ESANN 2026 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 22-24 April 2026, i6doc.com publ., ISBN 9782875870964. Available from http://www.i6doc.com/en/.

Beyond Performance: Comprehensive Evaluation Strategies for Impactful Machine Learning

Valerie Vaquet[1] and Ulrike Kuhl[1] and Saša Brdnik[2] and Benjamin Paaßen[1] *

1- Faculty of Technology - Bielefeld University
Inspiration 1, 33619 Bielefeld - Germany

2- Faculty of Electrical Engineering and Computer Science - University of Maribor
Koroška cesta 46, 2000 Maribor - Slovenia

Abstract. Evaluation is an integral part of developing machine learning and AI-based systems for real-world applications. Given the transformative changes induced by ML/AI systems like large language models, evaluation needs to go beyond performance and include robustness, fairness, user perception, and legal compliance to ensure responsible usage. Further, evaluation practices need to also consider the full range of applications beyond the classic batch setting of machine learning, i.e., data streams, recommender systems, reinforcement learning, and foundation models. This paper provides an analysis of current evaluation practices and gaps across settings and dimensions, and argues for holistic, reproducible evaluation beyond benchmark performance.

Recently, machine learning (ML) and artificial intelligence (AI) research has been grappling with an evaluation paradox: while ML systems, especially large language models (LLMs), appear to perform ever better in benchmarks, outperforming humans in many cases, this performance does not translate to success of deployed ML/AI systems in the real world, with 95% of AI projects in industry failing to provide meaningful return on investment [1, 2, 3]. This highlights the need to rethink evaluation to keep pace with current developments. First, the appearance of foundation models, including LLMs, poses new challenges as foundation models are intended for, and hence need to be evaluated on, many different tasks at the same time [1, 3]. Being trained on huge amounts of data from the internet, data leakage between test benchmarks and training data becomes a considerable risk [4]. Second, we observe an increase in real-world applications beyond batch machine learning, inducing additional challenges, like noisy data and dynamic environments, which need to be reflected in the evaluation process [5]. Third, when applying ML/AI systems in settings that affect human users, there is a need for evaluation beyond performance: Considering robustness and safety, fairness, explainability, privacy, and legal aspects in addition to performance is paramount to ensure performant and accountable application of ML/AI systems [6].

In this tutorial paper, we will analyze the state of evaluation along two axes (see Table 1), namely evaluation dimensions and settings. In terms of

*VV, UK and BP gratefully acknowledge funding for the project KI-Akademie OWL, financed by the Federal Ministry of Research, Technology and Space (BMFTR) and supported by the VDI/VDE Innovation + Technik GmbH under grant no. 16IS24057A. SB gratefully acknowledges financial support from the Slovenian Research Agency (Research Core Funding No. P2-0057).

ESANN 2026 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 22-24 April 2026, i6doc.com publ., ISBN 9782875870964. Available from http://www.i6doc.com/en/.

	Performance	Robustness	Fairness & Biases	User Perception	Compliance
A Batch Setup	●Established evaluation practices with theory	●●Empirical adversarial robustness widely explored, with frameworks being available [7]; provable adversarial robustness infeasible for state-of-the-art networks [8]	●Multiple definitions and benchmarks with evaluation practices [9]; challenges: choosing the definition and evaluating actual impact on affected populations [10]	●Well studied [11, 12]; need for better human studies still present; currently wide use proxy metrics [12, 13, 14]; synthetic benchmarks exist	●Emerging system-level pipelines [15], technical standards [6], documentation artifacts [16, 17, 18, 19],and post-hoc audit approaches [20] under development
B Data Streams	●●Established evaluation schemes; lack of realistic benchmarks [21]	●●Established evaluation schemes; lack of realistic benchmarks [21]	●Few studies under drift [22]; almost no benchmarks nor practical evaluation standards; some theoretical results for dynamical systems [23]	●No benchmarks, few metrics [24, 25]; recognized research gap	●Continuous audit pipelines [26], auditable MLOps/logging approaches [27, 28, 29]; system-level compliance practices (row **A**) may apply
C Recommender Systems	●Established evaluation practices	●No agreed-upon evaluation scheme and benchmarks [30]	●Established metrics and benchmarks, still difficulties in realistic, long-term evaluations [31]	●A/B tests, faithfulness not evaluated, subjective, weak grounding in model behaviour [32]	●●Rare / domain-specific operationalizations of user-impact dimensions [33]; system-level compliance practices (row **A**) and continuous approaches (row **B**) may apply
D Reinforcement Learning	●Established evaluation practices	●No agreed-upon evaluation scheme and benchmarks [5]	●Clear concepts and difficulties in practical implementation [34]	●Indirect evaluation, few metrics, and few user studies exist [12, 25]	●Static documentation/ audits insufficient; standardized protocols for auditing policies in interaction missing; system-level compliance practices may apply (row **A**)
E Foundation Models and LLMs	●Potential data leakage [4]; specific benchmarks insufficient to evaluate general performance, how to evaluate hallucinations and factual errors? [1, 3]	●Many robustness issues: prompt stability, jailbreaking, prompt injection, distribution shift, with only limited benchmarks [1, 3]	●Wide range of metrics for stereotypes in models, but no consensus [35, 36]	●●No ground truth exists so evaluations are human preference-focused; emerging direction [25, 37]	●Multi-level audits and benchmarkable compliance criteria emerging [38, 39]; system-level compliance practices (row **A**) may apply

