**AI and Ethics manuscript No.**
(will be inserted by the editor)

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

**Thilo Stadelmann, Philipp H. Merkt, and Kasey Barr**

Received: 18.10.2025 / Accepted: 09.02.2026

**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

T. Stadelmann is a Fellow of the ECLT European Centre for Living Technology, 30123 Venice, Italy, and a member of the Scientific Council of the IAEAI Israeli Association for Ethics in Artificial Intelligence, Tel Aviv, Israel.

Thilo Stadelmann
Centre for Artificial Intelligence
Zurich University of Applied Science
8400 Winterthur, Switzerland
E-mail: stdm@zhaw.ch

Philipp H. Merkt
Research and Education Center for Extraordinary Tactical Situations and Strategic Resilience (18_RECESS), Fresenius University of Applied Sciences
65510 Idstein, Germany; and
Chair for the Education of Personal and Interpersonal Competencies in Healthcare, Faculty of Health (Department of Human Medicine), Witten/Herdecke University
Alfred-Herrhausen-Straße 50, 58455 Witten, Germany

Kasey Barr
Rubenstein Center for Constitutional Challenges
Reichman University
Herzliya, 4610101, Israel

factual correctness, and exhibit confident behaviour even when wrong. Written to provide guidance for non-technical stakeholders, specifically but not exclusively in the military domain, it posits that 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 co-learning frameworks that maintain meaningful human control through bidirectional information flow, and behavioural and functional awareness on the human side. We give an outlook to decentralized, co-learned AI system tailored to specific teams in dedicated co-learning labs to mitigate power concentration risks while preserving essential human capacities, including moral judgment to exercise mercy.

**Key words:** artificial intelligence, military decision-making, human-AI collaboration, co-learning, meaningful human control, strategic resilience

## 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, 2025b), 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 multistep 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 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 shall human operators deal with such fluctuation in order fulfillment by their AI systems? 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 collaboration as the standard application scenario. Indeed, specific challenges of human-AI collaboration exist (e.g., ensuring reasonable human agency (Waefler et al., 2025)), bias (Glüge et al., 2020) and other risks associated with AI (Stadelmann, 2025a) need careful mitigation, and how AI interacts with our humanity (Segessenmann et al., 2025) will hopefully be a major theme of future thought. But will AI's 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)) be automatically to the advantage of respective organizations once such challenges—those that are applicable for a task at hand—are handled well?

In this paper, we argue from the point of view that a *basic understanding of the foundational working mechanisms of this technology is necessary for everyone involved* to know the ramifications of its 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 hence will provide this understanding (Sec. 2), derive consequences for military and other high-consequential use cases (Sec. 3), and formulate recommendations (Sec. 4). These recommendations align with the literature on Meaningful Human Control (MHC), which argues that 'meaningful' control is not satisfied by nominal human presence or a formal veto. What matters is whether socio-technical systems remain appropriately responsive to human reasons and support responsibility attribution (Santoni de Sio and Van den Hoven, 2018; Mecacci and Santoni de Sio, 2020; Veluwenkamp, 2022). Rather than re-litigating the MHC debate, this paper translates the stochastic properties of contemporary machine learning (ML) into practical design, training, and governance implications that bear directly on when oversight is genuinely meaningful.

## 2 The nature of AI

Artificial intelligence has been defined as the simulation of intelligent behaviour with a computer (Merriam-Webster, 2021; Fuchs, 2024; Stadelmann, 2025b)(cp. the opening hypothesis in McCarthy et al., 1955). 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 which stand for some semantic object) 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 1980's and 1990'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" (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).

### 2.2 Subsymbolic AI: Statistical machine learning

Since the mid-1990s, the predominant part of the AI toolbox is 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

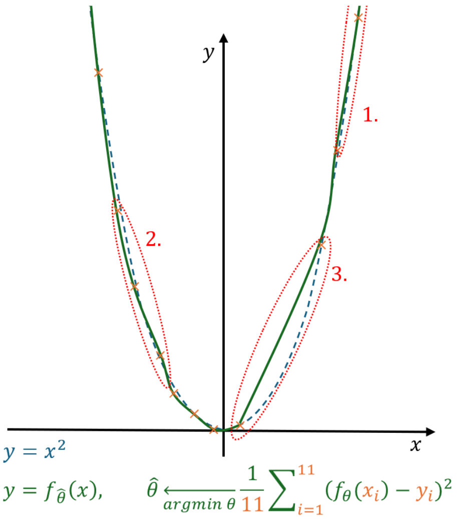

**Fig. 1** In blue, dotted: a plot of a parable $y = x^2$ as a toy example of some real-world event which's outcomes ($y$) shall be predicted by AI. A symbolic AI model can be thought of as having access to this true equation, from which it is able to perfectly reproduce each $y$ for any given $x$. In 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 $x_9$ .

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. In any case, what has been learned (implicitly or explicitly) is the probability density of outputs given the inputs.

This means that the resulting function (also called 'model' in AI and ML) has to 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 accurate 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 original 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 become 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 usually does not learn continually: while there is a subfield of ML called 'continual learning', there are still numerous open challenges to solve (Purushwalkam et al., 2022; Verwimp et al., 2024). Hence, the vast majority of deployed ML-based systems, for example all widely known GenAI systems like large language models (LLMs), do not learn continually (Kontogianni et al., 2024) (actually, we are not aware of any that does). Rather, they are iteratively trained on the training data until the model's fit is sufficiently good. Then, the parameters are fixed and the model is deployed 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 currently 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 LLM (Stadelmann, 2025c): 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. For example, a DJ cannot produce what cannot be reached by mixing exisiting recordings, hence new musical genres like the Grunge of the 1980s or New Metal in the 1990s would never emanate from them. Similarly, AI does not simulate the way intelligent human behaviour is produced, but certain carefully designed aspects of human behavioural outcomes, 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 errors that have decidedly different error patterns than human experts. We will discuss ensuing general implications for any high-consequence use of AI in Sec. 3.1, from which we derive the need for and foundations of meaningful human-AI collaboration in Sec. 3.2–3.3 before outlining the necessity for strategic resilience with respect to novel security issues in Sec. 3.4.

### 3.1 General implications

While the field of AI has developed various mitigation strategies for issues stemming from the discrepancy between being a probability density function and being perceived as humanlike 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'.* Models 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 competing (or conflicting) parties rely on basically similar AI decision support (think of the same advisor working for all parties).

*AI results have a certain likelihood of failure;* being error-free is not part of the methodology. That predictions are statistically plausible does not prevent them from still possibly being wrong.

*It is not known to a model if its current output is right or wrong,* and it is difficult to predict the correctness technically. Anyway, any result will be reported with optimistic confidence by an AI system. Results 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.

