1801.00631*, 2018.

John McCarthy, Marvin L Minsky, Nathaniel Rochester, and Claude E Shannon. A proposal for the Dartmouth summer research project on artificial intelligence. Research proposal, August 1955.

Herwin W Meerveld, RHA Lindelauf, Eric O Postma, and M Postma. The irresponsibility of not using AI in the military. *Ethics and Information Technology*, 25(1):14, 2023.

Tom M Mitchell. *Machine learning*. McGraw-Hill New York, 1997.

Marco Mussi, Alberto Maria Metelli, Marcello Restelli, Gianvito Losapio, Ricardo J. Bessa, Daniel Boos, Clark Borst, Giulia Leto, Alberto Castagna, Ricardo Chavarriaga, Duarte Dias, Adrian Egli, Andrina Eisenegger, Yassine El Manyari, Anton Fuxjäger, Joaquim Geraldes, Samira Hamouche, Mohamed Hassouna, Bruno Lemetayer, Milad Leyli-Abadi, Roman Liessner, Jonas Lundberg, Antoine Marot, Maroua Meddeb, Viola Schiaffonati, Manuel Schneider, Thilo Stadelmann, Julia Usher, Herke Van Hoof, Jan Viebahn, Toni Waefler, and Giacomo Zanotti. Human-AI interaction in safety-critical network infrastructures. *iScience*, page 113400, 2025. ISSN 2589-0042. doi: https://doi.org/10.1016/j.isci.2025.113400. URL `https://www.sciencedirect.com/science/article/pii/S258900422501661X`.

Arvind Narayanan and Sayash Kapoor. AI as normal technology. 25-09 Knight First Amend. Inst. (Apr. 14, 2025), `https://perma.cc/HVN8-QGQY`, 2025.

Daniel Neururer, Volker Dellwo, and Thilo Stadelmann. Deep neural networks for automatic speaker recognition do not learn supra-segmental temporal features. *Pattern Recognition Letters*, 181: 64–69, 2024.

Vu-Linh Nguyen, Mohammad Hossein Shaker, and Eyke Hüllermeier. How to measure uncertainty in uncertainty sampling for active learning. *Machine Learning*, 111(1):89–122, 2022.

Nils J Nilsson. *The quest for artificial intelligence*. Cambridge University Press, 2009.

Simon JD Prince. *Understanding deep learning*. MIT press, 2023.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019.

Dano Roost, Ralph Meier, Stephan Huschauer, Erik Nygren, Adrian Egli, Andreas Weiler, and Thilo Stadelmann. Improving sample efficiency and multi-agent communication in RL-based train rescheduling. In *2020 7th Swiss Conference on Data Science (SDS)*, pages 63–64. IEEE, 2020.

Stuart J Russell and Peter Norvig. *Artificial intelligence: a modern approach, 4^{th} edition*. Pearson, 2022.

Pascal J. Sager, Jan M. Deriu, Benjamin F. Grewe, Thilo Stadelmann, and Christoph von der Malsburg. The cooperative network architecture: Learning structured networks as representation of sensory patterns, 2025a. URL `https://arxiv.org/abs/2407.05650`.

Pascal J Sager, Benjamin Meyer, Peng Yan, Rebekka von Wartburg-Kottler, Layan Etaiwi, Aref Enayati, Gabriel Nobel, Ahmed Abdulkadir, Benjamin F Grewe, and Thilo Stadelmann. A comprehensive survey of agents for computer use: Foundations, challenges, and future directions. *arXiv preprint arXiv:2501.16150*, 2025b.

Jürgen Schmidhuber. Deep learning in neural networks: An overview. *Neural networks*, 61:85–117, 2015.

Jan Segessenmann, 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(1):1–32, 2025.

David Silver and Richard S Sutton. Welcome to the era of experience. In *Designing an Intelligence*. MIT Press, 2025.

Thilo Stadelmann. A guide to AI. *Global Resilience White Papers*, 2025a.

Thilo Stadelmann. How not to fear AI. TEDxZHAW (Apr. 10, 2025), `https://stdm.github.io/How-not-to-fear-AI/`, 2025b.

Thilo Stadelmann, Mohammadreza Amirian, Ismail Arabaci, Marek Arnold, Gilbert François Duivesteijn, Ismail Elezi, Melanie Geiger, Stefan Lörwald, Benjamin Bruno Meier, Katharina Rombach, et al. Deep learning in the wild. In *IAPR Workshop on Artificial Neural Networks in Pattern Recognition*, pages 17–38. Springer, 2018.