Table 1: Evaluation practices, challenges, and maturity per aspects. Maturity levels indicate the current state of evaluation practice, not intrinsic method quality (●- mature, ●- moderate, ●- underdeveloped).

ESANN 2026 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 22-24 April 2026, i6doc.com publ., ISBN 9782875870964. Available from http://www.i6doc.com/en/.

dimensions, we consider performance, compliance with regulatory frameworks (as discussed in Section 4), user perception (Section 3), fairness (Section 2), and robustness (Section 1). In terms of settings, next to the traditional batch setup, we consider all these criteria for non-stationary (streams), foundation models and LLMs, recommender systems, and reinforcement learning. Next to summarizing the current evaluation strategies, we identify weak spots and research gaps.

1 Robustness

Besides the accuracy of a model, its robustness is a key requirement when applying ML models in real-world settings. Moving from carefully curated datasets or simulated data to real-world applications, we have to consider more noisy data due to the nature of the applications, a potential sim2real gap, or imperfect sensors and measurement procedures [5, 30]. In case of minor variation in the observed data, we still require our models to work reliably and behave as expected to ensure their safe and compliant usage. In other words, we require models to be robust [5]. However, most ML models are not robust by design. In particular, many deep neural network-based approaches are vulnerable to *adversarial attacks* [40], meaning that a slight perturbation in the input, which is usually bounded by ϵ with respect to some norm ℓ, results in a change of the model output, e.g., in the prediction of a different class. Thereby, the changes are sufficiently subtle, such that they are often overlooked by humans.

To empirically evaluate the adversarial robustness of a ML model, in a first step, the threat model needs to be defined. This includes the goals of the adversary, their capabilities, and their knowledge about the attacked ML system. Based on that threat model, corresponding attacks are attempted to break the ML model and thereby evaluate its robustness. Using adaptive adversaries, i.e., considering a dynamic setup where the adversary can adapt to changes implemented in the model, is crucial to get a good estimate of the real-world adversarial robustness [7]. Evaluation in the batch setting frequently benefits from reproducibility (including publishing of source code and pre-trained models, alongside the usage of several frameworks). [7] provide a practical guide on evaluating adversarial robustness and common evaluation pitfalls. An alternative regime to empirical evaluation is so-called formal verification [41, 8]. Adversarial robustness (and some other properties, e.g., fairness metrics) can be formally proved or disproved (and thereby be positively or negatively evaluated). Originating from traditional program verification, formal verification relies on mathematical methods. While it provides formal guarantees, it has a much higher computational cost than empirical evaluation. Thus, the evaluation of large state-of-the-art artificial neural networks is currently infeasible [8].

Evaluating LLMs introduces a much broader landscape of what is meant by robustness: For instance, we would like the models to be robust with respect to minor modifications in the prompt (e.g., to rephrasing, usage of synonyms, or different input formats). Besides, the models should be robust to prompt injection and jailbreaking as well as distributional changes or drift. While the first benchmarks have been proposed to evaluate these robustness aspects, e.g.,

ESANN 2026 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 22-24 April 2026, i6doc.com publ., ISBN 9782875870964. Available from http://www.i6doc.com/en/.

with focus on adversarial robustness [1], more work is required. In particular, considering dynamic evaluation strategies and covering domain-specific scenarios is crucial to get an estimate of how robust LLMs are in real-world scenarios [1, 3].

Learning from non-stationary datastreams requires models that robustly adapt to distribution shifts. Thus, evaluating the performance of adaptive models under drift can be understood as a robustness evaluation. To better estimate how such models react to changes in real-world scenarios, there is a need for better datasets beyond the currently limited synthetic and real-world datasets [21].