On the systemic level, it is noteworthy that AI systems are powerful tools wielded (ultimately) by individual humans. This leads to higher concentrations of power in these individuals. In high consequence settings that are characterized by stressfulness and life and death decisions (e.g., military use on the battlefield), 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.

The next sections will indicate directions for mitigation regarding these implications with increasing focus on uses of AI in military contexts.

### 3.2 The case for ML literacy, co-learning, and decentralized systems

AI results, which are typically meant to make analyses more precise, e.g. in a military setting with respect to intelligence (in the sense of 'knowledge gathering') and targeting (King, 2024), are 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. Here, our assertion that a "basic understanding of the foundational working mechanisms of this technology is necessary for everyone involved" (cp. Sec. 1 and Tigard (2025)) plays out: with it, one *expects* the above implications in working with such a system.

For example, one is not surprised that a high-confidence recommendation by a purely ML-based *situational awareness* system can turn out sub-optimal, because it fuses information channels in a shallower way a human would Liu et al. (2025), namely based on low-level statistical signals rather than underlying meaning or causation. An operator with said basic understanding of ML would hence often check some of those signals that a human typically looks for and compare these results with the system's to gain further practical understanding for the kind of situations where judgment coincides. The basic understanding thus provides a generalizable layer of initial, realistic expectations that can further be adapted to specific cases through additional study or lived experience. It constrains and guides this adaptation and thus does much more than accelerating the process of arriving at realistic expectations—it ensures them, by changing the human's internal estimates of what are likely ML outputs based on an understanding of their causes. The following hypothetical examples illustrate this further:

*Target prioritization under pattern saturation:* An AI system designed to assist with target prioritization ranks objects or locations based on patterns learned from prior conflicts. The system may assign high confidence to a particular site because its features closely match historical training examples associated with hostile activity. From the model's perspective, the recommendation is statistically well supported. A human operator familiar with recent changes on the ground may recognize that the same pattern now reflects a civilian logistics function that emerged after the training data was collected, and chose to revisit the ML system's assessment of it, knowing that the system's output may not be erroneous in a technical sense: it correctly identifies a familiar pattern. Yet it is operationally misleading because statistical similarity is mistaken for present relevance. This illustrates how ML systems can produce confident but contextually incorrect recommendations, and why oversight requires judgment (based on understanding) rather than simple verification.

*Anomaly detection and quiet failure:* In anomaly-detection applications, an AI system may flag only what deviates sharply from learned norms. Gradual changes, such as slow shifts in movement patterns or communication behavior, may remain unremarkable to the system while raising concern for experienced analysts. The system does not miss the signal. It never learned to treat slow drift as meaningful. Here, the limitation arises not from lack of data, but from the statistical framing of relevance itself. This type of quiet failure is particularly challenging for both automated detection and human oversight. It can be anticipated, however, if analysist using the AI system understand its working mechanism.

Hence, the typical mode of operation in the abovementioned and other high-consequence application scenarios is to build human-AI 'teams',[1] with the final responsibility with the human (in terms of Davidovic (2025): for the prupose of safety, with meaningful human control and judgment as the type of engagement, analysing the whole process). 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 *co-learning* 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). Co-learning 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)) ML 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. [3] This happens within a long-term, iterative process of co-adaptation through

[1] As with many of the terms used to describe AI systems (e.g., 'intelligence', 'learning', 'thinking'), the line for undue antropomorphisation is crossed with 'teaming' (Seeber et al., 2020; National Academies of Sciences, Engineering, and Medicine, 2021; Gunkel and Wales, 2021). We adopt it here as an established technical term when we connect to the relevant literature or want to emphasize that the AI tool in this collaboration acts as a very 'active appliance' (Shneiderman, 2022); otherwise, we prefer the umbrella term 'human-AI collaboration:' Although duely criticized for the very same reasons (Evans et al., 2025), broadly accepted better alternatives are lacking beyond 'human-AI interaction,' which does not convey the 'active appliance' aspect and is thus deemed insufficient here: you can also interact, for example, with a hammer, but we are concerned with a different kind of—a different quality or level of—interaction here.

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

[3] Explainability through AI systems is a philosophically hotly debated topic: do technical artifacts even possess what it takes to explain (O'Hara, 2020; Mattioli et al., 2024)? Certainly they do not, in the human sense, and yet the term is helpful as an umbrella for methods such as Grad-CAM for visual inputs (Selvaraju et al., 2017), which help humans gain intuition into the 'why' behind respective results, helping with their interpretation.

interaction that leads to co-learning. 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.

Also 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 co-learning and decentralized systems might address these challenges—questions that merit serious research attention in future work (cp. Sec. 4): 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 co-learn (cp. (van den Bosch et al., 2019)).

### *3.2.1 Illustrative examples for co-learning systems in military practice*

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 unit would 'know' and complement each other's *specific* weaknesses (because it has co-learned and thus co-adapted to each other). This makes this AI system, without implying any anthropomorphization,[4] 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 what in a human-human collaboration would be called a joint 'team spirit:' For example, AI recommendations on ethics would be more likely to be followed by a human 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

[4] As Sec. 2 showed, AI systems 'know' (etc.) in a very different sense than humans do. Why then this anthropomorphism? Apart from hype and fallacy (Placani, 2024), finding helpful language in the context of AI, which is deliberately designed to mimick human behaviour at the surface, is difficult: it necessitates balancing the painting of useful word pictures for naturally human behavioural contexts while not making the technology appear more than it is. Within the scientific fields of AI/ML, the shared implicit understanding is that most of these terms are used as technical terms with figurative meaning. This leads to problems in interdisciplinary or public dialogue where this practice is not known or shared. Therefore, we caution readers to understand terms figuratively that are commonly used in a human context but appear here in relation to AI systems.

human coping mechanisms with stress and differing opinions by social means would translate to the AI part of the collaboration.

A second scenario shall serve to illustrate how the postulated co-adaptation benefits the involved humans and the overall quality of the result: the *'Recover' task within a Personnel Recovery (PR) operation* (cp. Holewijn (2011), especially Fig. 3 there). It comprises the complex and complicated task of identifying with high confidence, on the ground under immense time pressure and adverse circumstances, an isolated member of one's own forces, e.g., within a Combat Search & Rescue mission: The decision is to be taken whether the target is taken up (because the person can be identified with certainty, and the situation is reasonably safe) or left behind (because there are doubts on the person's identity or whether the situation constitutes an ambush). Whether the decision is ultimately taken by the commander on the ground or in the Tactical Operation Center, AI systems can contribute important cues with respect to situational awareness (assessing safety, cp. U.S. Department of Defense (2008)) and biometrics (adding to identity recognition of persons that may be wounded or disfigured) by means of their pattern recognition powers, i.e., the AI system's recommendation reflects statistical aggregation across prior examples. But the human may hesitate because of causal and situational reasoning about how the present case departs from those examples. The difference is not one of speed or accuracy alone, but of mode of judgment, and the resulting error risks are different in kind.

