

From Post-Hoc Explanations to Assurance-Oriented XAI in Safety-Critical Automatic Target Recognition

Vanessa Buhrmester¹, David Muench¹ and Michael Arens¹

¹Fraunhofer IOSB, Fraunhofer Institute of Optronics, System Technologies and Image Exploitation, Ettlingen, Germany

Abstract

Automatic Target Recognition (ATR) systems increasingly rely on deep learning models for detection, classification, and tracking in complex sensor data. In safety-critical and security-sensitive settings, explainability is therefore closely linked to transparency, human oversight, trust calibration, and assurance. However, common post-hoc explanation methods, such as saliency maps, attention visualizations, or detector-specific heatmaps, provide only limited evidence about correctness, robustness, or causal validity. We take the position that post-hoc explanations should not be treated as evidence of trustworthy behavior unless they are embedded in an assurance process that evaluates their faithfulness, robustness, uncertainty context, and domain plausibility. We propose a conceptual XAI-enabled reference framework for organizing explanation outputs as assurance evidence in ATR. The framework separates model prediction from assurance status and connects explanation methods with uncertainty estimation, explanation quality tests, domain plausibility checks, human oversight, and continuous monitoring.

Keywords

Automatic Target Recognition, Explainable AI, AI Assurance, Safety-Critical AI, Human Oversight, Trust Calibration

1. Introduction

Automatic Target Recognition (ATR) systems detect, classify, track, and identify objects in sensor data such as electro-optical imagery, infrared data, radar, or synthetic aperture radar (SAR). Modern ATR increasingly relies on deep neural networks and transformer-based architectures, which can achieve high detection performance but often remain opaque.

In safety-critical and security-sensitive environments, this opacity is problematic. ATR systems may support decisions under uncertainty, incomplete sensor coverage, occlusion, adverse environmental conditions, or adversarial manipulation. In such contexts, explainability is not merely a tool for model debugging, but becomes closely related to trust calibration, human oversight, accountability, and assurance. Operators and auditors must understand not only what a system detects, but also why a detection was produced, how reliable the prediction is, and under which conditions the system may fail.

Although ATR systems are not generative models, they are autonomous perception and decision-support systems in which explanations must support oversight, monitoring, intervention, and accountability. ATR therefore provides a safety-critical case study for broader challenges also faced by agentic and autonomous AI systems.

Current explainable AI (XAI) methods provide useful diagnostic tools for analyzing deep learning models. Saliency maps, Grad-CAM variants, attention visualizations, surrogate models, concept-based explanations, and detection-aware explanation methods can reveal relevant regions, features, or object-level evidence. However, these explanations do not automatically demonstrate that a model decision is correct, robust, or causally grounded. Prior work has shown that visually plausible explanations may be unstable, insensitive to model parameters, or vulnerable to manipulation [1, 2, 3]. This creates a critical gap between explanation as a visual aid and explanation as evidence for safety-critical assurance. This gap reflects a broader shift in XAI research from isolated explanation techniques toward interdisciplinary, context-aware, and evaluation-oriented approaches to explainability [4].

XAI.it 2026 - 6th Workshop on Explainable AI, Generative and Agentic Systems: Trust, Transparency, and Human Oversight, co-located with AxiA 2026, October 6-9, 2026, Perugia, Italy



© 2026 Copyright © 2026 for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

This position paper argues that, in high-assurance ATR, XAI should not be treated as an isolated post-hoc visualization layer. Instead, explanations should be embedded into a structured assurance process that combines model evaluation, uncertainty estimation, robustness testing, domain plausibility assessment, human-centered presentation, logging, and continuous monitoring.

In this paper, we use assurance to denote justified confidence that an ATR output is sufficiently supported by available evidence for its intended operational use. Assurance is therefore not synonymous with model confidence or explanation quality. Rather, it results from considering multiple sources of evidence, including sensor quality, predictive uncertainty, explanation quality, and domain plausibility. The proposed framework does not constitute a safety case or certification process, but structures XAI-related evidence that may contribute to broader verification, validation, and assurance activities.