The robustness of recommender systems is of particular interest to keep user experience, and, hence, business value, when faced with very noisy and malicious input. While many works explore more robust techniques, evaluation lags behind. In particular, it is done inconsistently with no agreed-upon benchmark datasets [30]. Similarly, robustness is a crucial aspect when it comes to reinforcement learning: When deploying RL-based systems in real-world settings, changes to the environment or other perturbations are to be accounted for. While an increasing body of work focusing on developing robust methods, again, the evaluation strategies are still very scattered, and for instance, lacking standard benchmark tasks and common baselines [5].

2 Fairness

Kicked off by high-profile scandals about apparent discrimination due to automatic decision-making systems in hiring, banking, and criminal justice, fairness research in machine learning has exploded over the last decade [9]. Most fairness research has focused on the batch classification setting with binary labels where the positive class is associated with a positive outcome (e.g. getting a job, getting a loan from a bank). In such a setting, a metric of (un-)fairness is the difference in rates of positive decisions between a protected group (e.g. women) and everyone else, which relates to the fairness definition of demographic parity [9]. Evaluating such a metric is relatively straightforward (provided that data about protected group membership is available) and can be performed via automatic tools, such as Aequitas [42]. Hence, such evaluation practices have already found their way into regulation, with the AI act providing a special provision to collecting data to evaluate bias (relating to group fairness) in ML/AI systems (Art. 10.5). However, different fairness definitions yield different fairness metrics, and definitions can contradict each other [9]. For example, equalized odds and equal opportunity consider classification errors instead of rates of positive decisions, which can be at odds with reducing demographic parity. Even worse, many fairness definitions are hard to evaluate in the first place. For example, individual fairness requires that similarly situated individuals are treated similarly, which can only be evaluated if an agreed-upon definition of similarity exists that is hard to construct in practice [43]. Causal fairness requires that protected attributes do not causally influence the prediction of the classifier, which can only be evaluated if a full causal model of the scenario is available, which is hard

ESANN 2026 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 22-24 April 2026, i6doc.com publ., ISBN 9782875870964. Available from http://www.i6doc.com/en/.

to obtain in practice [44]. For a full overview, we point to [9]. For more details on the workflow for fairness evaluation, refer to [45]. Further, Birhane has criticized fairness research for failing to take the perspective of affected, marginalized groups into account, who should arguably be the evaluators of whether they are treated fairly instead of automated metrics [10].

While most research on fairness in machine learning has focused on a batch, binary classification setting, other settings have also been studied. Fairness in recommender systems, in particular, has been a focus early on with the overarching goal of ensuring that the generated recommendations benefit everyone equally [43, 31]. Importantly, fairness does not only concern the users who receive recommendations but (perhaps more so) the content providers who have an interest in their content being recommended. The underlying definitions (and hence, evaluations) are often translated from the classification setup, with group fairness and individual fairness being the most prominent [31].

Fairness research regarding foundation models has mostly focused on stereotyped outputs, such as image generators generating images of white men in response to queries like "doctor" or large language models imputing stereotypical genders when translating sentences [36]. Evaluation metrics of fairness typically take into account the strength of stereotypical associations, either in the embedding space of the model or in the generated outputs, e.g. how strongly masculine versus feminine words are associated with certain occupations (like doctor). Still, the community has not agreed on a metric and all common metrics are imperfect operationalizations of the actual harm.

A notoriously challenging setting for fairness evaluation are dynamic setups. [46] provide a first framework for generating realistic inherently unfair, and drifting data streams. Beyond this, more work on drift-aware evaluation protocols is required [22]. For fairness, feedback loops are particularly important: For example, if a bank decides to grant a loan, this decision contributes to the applicant having enough resources to pay back the loan, introducing a self-fulfilling prophecy effect that has also been dubbed performative prediction [47, 23]. The most mature approaches to handle this direction of study is, perhaps, most advanced in reinforcement learning, where fairness can be defined as equal rewards across actors (and related notions) [34]. Still, the practical evaluation of fairness remains challenging as it requires long observations that representatively cover likely scenarios to evaluate the reward distribution across actors comprehensively. As such, many studies remain constrained to simulation.

3 User Perception

Human-centered evaluation of ML/AI systems covers a wide range of aspects, including interface design and usability, model-side factors, and interaction aspects between humans and models [48, 49]. Explainability has emerged as a prominent focus directly mediating between users and AI, and is audience-centered by design. An explanation only works if it makes an ML/AI system understandable in the context of the task to the people who must interpret or act on it [50, 14]. Good explanations are faithful to the underlying system, understandable to

ESANN 2026 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 22-24 April 2026, i6doc.com publ., ISBN 9782875870964. Available from http://www.i6doc.com/en/.

the intended audience, useful and actionable for the observed task, and stable enough to support reliable reasoning across similar cases [13, 50]. Therefore, explanation design and evaluation are artifact- and stakeholder-specific, meaning there is no one best explanation format that fits them all [51, 13]. To identify the best explanation for each use case, studies typically rely on user questionnaires (most common instruments summarized in [52, 14]). Interviews, observation of use, simulations, case studies, expert studies, expert evaluation, self-reports, and visual assessments have also been utilized in prior works [11]. The dimensions guiding the evaluation are not standardized and vary across studies. However, surveys show trustworthiness is most commonly evaluated [12, 52]. Other dimensions in focus are informativeness, user satisfaction, causality, transferability, confidence, fairness, accessibility, interactivity, and privacy awareness [12, 13, 50].