How can this judgment be honed and the AI system's cues (that could be incorrect at any time, either due to error or because certain input modalities were not taken into account, like smell) be duely incorporated? We argue: By joint experience gained in *co-learning labs* (see below and Sec. 3.3.1) will the AI system have (machine-) learned, for example, what specific foci of attention 'Sgt. Snuffy', who leads the operation on the ground, has. Hence, the system will use this training to provide her with a tailored priorization of inputs targeted at complementing this specific human's judgment in the most efficient way that does not lead to cognitive overload. Through the same joint training, Sgt. Snuffy has learned in what kind of situations the AI recommendations are reliable and what environmental cues are indicative of the system missing something or appearing too confident. As decisions need to be taken in split seconds, this fine tuning of tool and human to their specific cooperation can lead to vital advantages while not making such an AI system (as a technical equipment item) less controllable or safe for the organization deploying it.

How did the fine tuning take place? Sgt. Snuffy and her team exercised several training PR missions in a co-learning lab (think of a Urban Warfare Training Center with additional focus on AI), having the respective AI system as part of their euqipment. Using it in training gave rise to the mentioned experience for the humans, while the lab environment captured multi-sensory data (e.g., from fixed and helmet cameras, radio, tracking systems installed in the facility, etc.) that have later been used to perform 'fine tuning' on the ML models underlying the AI system (cp. (Koedijk et al., 2026) and how ML models are fine-tuned for a different purpose by Ruiz and Sell (2024)).

### 3.3 Behavioral and functional foundations of meaningful human control

Building on the co-learning framework discussed above, sustaining MHC in high-consequence environments through co-learning requires that human operators cultivate awareness of both *behavioral* and *functional* factors. This emphasis is aligned with the MHC literature, which argues that 'meaningful' control depends on whether the overall human–AI arrangement remains appropriately responsive to human reasons and supports responsibility attribution in practice (Chengeta, 2016; Ekelhof, 2019; Mecacci and Santoni de Sio, 2020; Santoni de Sio and Van den Hoven, 2018).

*Behavioral* awareness involves understanding how stress, time pressure, and cognitive bias influence human judgment when interacting with probabilistic systems. Decades of research in behavioral decision science demonstrate that even experts, individually and in groups (Barr and Mintz, 2022, 2018), are prone to overconfidence, anchoring, and inconsistent evaluations under uncertainty (Kahneman and Tversky, 1979; Kahneman et al., 2021; Dror, 2020). High workload and time pressure further degrade attention and vigilance, changing how operators notice, interpret, and respond to automated cues (Endsley, 1995). These conditions also intensify well-documented patterns of automation misuse, including complacency and automation bias, in which users defer to an aid even when it is wrong or context-mismatched (Parasuraman and Manzey, 2010; Parasuraman and Riley, 1997; Merritt and Ilgen, 2008). Because reliance is not static, behavioral awareness must include trust calibration: The capacity to maintain appropriately proportional reliance as performance, context, and incentives shift (Lee and See, 2004).

*Functional* awareness complements this behavioral awareness: Operators must grasp how machine-learning systems represent uncertainty, how their reliability changes with context, and where their confidence diverges from causal truth (Lyons et al., 2021; Gao et al., 2023). Transparency and explainability are critical to functional awareness as they enable operators to interpret how AI systems reason under uncertainty and challenge their recommendations appropriately (Miller, 2019).

In high-consequence domains, 'accuracy in testing' is not sufficient as a proxy for trustworthiness in deployment: Dataset shift, underspecification, and out-of-distribution conditions can produce brittle or unstable behavior that is not visible in routine validation (Ovadia et al., 2019; D'Amour et al., 2022; Hendrycks and Gimpel, 2017). Functional awareness therefore includes (a) understanding calibration limits and what probabilistic confidence does (and does not) mean (Guo et al., 2017), and (b) understanding why interpretability and explanation are not merely transparency virtues but practical tools for contestation, error detection, and bounded reliance (Rudin, 2019; Doshi-Velez and Kim, 2017). When operators appreciate both their own cognitive dynamics *and* the variability and boundary conditions of AI performance, human–AI collaboration can evolve from passive oversight into an adaptive process of mutual calibration.

This functional awareness is central to this article's core claim: Because contemporary ML is stochastic and context-sensitive, meaningful human oversight depends on operators understanding how uncertainty is represented. When reliability shifts across contexts, and where system confidence can be overrepresented, reliance must be a matter of judgment based on clear understanding of the boundaries within which the system can perform with accuracy and predictability. How can this be implemented?

### *3.3.1 Towards implementing MHC via co-learning*

Co-learning scenarios should aim to build *reciprocal situational awareness*: A two-way understanding in which humans learn how AI performance varies across operational conditions, and AI systems (through design, training, interface, and feedback loops) are shaped to anticipate predictable human vulnerabilities and restrictions under stress, ambiguity, and tempo. This concept builds on established work on situation awareness and human–automation interaction, while emphasizing co-adaptation rather than one-directional 'user training' alone (Endsley, 1995; Parasuraman and Riley, 1997; Parasuraman and Manzey, 2010).

Returning to the concept of co-learning labs from Sec. 3.2.1, they can now more generally be understood as real-world training and evaluation environments in which operational situations (including high-tempo decision contexts) are simulated through enactment, enabling teams to practice with AI-enabled decision support while generating structured evidence about where human and machine reliabilities intersect, diverge, or degrade: As noted, machines excel at rapid pattern detection and probabilistic reasoning, yet lack causal comprehension and moral sensitivity. Humans bring contextual judgment and ethical evaluation but are vulnerable to fatigue, stress, framing effects, and diffusion of responsibility. Hence, training environments that expose where human and machine reliability cross and interact need to be established that can foster the kind of co-adaptive awareness needed for safe deployment.

These environments are not simply 'training ranges.' They are governance-relevant infrastructures that (a) reveal predictable failure modes (behavioral and functional), (b) support the refinement of interfaces, escalation protocols, and contestation pathways, and (c) produce auditable learning artifacts that can feed test-and-evaluation, doctrine, and accountability. In this sense, co-learning supports MHC by operationalizing the conditions under which oversight is *actually* meaningful: it strengthens the operator's capacity to contest, recalibrate, and redirect reliance, and it strengthens the system's design alignment with human reasons, responsibility practices, and institutional review.