Thilo Stadelmann, Vasily Tolkachev, Beate Sick, Jan Stampfli, and Oliver Dürr. Beyond ImageNet: deep learning in industrial practice. In *Applied data science: lessons learned for the data-driven business*, pages 205–232. Springer, 2019.

Thilo Stadelmann, 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(2), 2022.

Thilo Stademann. Debate: Evidence-based AI risk assessment for public policy. *Public Money & Management*, 2025.

Stephen M Stigler. Regression towards the mean, historically considered. *Statistical methods in medical research*, 6(2):103–114, 1997.

Yunlong Tang, Jing Bi, Siting Xu, Luchuan Song, Susan Liang, Teng Wang, Daoan Zhang, Jie An, Jingyang Lin, Rongyi Zhu, Ali Vosoughi, Chao Huang, Zeliang Zhang, Pinxin Liu, Mingqian Feng, Feng Zheng, Jianguo Zhang, Ping Luo, Jiebo Luo, and Chenliang Xu. Video understanding with large language models: A survey. *IEEE Transactions on Circuits and Systems for Video Technology*, pages 1–1, 2025. doi: 10.1109/TCSVT.2025.3566695.

Katherine Tian, Eric Mitchell, Allan Zhou, Archit Sharma, Rafael Rafailov, Huaxiu Yao, Chelsea Finn, and Christopher D Manning. Just ask for calibration: Strategies for eliciting calibrated confidence scores from language models fine-tuned with human feedback. In *Proceedings of EMNLP*, 2023.

Lukas Tuggener, Mohammadreza Amirian, Katharina Rombach, Stefan Lörwald, Anastasia Varlet, Christian Westermann, and Thilo Stadelmann. Automated machine learning in practice: state of the art and recent results. In *2019 6th Swiss Conference on Data Science (SDS)*, pages 31–36. IEEE, 2019.

Lukas Tuggener, Raphael Emberger, Adhiraj Ghosh, Pascal Sager, Yvan Putra Satyawan, Javier Montoya, Simon Goldschagg, Florian Seibold, Urs Gut, Philipp Ackermann, et al. Real world music object recognition. *Transactions of the International Society for Music Information Retrieval*, 7(1):1–14, 2024.

Karel van den Bosch, Tjeerd Schoonderwoerd, Romy Blankendaal, and Mark Neerincx. Six challenges for human-AI co-learning. In *International Conference on Human-Computer Interaction*, pages 572–589. Springer, 2019.

Vladimir N Vapnik. An overview of statistical learning theory. *IEEE transactions on neural networks*, 10(5):988–999, 1999.

Christoph von der Malsburg, Thilo Stadelmann, and Benjamin F Grewe. A theory of natural intelligence. *arXiv preprint arXiv:2205.00002*, 2022.

Toni Waefler, Samira Hamouche, and Andrina Eisenegger. The Supportive AI framework: From recommending to supporting. In Dylan D. Schmorrow and Cali M. Fidopiastis, editors, *Augmented Cognition*, pages 303–317, Cham, 2025. Springer Nature Switzerland. ISBN 978-3-031-93724-8.

Hao Wang and Dit-Yan Yeung. A survey on Bayesian deep learning. *ACM computing surveys (csur)*, 53(5):1–37, 2020.

Liyuan Wang, Xingxing Zhang, Hang Su, and Jun Zhu. A comprehensive survey of continual learning: Theory, method and application. *IEEE transactions on pattern analysis and machine intelligence*, 46(8):5362–5383, 2024.

Fabio Massimo Zanzotto. Human-in-the-loop artificial intelligence. *Journal of Artificial Intelligence Research*, 64:243–252, 2019.

Changxi Zhu, Mehdi Dastani, and Shihan Wang. A survey of multi-agent deep reinforcement learning with communication. *Autonomous Agents and Multi-Agent Systems*, 38(1):4, 2024.

Nikola Žigulić, Matko Glučina, Ivan Lorencin, and Dario Matika. Military decision-making process enhanced by image detection. *Information*, 15(1), 2024. ISSN 2078-2489. doi: 10.3390/info15010011. URL `https://www.mdpi.com/2078-2489/15/1/11`.