The main gap in research is the lack of standardized evaluation frameworks and metrics, leading to weak comparability across studies [25, 53, 52]. Some controlled evaluation frameworks that fix most study conditions while varying specific explanation factors to enable systematic comparison have been proposed [54], but such frameworks remain sparse and typically limited to narrow tasks or explanation types. This is compounded by the heterogeneity of XAI goals and domains, where the evaluation setup (target users, task, medium, and explanation format) strongly shapes what can be concluded [12]. The challenge is amplified in non-stationary settings when faced with concept drift and uncertainty, as explanations and trust calibration (i.e., aligning users' confidence and reliance with the system's actual reliability) can degrade over time [25]. Relying primarily on (standardized and validated) questionnaires enables comparability, but narrows data collection; adding interviews and observational methods would yield richer data [12]. There is no lack of validated instruments for popular goals (trust, informativeness, satisfaction), but fewer validated measures for less-studied goals [12, 53], alongside weaknesses in participant recruitment and sampling. Studies often draw from easily accessible samples (crowdsourcing platforms, general audiences, and students), leading to lower external validity [12] and weakening the results [53]. Participatory design and inclusion of impacted or vulnerable stakeholders, motivated by the power imbalance and accountability concerns, are still relevant [25]. Finally, the field is dominated by short, single-session evaluations; long-term or longitudinal evidence on whether explanations support calibrated use over time remains rare [12]. Intriguingly, the limited longitudinal studies that do exist reveal unexpected or even adverse effects of explanations [55]. For emerging LLM-based conversational explanations, evaluations remain dominated by exploratory user studies and questionnaires, with limited objective task-based protocols and fragmented evaluation procedures across systems [37].

User-evaluated XAI systems most often use deep learning and classification models [12]. Batch classification and regression models have the most mature evaluation toolbox (Table 1) of proxy metrics and tests, but key issues remain: mixing plausibility with faithfulness and using proxy scores to probe human usefulness [53, 56]. For foundation models and LLMs, evaluations are less mature, as

ESANN 2026 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 22-24 April 2026, i6doc.com publ., ISBN 9782875870964. Available from http://www.i6doc.com/en/.

explanations are generally post-hoc and lack ground truth, so studies frequently build on qualitative assessment of human preference judgments, rather than faithfulness protocols [25, 37]. For streaming or non-stationary classification, evaluations must address the temporal reliability of explanations, yet drift-aware benchmarks and standardized protocols remain sparse [24, 25]. In recommender systems, evaluation is largely user-centered (focused on transparency, trust, effectiveness, efficiency, and satisfaction) with some A/B tests [32]. The studies frequently measure explainability coverage (e.g., fraction of explainable items, mean explainability precision and recall, feature matching/coverage/diversity) rather than faithfulness or helpfulness of explanations in user decision-making. Key gaps are non-standardized measures for responsiveness and cognitive load, and fragmented HCI taxonomies [32]. Finally, in reinforcement learning, explanations (commonly counterfactual or causal) are typically evaluated indirectly, through task performance of user perception with few standardized metrics, and recurring reproducibility and comparability concerns [12, 25].

User evaluation of explanations remains methodologically fragmented. Studies frequently rely on proxy metrics and questionnaires, rarely validate results longitudinally or under realistic deployment conditions (e.g., concept drift), and lack a universal evaluation protocol. We therefore argue against single-metric evaluation of explanations and recommend a layered, reproducible strategy that aligns targets with stakeholder goals and artifacts, combines proxy checks with task-based mixed-method user studies, and standardizes reporting to enable replication and cross-study comparison [14, 25, 57].

4 Compliance

The decisive question regarding the evaluation of ML/AI systems in terms of regulatory compliance is whether it can be deployed defensibly under the applicable rules, and whether evidence for compliance can be produced on demand. For high-risk systems under the EU AI Act,[1,2] this largely translates to producing technical documentation and demonstrating adherence to relevant technical standards (e.g., ISO/IEC standards [6]), including documentation of system design and overall algorithmic logic [58]. Importantly, legal regulations seldom mandate the trustworthy AI approaches of the machine learning literature (such as explainability, fairness, and transparency). Rather, regulations require providing appropriate information/documentation to users, and putting in place suitable oversight measures [59], such as assessments, audits, disclosures, inventories, and red teaming [60]. Consequently, compliance-oriented evaluation emphasizes auditable artifacts and life-cycle evidence rather than ML metrics.