The co-learning framework advocated for here is not intended to replace established concepts such as calibrated trust (Lee and See, 2004). Rather, the two operate at different analytical levels. Calibrated trust refers to a state of appropriate proportional reliance on an automated system at a given time, based on perceived competence, predictability, and context. Co-learning, by contrast, describes an iterative process through which such calibration is developed and maintained over time. By repeatedly exposing human operators to system behavior across varying conditions—including uncertainty, degradation, and failure, co-learning environments support the ongoing adjustment of reliance as system performance and operational contexts evolve. In this sense, co-learning functions as a practical mechanism for sustaining calibrated trust in settings where static training or one-time validation is insufficient.

The behavioral–functional approach outlined above clarifies why MHC cannot be secured by procedural oversight alone. It requires disciplined operator judgment *and* system-level literacy about how probabilistic models behave at and beyond their domain limits. Then, co-learning labs provide a practical bridge between these requirements by turning abstract commitments to control into trained competencies, validated boundaries, and institutionally usable evidence of when and why reliance is warranted: research on human-autonomy teaming highlights that robust collaboration depends on dynamic trust, transparency, and shared (or, in the case of AI: appropriately aligned) mental models that allow both sides to anticipate one another's limitations (Lyons et al., 2021; O'Neill et al., 2022). Co-learning can be the mechanism for strengthening the conditions of MHC (i.e., sustained engagement, calibrated reliance, and contestability) given well-documented risks of complacency and automation bias in human–automation interaction (Parasuraman and Riley, 1997; Parasuraman and Manzey, 2010; Lee and See, 2004).

Humans remain the central decision-makers, exercising authority most effectively when they understand their own cognitive limits and the probabilistic nature of AI reasoning. In this sense, co-learning can support calibrated agency, enabling mutual adaptation while preserving the moral and operational accountability of human operators. Training environments that reveal where human and machine reliability intersect provide a plausible path to cultivating the co-adaptive awareness required for safe deployment. A behavioral-functional design provides a foundation for responsible co-learning and for ensuring that human judgment remains active, informed, and accountable even within stochastic, high-tempo decision systems.

### *3.3.2 The limits of human oversight*

Human oversight is often treated as an inherent safeguard, yet the human-automation literature shows that oversight can be protective or counterproductive depending on operational and cognitive conditions. Oversight is most likely to *help* when operators can independently evaluate the system's output, when the task tempo permits verification, and when interfaces support situation awareness and informed contestation rather than passive acceptance (Endsley, 1995; Lee and See, 2004). In these conditions, human judgment functions as a meaningful check on probabilistic outputs, particularly under uncertainty or domain shift.

However, oversight can *harm* when cognitive workload, time pressure, or organizational incentives push operators toward shallow review, producing 'rubber-stamping' rather than genuine evaluation. Under high perceived automation reliability, users tend to drift into complacency and automation bias, deferring to system outputs even when they are incorrect or context-mismatched (Parasuraman and Riley, 1997; Parasuraman and Manzey, 2010; Merritt and Ilgen, 2008). This risk is compounded when operators lack functional awareness about what model confidence signals mean, or when the human role becomes supervisory monitoring in low-engagement conditions that predict vigilance decrements (Endsley, 1995; Parasuraman and Manzey, 2010). In short, 'human in the loop' or 'human on the loop' is not a sufficient condition for meaningful control if the loop is cognitively thin or institutionally pressured.

These findings imply that the value of oversight is conditional and predictable. Factors that tend to improve oversight include time and workload margins, clear contestation pathways, feedback that enables calibration, and interface design that supports active verification (Lee and See, 2004).

### 3.4 Strategic resilience to mitigate hybrid security threats

ML's proneness to bias, confusing of plausibility with factuality, and signalling a high degree of self-confidence even when producing incorrect results also open up new possibilities for attack and disruption. These become strategically relevant for military use in light of hybrid threats (Kambouris, 2024): For example, AI systems intended as aids in the MDMP can be deliberately thrown off balance by manipulating data streams. The hybridization of conflict, in which conventional operations are interwoven with cyberattacks, disinformation, and bioterrorist scenarios, makes AI a double-edged sword in military use: It is both an efficiency enhancer and a target for attack. For effective preventive threat prevention, Badalič (2024) hence emphasizes that threats must be addressed ex ante before they become effective in combat, which can in part be addressed by meaningful human oversight.

Jonsson and Käihkö (2025) expand the view of conflict arenas beyond the battlefield with their approach to non-military warfare. The challenges are particularly acute where open-source intelligence (OSINT) data is used for training and mission evaluation by AI: Open sources are heterogeneous, manipulable, and often express aggressive narratives. When AI systems that operate on the principle of statistical plausibility learn from OSINT data, this can distort the situation assessment in lieu of supposed operational effectiveness (cf. Ziehr and Merkt, 2024). Again, meaningful human oversight is the required counter measure.

Marquis (1997) shows that asymmetric actors are particularly successful in exploiting the weaknesses of superior systems. Freudenberg (2010)'s theory of irregular warfare makes it clear that asymmetric warfare operates through methods such as deception, infiltration, or overstretching of enemy forces. Applied to the above analysis of AI error modes, this means that irregular opponents can deliberately provoke the inherent susceptibility to bias of probabilistic models—for example, through 'poisoned' OSINT data or simulated patterns—in order to gain disproportionate influence over operational decisions.

To counter these threats systemically, strategic resilience is required, which can be located on three levels (Souchon, 2020): at the micro level among soldiers, analysts, and AI assistance systems; at the meso level among military organizations and networks; and at the macro level among the state, society, and the international order. Resilience means more than just resistance: it encompasses the ability to address threats ex ante preventively, ex nunc at the moment of action, and ex post in the sense of organizational learning. Constellation analysis (Ohlhorst and Schön, 2015) is a suitable methodological tool for this purpose, as it reveals the interactions between actors, means, goals, and dynamics and, in conjunction with the hybrid methods of asymmetric warfare described by Freudenberg, shows where AI systems are vulnerable and how their integration into military constellations can be designed responsibly (cf. Merkt et al., 2025).

## 4 Conclusions

The non-negligible likelihood of AI errors in any one situation necessitates the implementation of processes to ensure human operators participating in human-AI collaboration understand the failure modes of their tools and are properly integrated in decision-making. Co-learning schemes can help herewith and at the same time train respective AI systems to compensate for specific errors and limitations of 'their' human users. Furthermore, AI is not only a tool, but also a potential target for attacks in asymmetric conflicts. Only when meaningful human oversight and judgment, decentralized architectures, and resilient organizational setups work together can probabilistic systems be prevented from becoming targets and catalysts for strategic misjudgments in highly consequential situations.