Importantly, compliance refers to the entire ML/AI system (in a certain application context) and is typically not evaluated for a machine learning model in

[1]EU AI Act: https://artificialintelligenceact.eu/the-act, accessed on February 4th, 2026.

[2]Beyond the EU AI Act, global AI governance is shaped by a mix of treaty-level instruments (e.g., the Council of Europe Framework Convention on AI), voluntary governmental risk frameworks (e.g., the NIST AI Risk Management Framework), and international principles (e.g., the OECD AI Principles, the UNESCO Recommendation on the Ethics of AI).

ESANN 2026 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 22-24 April 2026, i6doc.com publ., ISBN 9782875870964. Available from http://www.i6doc.com/en/.

isolation. A requirements-to-metrics perspective treats compliance as a translation problem from legal obligations to testable predictive and compliance proxies and supports trade-off mapping for auditable model selection [15]. To address frequent system updates (e.g., drift and continuously updated recommenders), Continuous Audit-Based Certification (CABC) [26] shifts from point-in-time audits to automated, continuous auditing against predefined quality requirements, using MLOps artifacts as evidence inputs and enabling renewal/revocation decisions. Lifecycle audit catalogs (e.g., TÜV AUSTRIA Trusted AI framework [61], ML auditing core criteria catalog [62]) structure evaluation as a system-wide evaluation across security, functionality, and governance. Where transparency conflicts with confidentiality (e.g., proprietary models or sensitive data), ZKMLOps [63] proposes cryptographic compliance evidence via Zero-Knowledge Proofs integrated into the MLOps lifecycle. Domain-specific pipelines operationalize fairness, privacy, and oversight in their particular area (e.g., education) [33]. For LLMs, required auditing pipelines are inherently multi-level, distributing compliance requirements across provider-, model-, and application-level components [38]. Finally, data-quality checklists [64] complement end-to-end pipelines by structuring compliance-relevant judgments, but often leave concrete measurements and acceptance criteria to domain experts, which limits comparability across use cases.

A recurring practical issue is that compliance requires evidence artifacts, yet documentation is often sparse in practice, making auditing difficult [29]. Proposed artifacts range from documentation templates [16, 17, 18] and structured metadata [19] to tamper-resistant logging mechanisms [27, 28] and audit reports that summarize compliance-relevant evidence for human validation [20]. Thus, in practice, batch classification settings may operationalize compliance via dataset and model documentation and post-hoc audits, whereas recommender- and streaming settings depend more on continuous monitoring logs and change management. Reinforcement learning systems additionally require documenting the system's operating constraints and oversight in deployment since behavior emerges through interaction.

Recurring gaps in the literature include what constitutes "enough evidence" for a sufficient compliance claim beyond checklists, and the lack of systematic processes for translating legal texts into technical requirements. For LLMs, early regulation-oriented suites exist (COMPL-AI[3] [39]), but remain rare. Further challenges include strong domain dependence of operationalizations and—despite efforts such as CABC [26]—standardizing methods for dealing with dynamic model changes. Finally, when human oversight is needed, it is not clear how to integrate it effectively, and how to decide upfront when the human-in-the-loop is required rather than optional.

[3] https://compl-ai.org

ESANN 2026 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 22-24 April 2026, i6doc.com publ., ISBN 9782875870964. Available from http://www.i6doc.com/en/.

5 Conclusion

In this paper, we analyzed the state of evaluation across evaluation settings and dimensions. We found that evaluation in the batch setting is very advanced across the different dimensions, with only compliance lagging behind. Similarly, evaluation of performance is mature across most considered settings (with the notable exception of foundation models). However, evaluation beyond performance and the batch setting is still very preliminary.

Moving forward, holistic evaluation requires a shift from scalar scores to entire evidence packages, including user and field studies, considering multiple dimensions and providing sufficient documentation. Following this path, trade-offs between different scores and criteria need to be accounted for. We see immense opportunities in realizing such evaluation processes by relying on interdisciplinary efforts, including experts from psychology, social sciences and humanities, IT security, and law. Finally, next to the need for a life-cycle view on evaluation, including continuous monitoring and longitudinal evidence, increasing reproducibility by sharing protocols, benchmarks, and reporting templates is crucial to translate benchmark success into deployment and resolve the evaluation paradox.

References

[1] Sourav Banerjee, Ayushi Agarwal, and Eishkaran Singh. The Vulnerability of Language Model Benchmarks: Do They Accurately Reflect True LLM Performance? 2024.
[2] Aditya Challapally, Chris Pease, Ramesh Raskar, et al. *The GenAI Divide – State of AI in Business 2025*. Tech. rep. MIT NANDA, 2025.
[3] Zishan Guo, Renren Jin, Chuang Liu, et al. *Evaluating Large Language Models: A Comprehensive Survey*. 2023.
[4] Sayash Kapoor and Arvind Narayanan. Leakage and the reproducibility crisis in machine-learning-based science. *Patterns* 4(9) 2023.
[5] Janosch Moos, Kay Hansel, Hany Abdulsamad, et al. Robust Reinforcement Learning: A Review of Foundations and Recent Advances. en. *Machine Learning and Knowledge Extraction* 4(1) 2022, pp. 276–315.
[6] Jesus Oviedo, Moises Rodriguez, Andrea Trenta, et al. ISO/IEC quality standards for AI engineering. *Computer Science Review* 54 2024, p. 100681.
[7] Nicholas Carlini, Anish Athalye, Nicolas Papernot, et al. *On Evaluating Adversarial Robustness*. arXiv:1902.06705. 2019.
[8] Mark Huasong Meng, Guangdong Bai, Sin Gee Teo, et al. Adversarial Robustness of Deep Neural Networks: A Survey from a Formal Verification Perspective. *IEEE Transactions on Dependable and Secure Computing* 2024, pp. 1–1.
[9] Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, et al. A Survey on Bias and Fairness in Machine Learning. *ACM Comput. Surv.* 54(6) 2021.
[10] Abeba Birhane, Elayne Ruane, Thomas Laurent, et al. The Forgotten Margins of AI Ethics. *Proceedings of the FAccT*. FAccT '22. Seoul, Republic of Korea, 2022, pp. 948–958.
[11] Saša Brdnik and Boštjan Šumak. Current Trends, Challenges and Techniques in XAI Field; A Tertiary Study of XAI Research. *2024 47th MIPRO ICT and Electronics Convention (MIPRO)*. 2024, pp. 2032–2038.
[12] Noor Al-Ansari, Dena Al-Thani, and Reem S. Al-Mansoori. User-Centered Evaluation of Explainable Artificial Intelligence (XAI): A Systematic Literature Review. *Human Behavior and Emerging Technologies* 2024(1) 2024, p. 4628855.
[13] Robert R. Hoffman, Shane T. Mueller, Gary Klein, et al. Measures for explainable AI: Explanation goodness, user satisfaction, mental models, curiosity, trust, and human-AI performance. *Frontiers in Computer Science* 5 2023.
[14] Robert R. Hoffman, Shane T. Mueller, Gary Klein, et al. *Metrics for Explainable AI: Challenges and Prospects*. 2019.
[15] Mathias Hanson, Gregory Lewkowicz, and Sam Verboven. Engineering the law-machine learning translation problem: Developing legally aligned models. *Computer Law & Security Review* 60 2026, p. 106252.
[16] Margaret Mitchell, Simone Wu, Andrew Zaldivar, et al. Model cards for model reporting. *Proceedings of the conference on fairness, accountability, and transparency*. 2019, pp. 220–229.
[17] Matthew Arnold, Rachel KE Bellamy, Michael Hind, et al. FactSheets: Increasing trust in AI services through supplier's declarations of conformity. *IBM Journal of Research and Development* 63(4/5) 2019, pp. 6–1.
[18] Laura Lucaj, Alex Loosley, Håkan Jonsson, et al. TechOps: Technical Documentation Templates for the AI Act. *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*. Vol. 8. 2. 2025, pp. 1647–1660.
[19] Bill Marino, Preslav Aleksandrov, Carwyn Rahman2 Yulu Pi, et al. Compliance Cards: Computational Artifacts for Automated AI Regulation Compliance. *arXiv preprint arXiv:2406.14758* 2024.
[20] Sahaj Tushar Gandhi. AI-Driven Compliance Audits: Enhancing Regulatory Adherence in Financial and—Legal Sectors. *International Journal of Advanced Research in Computer Science & Technology (IJARCST)* 6(5) 2023, pp. 8981–8988.
[21] Fabian Hinder, Johannes Brinkrolf, Kathrin Lammers, et al. Real vs. Virtual Drift: Creating Realistic Stream Learning Benchmarks. *ESANN 2026* (Bruges, Belgium). 2026.
[22] Kathrin Lammers, Valerie Vaquet, and Barbara Hammer. Continuous Fair SMOTE – Fairness-Aware Stream Learning from Imbalanced Data. *Proceedings of the ICANN*. 2026, pp. 336–348.
[23] Benjamin Paaßen, Astrid Bunge, Carolin Hainke, et al. Dynamic fairness – Breaking vicious cycles in automatic decision making. *Proceedings of the 27th European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning (ESANN 2019)* (Bruges, Belgium). 2019, pp. 477–482.