### 4.1 Limitations and future work

This article has argued that meaningful oversight based on behavioural and functional (including ML) awareness is necessary for high-consequence AI deployment, but it does not claim that oversight is uniformly beneficial across contexts (see Sec. 3.3.2). The effectiveness of oversight is conditional and depends on operational constraints, human cognitive capacity, and task properties. The following limitations also address directions for future work:

*Human capacity constraints:* Human oversight can degrade under bandwidth limits, cognitive load, and information volume. In high-tempo settings, monitoring roles can predict vigilance decrements and 'rubber-stamping,' especially when verification is effortful and accountability is diffuse. Thus, future work should model oversight as a scarce resource and specify minimum conditions for substantive review (for example, time-to-verify margins, workload thresholds, and interface support for contestation). At the same time, succeeding to implement co-learning frameworks is expected to mitigate a larger proportion of this issue through the improved bi-directional information flow between human and AI, mediated by mutual adaptation.

*Expertise differentiation:* Oversight competence is likely heterogeneous. Domain expertise and what we term *functional awareness* (ability to interpret uncertainty, calibration limits, and domain shift) may shape error detection, but the distribution and magnitude of these effects are not established here. Future research should distinguish (a) domain expertise (operational knowledge), (b) system literacy (knowing what confidence does and does not mean), and (c) procedural expertise (ability to execute verification and escalation under time pressure), and test which combinations predict reliable contestation of AI outputs.

*Expertise transmission:* The co-learning laboratories proposed in this article assume that at least some oversight-relevant skills can be trained, but the pedagogical limits are not yet well characterized. Training may improve calibration and contestation, yet it may also induce overconfidence or brittle heuristics. A priority research agenda is therefore to evaluate training regimes experimentally, with outcomes that include calibration, detection of out-of-distribution failure, and appropriate override behavior under stress and time pressure.

*Failure modes of oversight:* Moreover, human involvement can worsen outcomes when deliberation introduces harmful delays; when cognitive biases dominate under uncertainty; when users over-ride correct AI outputs due to misplaced confidence; or when the human lacks the domain basis to evaluate the recommendation. These risks reinforce a central point of this article: 'human-in-the-loop' is not a sufficient condition for meaningful control. Oversight must be designed so that review is feasible, contestation pathways are clear, and the human role remains cognitively substantive rather than formal. Co-learning labs are proposed as one mechanism for identifying these boundary conditions in practice and for generating auditable evidence about where reliance should be bounded, deferred, or escalated.

In addition to these human factors, open *questions on the ML side* need to be addressed thoroughly in the future to realize co-learning enabled MHC in high-consequence scenarios: how to transform *any* practical ML system used in the respective high-consequence context into one that is continously learning, overcoming memory issues and catasrophic forgetting, and outputting meaningful confidence estimates? How to enable the information flow from AI system to human necessary for human adaptation via XAI, especially for data modalities other than vision? What ML algorithms are appropriate for the machine part of co-learning, and how to evaluate the joint learning progress? Finally: What aspects of psychology need to be incorporated to which degree to facilitate the human learning?

### 4.2 Recommendations

Because AI processing routinely exceeds human review capacity, meaningful verification must rely on selective validation strategies that target anomalies, boundary conditions, and irreversible decisions, rather than comprehensive replication of machine analysis. The following principles are offered as practical orientation precisely because oversight is conditional; they are meant to guide governance toward contexts where oversight is feasible and meaningful, and away from procedural 'human presence' that does not improve outcomes.

To keep the discussion actionable, the technical realities outlined in this article can be translated into practical high-level guiding principles for responsible innovation, forming the acronym *'GUARD'*:

**G**overnance. Assign clear decision authority and responsibility across the AI lifecycle, including who may deploy, modify, pause, or override the system, and under what conditions.

**U**phold Human Dignity. Ensure that high-stakes decisions remain responsive to human reasons and do not reduce persons to data points; where consequences are severe, preserve the ability to pause, reconsider, or defer.

**A**nticipate Error. Treat mistakes and misfit as expected under uncertainty and shifting conditions; build practices that help users recognize when outputs may be unreliable and require additional scrutiny; teach basic ML literacy to all stakeholders.

**R**etain Human Agency. Structure human involvement as active judgment rather than passive monitoring, including meaningful opportunities to question, contest, and redirect reliance on the system; *create, with high priority, co-learning labs to train this.*

**D**ocument Accountability. Maintain traceable records sufficient to reconstruct how the system's output was used and how human judgment was exercised, supporting audit, learning, and responsibility attribution.

Tab. 1 operationalizes GUARD. Absent these safeguards, reliance is more likely to drift from proportional, context-sensitive use toward routine and procedural acceptance. These principles provide a practical orientation for embedding AI decision support in ways that keep oversight meaningful, while leaving the operational specification of thresholds and procedures for future research and evaluation.

Not least, designing, developing, and deploying AI systems (e.g., the speculative ones of Sec. 3.2.1) according to these principles of decentralized, co-learned AI systems, specifically in military contexts, would ensure a proper place for the often unwanted but ultimately important human trait of having mercy.

## Acknowledgements

We are grateful for the helpful discussions with and input received from Helge Stadelmann, Julia Usher, Ricardo Chavarriaga, and Manuel Renold.

| GUARD principle | Oversight requirement | Implementation conditions |
|---|---|---|
| **G**overnance | Clarify decision authority and responsibility across the AI lifecycle (deployment, modification, suspension, override), including conditions under which each applies. | Make authority legible: specify who authorizes use, who can pause/suspend, who can override, and who is responsible for review when operating conditions shift or uncertainty rises (cp. (Diakopoulos et al., 2017)). |
| **U**phold Human Dignity | Ensure high-stakes decisions remain responsive to human reasons and do not reduce persons to data points; where consequences are severe, preserve the ability to pause, reconsider, or defer. | Require a brief human-reasons statement for high-consequence outputs (one sentence explaining the human-relevant grounds beyond the model's output) and preserve a protected pause/deferral pathway where consequences are irreversible or contested. |
| **A**nticipate Error | Treat mistakes and misfit as expected under uncertainty and shifting conditions; build practices that help users recognize when outputs may be unreliable and require additional scrutiny. | Teach basic ML literacy to all stakeholders; couple outputs to uncertainty awareness: require an uncertainty/confidence representation (where available), and include a short "conditions changed?" check (e.g., missing data, degraded inputs, novelty, time pressure) that triggers heightened verification, second review, or deferral. |
| **R**etain Human Agency | Structure human involvement as active judgment rather than passive monitoring, including meaningful opportunities to question, contest, and redirect reliance on the system. | *Create, with high priority, co-learning labs to train this*; preserve contestability by design: for designated high-consequence outputs, require an active verification move (cross-check, alternative hypothesis check, or second reviewer) so reliance is not reduced to rubber-stamping under tempo or perceived system authority. |
| **D**ocument Accountability | Maintain traceable records sufficient to reconstruct how the system's output was used and how human judgment was exercised in order to support auditing, learning, and responsibility attribution. | Maintain a minimal standard of process preservation to include output consulted (and uncertainty representation, if present) and a brief rationale noting salient constraints (time pressure, missing inputs, uncertainty); this can be automated in co-learning lab environments. |

**Table 1** Operationalizing GUARD. The oversight requirements resemble the short description above. The implementation conditions give further insight towards operationalization.