ESANN 2026 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 22-24 April 2026, i6doc.com publ., ISBN 9782875870964. Available from http://www.i6doc.com/en/.

[24] Daniele Pelosi, Diletta Cacciagrano, and Marco Piangerelli. Explainability and Interpretability in Concept and Data Drift: A Systematic Literature Review. *Algorithms* 18(7) 2025.
[25] Luca Longo, Mario Brcic, Federico Cabitza, et al. Explainable Artificial Intelligence (XAI) 2.0: A manifesto of open challenges and interdisciplinary research directions. *Information Fusion* 106 2024, p. 102301.
[26] Dorian Knoblauch and Jürgen Großmann. Towards a risk-based continuous auditing-based certification for machine learning. *The Review of Socionetwork Strategies* 17(2) 2023, pp. 255–273.
[27] Krishna Mohan Kadambala. Auditable AI pipelines: Logging and verifiability in ML workflows. *Innovative Journal of Applied Science* 2025, pp. 35–35.
[28] Jimmy Joseph. Trust, but Verify: Audit-ready logging for clinical AI. *World Journal of Advanced Engineering Technology and Sciences* 2023, pp. 449–474.
[29] Patrick Loic Foalem, Leuson Da Silva, Foutse Khomh, et al. Logging requirement for continuous auditing of responsible machine learning-based applications. *Empirical Software Engineering* 30(3) 2025, p. 97.
[30] Kaike Zhang, Qi Cao, Fei Sun, et al. Robust Recommender System: A Survey and Future Directions. en. *ACM Computing Surveys* 58(1) 2026, pp. 1–38.
[31] Yifan Wang, Weizhi Ma, Min Zhang, et al. A Survey on the Fairness of Recommender Systems. *ACM Trans. Inf. Syst.* 41(3) 2023.
[32] Weiqing Li, Yue Xu, Yuefeng Li, et al. Display Content, Display Methods, and Evaluation Methods of the HCI in Explainable Recommender Systems: A Survey. *Big Data Mining and Analytics* 9(1) 2026, pp. 198–228.
[33] Mohammed Nazmul Islam Miah, Md Joshim Uddin, and Md Wasim Ahmed. Regulating Artificial Intelligence in Education: Analyzing Legal and Ethical Frameworks for the Deployment of AI and Machine Learning Models in US Educational Institutions. *Journal of Computer Science and Technology Studies* 7(11) 2025, pp. 387–404.
[34] Anka Reuel and Devin Ma. Fairness in Reinforcement Learning: A Survey. *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society* 7(1) 2024, pp. 1218–1230.
[35] Aida Kostikova, Zhipin Wang, Deidamea Bajri, et al. *LLLMs: A Data-Driven Survey of Evolving Research on Limitations of Large Language Models*. 2025.
[36] Isabel O. Gallegos, Ryan A. Rossi, Joe Barrow, et al. Bias and Fairness in Large Language Models: A Survey. *Computational Linguistics* 50(3) 2024, pp. 1097–1179.
[37] Dimitry Mindlin, Fabian Beer, Leonie Nora Sieger, et al. Beyond one-shot explanations: a systematic literature review of dialogue-based xAI approaches. *Artificial Intelligence Review* 58(3) 2025, p. 81.
[38] Jakob Mökander, Jonas Schuett, Hannah Rose Kirk, et al. Auditing large language models: a three-layered approach. *AI and Ethics* 4(4) 2024, pp. 1085–1115.
[39] Philipp Guldimann, Alexander Spiridonov, Robin Staab, et al. Compl-ai framework: A technical interpretation and llm benchmarking suite for the eu artificial intelligence act. *arXiv preprint arXiv:2410.07959* 2024.
[40] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, et al. *Intriguing properties of neural networks*. arXiv:1312.6199. 2014.
[41] Xuan Xie, Kristian Kersting, and Daniel Neider. *Neuro-Symbolic Verification of Deep Neural Networks*. 2022.
[42] Pedro Saleiro, Benedict Kuester, Loren Hinkson, et al. *Aequitas: A Bias and Fairness Audit Toolkit*. 2019.
[43] Cynthia Dwork, Moritz Hardt, Toniann Pitassi, et al. Fairness Through Awareness. *ITCS 2012*. Cambridge, Massachusetts, 2012, pp. 214–226.
[44] Niki Kilbertus, Mateo Rojas Carulla, Giambattista Parascandolo, et al. Avoiding Discrimination through Causal Reasoning. *NIPS 2017*. 2017, pp. 656–666.