## Declarations

Funding.

Aspects of this work have been developed within the AI4REALNET project, which has received funding from European Union's Horizon Europe Research and Innovation programme under the Grant Agreement No 101119527 and from the Swiss State Secretariat for Education, Research and Innovation (SERI). Views and opinions expressed are, however, those of the author(s) only and do not necessarily reflect those of the European Union and SERI. Neither the European Union nor the granting authority can be held responsible for them.

Competing Interests.

The authors have no competing interests to declare that are relevant to the content of this article.

Compliance with Ethical Standards.

Disclosure of potential conflicts of interest: none.
Research involving Human Participants and/or Animals: no.
Informed consent: not applicable.

## References

Badalič V (2024) Preventive Warfare: Hegemony, Power, and the Reconceptualization of War. Springer Nature

Barr K, Mintz A (2018) Public policy perspective on group decision-making dynamics in foreign policy. Policy Studies Journal 46:S69–S90

Barr K, Mintz A (2022) Groupthink, polythink, and con-div: Identifying group decision-making dynamics. In: Routledge Handbook of Foreign Policy Analysis Methods, Routledge, pp 269–287

Bhuyan BP, Ramdane-Cherif A, Tomar R, Singh T (2024) Neuro-symbolic artificial intelligence: a survey. Neural Computing and Applications 36(21):12809–12844

van den Bosch K, Schoonderwoerd T, Blankendaal R, Neerincx M (2019) Six challenges for human-AI co-learning. In: International Conference on Human-Computer Interaction, Springer, pp 572–589

Brooks R (2017) The seven deadly sins of AI predictions. MIT Technology review, https://www.technologyreview.com/2017/10/06/241837/the-seven-deadly-sins-of-ai-predictions/

Bubeck S, Chandrasekaran V, Eldan R, Gehrke J, Horvitz E, Kamar E, Lee P, Lee YT, Li Y, Lundberg S, et al. (2023) Sparks of artificial general intelligence: Early experiments with GPT-4. arXiv preprint arXiv:230312712

Chengeta T (2016) Defining the emerging notion of meaningful human control in weapon systems. NYUJ Int'l L & Pol 49:833

D'Amour A, Heller K, Moldovan D, Adlam B, Alipanahi B, Beutel A, Chen C, Deaton J, Eisenstein J, Hoffman MD, et al. (2022) Underspecification presents challenges for credibility in modern machine learning. Journal of Machine Learning Research 23(226):1–61

Davidovic J (2025) Rethinking human roles in AI warfare. Nature Machine Intelligence DOI 10.1038/s42256-025-01123-6

Diakopoulos N, Friedler S, Arenas M, Barocas S, Hay M, Howe B, Jagadish HV, Unsworth K, Sahuguet A, Venkatasubramanian S, Wilson C, Yu C, Zevenbergen B (2017) Principles for accountable algorithms and a social impact statement for algorithms. Online guideline document, URL `https://www.fatml.org/resources/principles-for-accountable-algorithms`, FAT/ML (Fairness, Accountability, and Transparency in Machine Learning)

Doshi-Velez F, Kim B (2017) Towards a rigorous science of interpretable machine learning. arXiv preprint arXiv:170208608

Dror IE (2020) Cognitive and human factors in expert decision making: six fallacies and the eight sources of bias. Analytical chemistry 92(12):7998–8004

Dwivedi R, Dave D, Naik H, Singhal S, Omer R, Patel P, Qian B, Wen Z, Shah T, Morgan G, et al. (2023) Explainable AI (XAI): Core ideas, techniques, and solutions. ACM computing surveys 55(9):1–33

Ekelhof M (2019) Moving beyond semantics on autonomous weapons: Meaningful human control in operation. Global Policy 10(3):343–348

Endsley MR (1995) Toward a theory of situation awareness in dynamic systems. Human Factors 37(1):32–64, DOI 10.1518/001872095779049543

Evans KD, Robbins SA, Bryson JJ (2025) Do we collaborate with what we design? Topics in Cognitive Science 17(2):392–411, DOI https://doi.org/10.1111/tops.12682

Freudenberg D (2010) Irreguläre Kräfte und der interessierte Dritte im modernen Kleinkrieg. In: Die Komplexität der Kriege, Springer, pp 179–187

Fuchs T (2024) Understanding sophia? on human interaction with artificial agents. Phenomenology and the Cognitive Sciences 23(1):21–42

Gao Q, Xu W, Shen M, Gao Z (2023) Agent teaming situation awareness (atsa): A situation awareness framework for human-ai teaming. arXiv preprint arXiv:230816785

Glüge S, Amirian M, Flumini D, Stadelmann T (2020) How (not) to measure bias in face recognition networks. In: IAPR Workshop on Artificial Neural Networks in Pattern Recognition, Springer, pp 125–137

Gunkel DJ, Wales JJ (2021) Debate: what is personhood in the age of ai? AI & society 36(2):473–486

Guo C, Pleiss G, Sun Y, Weinberger KQ (2017) On calibration of modern neural networks. In: International conference on machine learning, PMLR, pp 1321–1330

Hart PE, Nilsson NJ, Raphael B (1968) A formal basis for the heuristic determination of minimum cost paths. IEEE transactions on Systems Science and Cybernetics 4(2):100–107

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

Hendrycks D, Gimpel K (2017) A baseline for detecting misclassified and out-of-distribution examples in neural networks. In: International Conference on Learning Representations

Holewijn B (2011) Personnel recovery - a primer. Joint Air Power Competence Centre (JAPCC)

Hornik K, Stinchcombe M, White H (1989) Multilayer feedforward networks are universal approximators. Neural networks 2(5):359–366

Huang J, Xu Y, Wang Q, Wang QC, Liang X, Wang F, Zhang Z, Wei W, Zhang B, Huang L, et al. (2025) Foundation models and intelligent decision-making: Progress, challenges, and perspectives. The Innovation

Jonsson O, Käihkö I (2025) Non-Military Warfare: A War of Our Time. Taylor & Francis

Jordan MI, Mitchell TM (2015) Machine learning: Trends, perspectives, and prospects. Science 349(6245):255–260, DOI 10.1126/science.aaa8415

Kahneman D, Tversky A (1979) Prospect theory: An analysis of decision under risk. Econometrica 47(2):263–291

Kahneman D, Sibony O, Sunstein CR (2021) Noise: A flaw in human judgment. Hachette UK

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

Kambhampati S, Stechly K, Valmeekam K, Saldyt L, Bhambri S, Palod V, Gundawar A, Samineni SR, Kalwar D, Biswas U (2025) Stop anthropomorphizing intermediate tokens as reasoning/thinking traces! arXiv preprint arXiv:250409762

Kambouris ME (2024) Hybrid Warfare 2.2: Where Biothreats Meet Irregular Operations and Cyber Warriors in the 21st Century. Springer

Kaya M, Bilge HŞ (2019) Deep metric learning: A survey. Symmetry 11(9):1066

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

Koedijk M, Landman A, Bottenheft C, Fonken YM, Binsch O (2026) A virtual reality test to evaluate dismounted soldiers' cognitive and psychomotor performance in an operationally relevant setting. Frontiers in Psychology Volume 16 - 2025, DOI 10.3389/fpsyg.2025.1540936

Kontogianni T, Yue Y, Tang S, Schindler K (2024) Is continual learning ready for real-world challenges? arXiv preprint arXiv:240210130

Krizhevsky A, Sutskever I, Hinton GE (2012) ImageNet classification with deep convolutional neural networks. Advances in neural information processing systems 25

Kumar A, Clune J, Lehman J, Stanley KO (2025) Questioning representational optimism in deep learning: The fractured entangled representation hypothesis. arXiv preprint arXiv:250511581

LeCun Y, Bengio Y, Hinton G (2015) Deep learning. Nature 521(7553):436–444

Lee JD, See KA (2004) Trust in automation: Designing for appropriate reliance. Human factors 46(1):50–80

Lenat DB (1995) CYC: a large-scale investment in knowledge infrastructure. Commun ACM 38(11):33–38, DOI 10.1145/219717.219745

Liu M, Wei J, Liu Y, Davis J (2025) Human and ai perceptual differences in image classification errors. Proceedings of the AAAI Conference on Artificial Intelligence 39(13):14318–14326, DOI 10.1609/aaai.v39i13.33568

Lyons JB, Sycara K, Lewis M, Capiola A (2021) Human–autonomy teaming: Definitions, debates, and directions. Frontiers in Psychology Volume 12 - 2021, DOI 10.3389/fpsyg.2021.589585

von der Malsburg C, Stadelmann T, Grewe BF (2022) A theory of natural intelligence. arXiv preprint arXiv:220500002

Marcus G (2018) Deep learning: A critical appraisal. arXiv preprint arXiv:180100631

Marquis SL (1997) Unconventional warfare: Rebuilding US Special Operations Forces

Mattioli M, Cinà AE, Pelillo M (2024) Understanding xai through the philosopher's lens: A historical perspective. arXiv preprint arXiv:240718782

McCarthy J, Minsky ML, Rochester N, Shannon CE (1955) A proposal for the Dartmouth summer research project on artificial intelligence. Research proposal

Mecacci G, Santoni de Sio F (2020) Meaningful human control as reason-responsiveness: the case of dual-mode vehicles. Ethics and Information Technology 22(2):103–115

Meerveld HW, Lindelauf R, Postma EO, Postma M (2023) The irresponsibility of not using AI in the military. Ethics and Information Technology 25(1):14

Merkt PH, Ziehr S, Voigt T, Bickelmayer J, Toursarkissian M (2025) Bedarf und Konzept eines einheitlichen Curriculums für den Master Medic. Taktik+Medizin

Merriam-Webster (2021) Artificial intelligence. URL `https://web.archive.org/web/20210417130256/https://www.merriam-webster.com/dictionary/artificial%20intelligence`, archived April 17, 2021.

Merritt SM, Ilgen DR (2008) Not all trust is created equal: Dispositional and history-based trust in human-automation interactions. Human factors 50(2):194–210

Miller T (2019) Explanation in artificial intelligence: Insights from the social sciences. Artificial intelligence 267:1–38

Mitchell TM (1997) Machine learning. McGraw-Hill New York

Mussi M, Metelli AM, Restelli M, Losapio G, Bessa RJ, Boos D, Borst C, Leto G, Castagna A, Chavarriaga R, Dias D, Egli A, Eisenegger A, El Manyari Y, Fuxjäger A, Geraldes J, Hamouche S, Hassouna M, Lemetayer B, Leyli-Abadi M, Liessner R, Lundberg J, Marot A, Meddeb M, Schiaffonati V, Schneider M, Stadelmann T, Usher J, Van Hoof H, Viebahn J, Waefler T, Zanotti G (2025) Human-AI interaction in safety-critical network infrastructures. iScience p 113400, DOI https://doi.org/10.1016/j.isci.2025.113400

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

National Academies of Sciences, Engineering, and Medicine (2021) Human-AI teaming: State-of-the-art and research needs. The National Academies Press, Washington DC DOI 10.17226/26355

Neururer D, Dellwo V, Stadelmann T (2024) Deep neural networks for automatic speaker recognition do not learn supra-segmental temporal features.

Pattern Recognition Letters 181:64–69

Nguyen VL, Shaker MH, Hüllermeier E (2022) How to measure uncertainty in uncertainty sampling for active learning. Machine Learning 111(1):89–122

Nilsson NJ (2009) The quest for artificial intelligence. Cambridge University Press

O'Hara K (2020) Explainable ai and the philosophy and practice of explanation. Computer Law & Security Review 39:105474, DOI https://doi.org/10.1016/j.clsr.2020.105474

Ohlhorst D, Schön S (2015) Constellation analysis as a means of interdisciplinary innovation research–theory formation from the bottom up. Historical Social Research/Historische Sozialforschung pp 258–278

Ovadia Y, Fertig E, Ren J, Nado Z, Sculley D, Nowozin S, Dillon J, Lakshminarayanan B, Snoek J (2019) Can you trust your model's uncertainty? evaluating predictive uncertainty under dataset shift. Advances in neural information processing systems 32

O'Neill T, McNeese N, Barron A, Schelble B (2022) Human–autonomy teaming: A review and analysis of the empirical literature. Human factors 64(5):904–938

Parasuraman R, Manzey DH (2010) Complacency and bias in human use of automation: An attentional integration. Human factors 52(3):381–410

Parasuraman R, Riley V (1997) Humans and automation: Use, misuse, disuse, abuse. Human factors 39(2):230–253

Placani A (2024) Anthropomorphism in AI: hype and fallacy. AI and Ethics 4(3):691–698

Prince SJ (2023) Understanding deep learning. MIT press

Purushwalkam S, Morgado P, Gupta A (2022) The challenges of continuous self-supervised learning. In: European conference on computer vision, Springer, pp 702–721