[45] Zhenpeng Chen, Jie M. Zhang, Max Hort, et al. Fairness Testing: A Comprehensive Survey and Analysis of Trends. *ACM Trans. Softw. Eng. Methodol.* 33(5) 2024.
[46] Kathrin Lammers, Valerie Vaquet, Jonas Vaquet, et al. Realistic Benchmarks for Fair Stream Learning. *International Conference on Neural Information Processing*. Springer. 2025, pp. 169–184.
[47] Moritz Hardt and Celestine Mendler-Dünner. Performative prediction: Past and future. *Statistical Science* 40(3) 2025, pp. 417–436.
[48] Jan Philip Göpfert, Ulrike Kuhl, Lukas Hindemith, et al. Intuitiveness in Active Teaching. *IEEE Transactions on Human-Machine Systems* 52(3) 2021, pp. 458–467.
[49] Fabian Sperrle, Mennatallah El-Assady, Grace Guo, et al. A survey of human-centered evaluations in human-centered machine learning. *Computer Graphics Forum*. Vol. 40. 3. Wiley Online Library. 2021, pp. 543–568.
[50] Alejandro Barredo Arrieta, Natalia Díaz-Rodríguez, Javier Del Ser, et al. Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information Fusion* 58 2020, pp. 82–115.
[51] Annika Schuster and Florian J. Boge. How can we trust opaque systems? Criteria for robust explanations in XAI. *2025 International Joint Conference on Neural Networks (IJCNN)*. 2025, pp. 1–8.
[52] Yao Rong, Tobias Leemann, Thai-Trang Nguyen, et al. Towards Human-Centered Explainable AI: A Survey of User Studies for Model Explanations. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 46(4) 2024, pp. 2104–2122.
[53] Meike Nauta, Jan Trienes, Shreyasi Pathak, et al. From Anecdotal Evidence to Quantitative Evaluation Methods: A Systematic Review on Evaluating Explainable AI. *ACM Comput. Surv.* 55(13s) 2023.
[54] Ulrike Kuhl, André Artelt, and Barbara Hammer. Let's go to the Alien Zoo: Introducing an experimental framework to study usability of counterfactual explanations for machine learning. *Frontiers in Computer Science* 5 2023, p. 1087929.
[55] Ulrike Kuhl and Annika Bush. When Bias Backfires: The Modulatory Role of Counterfactual Explanations on the Adoption of Algorithmic Bias in XAI-Supported Human Decision-Making. *World Conference on Explainable Artificial Intelligence*. Springer. 2025, pp. 249–273.
[56] Finale Doshi-Velez and Been Kim. *Towards A Rigorous Science of Interpretable Machine Learning*. 2017.
[57] Gary Klein, Robert R Hoffman, William J Clancey, et al. "Minimum Necessary Rigor" in empirically evaluating human–AI work systems. *AI Magazine* 44(3) 2023, pp. 274–281.
[58] Francesco Sovrano, Salvatore Sapienza, Monica Palmirani, et al. Metrics, explainability and the European AI act proposal. *J* 5(1) 2022, pp. 126–138.
[59] Cecilia Panigutti, Ronan Hamon, Isabelle Hupont, et al. The role of explainable AI in the context of the AI Act. *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*. FAccT '23. Chicago, IL, USA: ACM, 2023, pp. 1139–1150.
[60] Mona Sloane and Elena Wüllhorst. A systematic review of regulatory strategies and transparency mandates in AI regulation in Europe, the United States, and Canada. *Data & Policy* 7 2025, e11.
[61] Kajetan Schweighofer, Barbara Brune, Lukas Gruber, et al. Safe and Certifiable AI Systems: Concepts, Challenges, and Lessons Learned. *arXiv preprint arXiv:2509.08852* 2025.
[62] Markus Schwarz, Ludwig Christian Hinske, Ulrich Mansmann, et al. Designing an ml auditing criteria catalog as starting point for the development of a framework. *IEEE access* 12 2024, pp. 39953–39967.
[63] Filippo Scaramuzza, Renato Cordeiro Ferreira, Tomaz Maia Suller, et al. " Show Me You Comply... Without Showing Me Anything": Zero-Knowledge Software Auditing for AI-Enabled Systems. *arXiv preprint arXiv:2510.26576* 2025.
[64] Daniel Schwabe, Katinka Becker, Martin Seyferth, et al. The METRIC-framework for assessing data quality for trustworthy AI in medicine: a systematic review. *NPJ digital medicine* 7(1) 2024, p. 203.