Radford A, Wu J, Child R, Luan D, Amodei D, Sutskever I, et al. (2019) Language models are unsupervised multitask learners. OpenAI blog 1(8):9

Roost D, Meier R, Huschauer S, Nygren E, Egli A, Weiler A, Stadelmann T (2020) Improving sample efficiency and multi-agent communication in RL-based train rescheduling. In: 2020 7th Swiss Conference on Data Science (SDS), IEEE, pp 63–64

Rudin C (2019) Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. Nature machine intelligence 1(5):206–215

Ruiz DC, Sell J (2024) Fine-tuning and evaluating open-source large language models for the army domain. arXiv preprint arXiv:241020297

Russell SJ, Norvig P (2022) Artificial intelligence: a modern approach, $4^{th}$ edition. Pearson

Sager PJ, Deriu JM, Grewe BF, Stadelmann T, von der Malsburg C (2025a) The cooperative network architecture: Learning structured networks as representation of sensory patterns. arXiv preprint arXiv:240705650

Sager PJ, Meyer B, Yan P, von Wartburg-Kottler R, Etaiwi L, Enayati A, Nobel G, Abdulkadir A, Grewe BF, Stadelmann T (2025b) A comprehensive survey of agents for computer use: Foundations, challenges, and future directions. arXiv preprint arXiv:250116150

Schmidhuber J (2015) Deep learning in neural networks: An overview. Neural networks 61:85–117

Seeber I, Bittner E, Briggs RO, de Vreede T, de Vreede GJ, Elkins A, Maier R, Merz AB, Oeste-Reiß S, Randrup N, Schwabe G, Söllner M (2020) Machines as teammates: A research agenda on ai in team collaboration. Information & Management 57(2):103174, DOI https://doi.org/10.1016/j.im.2019.103174

Segessenmann J, Stadelmann T, Davison A, Dürr O (2025) Assessing deep learning: a work program for the humanities in the age of artificial intelligence. AI and Ethics 5(1):1–32

Selvaraju RR, Cogswell M, Das A, Vedantam R, Parikh D, Batra D (2017) Grad-cam: Visual explanations from deep networks via gradient-based localization. In: Proceedings of the IEEE international conference on computer vision, pp 618–626

Shneiderman B (2022) Human-centered AI. Oxford University Press

Silver D, Sutton RS (2025) Welcome to the era of experience. In: Designing an Intelligence, MIT Press

Santoni de Sio F, Van den Hoven J (2018) Meaningful human control over autonomous systems: A philosophical account. Frontiers in Robotics and AI 5:323836

Souchon L (2020) Strategy in the 21st Century: The Continuing Relevance of Carl von Clausewitz. Springer Cham

Stadelmann T (2025a) Debate: Evidence-based AI risk assessment for public policy. Public Money & Management

Stadelmann T (2025b) A guide to AI. Global Resilience White Papers

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

Stadelmann T, Amirian M, Arabaci I, Arnold M, Duivesteijn GF, Elezi I, Geiger M, Lörwald S, Meier BB, Rombach K, et al. (2018) Deep learning in the wild. In: IAPR Workshop on Artificial Neural Networks in Pattern Recognition, Springer, pp 17–38

Stadelmann T, Tolkachev V, Sick B, Stampfli J, Dürr O (2019) Beyond ImageNet: deep learning in industrial practice. In: Applied data science: lessons learned for the data-driven business, Springer, pp 205–232

Stadelmann T, Klamt T, Merkt PH (2022) Data centrism and the core of data science as a scientific discipline. Archives of Data Science, Series A 8(2)

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

Tang Y, Bi J, Xu S, Song L, Liang S, Wang T, Zhang D, An J, Lin J, Zhu R, Vosoughi A, Huang C, Zhang Z, Liu P, Feng M, Zheng F, Zhang J, Luo P, Luo J, Xu C (2025) Video understanding with large language models: A survey. IEEE Transactions on Circuits and Systems for Video Technology pp 1–1, DOI 10.1109/TCSVT.2025.3566695

Tian K, Mitchell E, Zhou A, Sharma A, Rafailov R, Yao H, Finn C, Manning CD (2023) Just ask for calibration: Strategies for eliciting calibrated confidence scores from language models fine-tuned with human feedback. In: Proceedings of EMNLP

Tigard DW (2025) On bullshit, large language models, and the need to curb your enthusiasm. AI and Ethics pp 1–11

Tuggener L, Amirian M, Rombach K, Lörwald S, Varlet A, Westermann C, Stadelmann T (2019) Automated machine learning in practice: state of the art and recent results. In: 2019 6th Swiss Conference on Data Science (SDS), IEEE, pp 31–36

Tuggener L, Emberger R, Ghosh A, Sager P, Satyawan YP, Montoya J, Goldschagg S, Seibold F, Gut U, Ackermann P, et al. (2024) Real world music object recognition. Transactions of the International Society for Music Information Retrieval 7(1):1–14

US Department of Defense (2008) DD Form 1833: Isolated Personnel Report (ISOPREP). Department of Defense Form, issued by the U.S. DoD; earliest public form copy dated May 2008

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

Veluwenkamp H (2022) Reasons for meaningful human control. Ethics and Information Technology 24(4):51

Verwimp E, Aljundi R, Ben-David S, Bethge M, Cossu A, Gepperth A, Hayes TL, Hüllermeier E, Kanan C, Kudithipudi D, Lampert CH, Mundt M, Pascanu R, Popescu A, Tolias AS, van de Weijer J, Liu B, Lomonaco V, Tuytelaars T, van de Ven GM (2024) Continual learning: Applications and the road forward. Transactions on Machine Learning Research

Waefler T, Hamouche S, Eisenegger A (2025) The Supportive AI framework: From recommending to supporting. In: Schmorrow DD, Fidopiastis CM (eds) Augmented Cognition, Springer Nature Switzerland, Cham, pp 303–317

Wang H, Yeung DY (2020) A survey on Bayesian deep learning. ACM computing surveys (csur) 53(5):1–37

Wang L, Zhang X, Su H, Zhu J (2024) A comprehensive survey of continual learning: Theory, method and application. IEEE transactions on pattern analysis and machine intelligence 46(8):5362–5383

Zanzotto FM (2019) Human-in-the-loop artificial intelligence. Journal of Artificial Intelligence Research 64:243–252

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

Ziehr S, Merkt PH (2024) Strategic resilience in human performance in the context of science and education - perspective. Frontiers in Psychiatry Volume 15 - 2024, DOI 10.3389/fpsyt.2024.1410296

Žigulić N, Glučina M, Lorencin I, Matika D (2024) Military decision-making process enhanced by image detection. Information 15(1), DOI 10.3390/info15010011