The contributions of this paper are threefold:

1. We analyze why post-hoc XAI methods are insufficient as standalone assurance evidence for safety-critical ATR systems.
2. We formulate assurance-oriented evaluation criteria for XAI in ATR, emphasizing interpretability, robustness, faithfulness, manipulation resistance, verification and validation suitability, real-time feasibility, uncertainty awareness, and domain plausibility.
3. We propose a testable reference framework for organizing XAI outputs as assurance evidence in ATR. Unlike conventional XAI use, the framework separates model prediction from assurance status by linking uncertainty estimation, explanation quality testing, domain plausibility checks, human oversight, and auditability.

2. Why Post-Hoc XAI Is Insufficient for Safety-Critical ATR

Post-hoc explanation methods are widely used because they can be applied to already trained black-box models without requiring fundamental architectural changes. In computer vision and ATR, prominent examples include gradient-based saliency, Integrated Gradients, Grad-CAM, detector-specific CAM variants, attention visualizations, LIME, SHAP, and concept-based approaches such as TCAV [5, 6, 7, 8, 9]. However, in high-stakes decision contexts, relying on post-hoc explanations for black-box models has been criticized because such explanations may provide only an approximation of model behavior rather than a transparent and reliable decision mechanism [10].

These methods can support model inspection and failure analysis. For instance, heatmaps may indicate whether an ATR system focuses on the target object, background clutter, sensor artifacts, or irrelevant contextual cues. Detection-aware explanation methods are particularly relevant because they link explanations to object-level predictions such as bounding boxes, detection scores, or detected instances [11, 12].

However, an explanation that is useful for human inspection is not necessarily reliable as assurance evidence. Three limitations are particularly important.

2.1. Spurious and Unfaithful Explanations

A visually plausible explanation may fail to reflect the actual decision mechanism of the model. Sanity checks have shown that some saliency methods can remain surprisingly similar even when model parameters are randomized [1]. This indicates that explanation maps may partly reflect input structure or visualization artifacts rather than learned model reasoning. In ATR, this is problematic because explanations may appear operationally meaningful while the model relies on spurious correlations, background patterns, or sensor-specific artifacts. For example, an ATR detector may assign the correct class to a vehicle while the explanation highlights road markings, shadows, clutter, or sensor artifacts rather than vehicle-specific evidence. Such cases show why visual plausibility alone is insufficient for assurance [1].

2.2. Instability and Manipulation

Safety-critical ATR systems operate under noise, occlusion, aspect-angle variation, sensor degradation, and domain shift. Explanations that change substantially under minor input perturbations are therefore difficult to use as reliable evidence. Moreover, post-hoc explanations may be manipulated or constructed in ways that appear plausible without faithfully representing model behavior [3]. Such instability is especially critical in adversarial or security-sensitive environments. In multisensor ATR, small changes in thermal contrast, viewing angle, radar speckle, or preprocessing may alter the explanation although the prediction remains unchanged. This makes stability and manipulation resistance essential for high-assurance ATR [3].

2.3. Overtrust and Human Oversight

XAI is often expected to support human oversight. However, visually persuasive explanations may also foster overtrust. Operators may interpret heatmaps or attention maps as causal evidence, even when the explanation method provides only correlational or heuristic information. This risk is amplified in detection scenarios because bounding boxes and object-level visualizations can create a strong impression of precision. In safety-critical ATR, XAI should therefore calibrate trust rather than simply increase it. For instance, a bounding box combined with a sharp heatmap may suggest object-level precision even if the attribution is only weakly related to the detector’s decision process. Such explanations may support situational awareness, but also reinforce automation bias and overreliance [2, 13].

3. Assurance-Oriented Evaluation Dimensions

For safety-critical ATR, explanation quality should be evaluated beyond visual plausibility. We use the following evaluation dimensions as a conceptual basis for assurance-oriented XAI. Table 1 summarizes these dimensions and relates them to safety-critical ATR requirements.

Table 1: Assurance-oriented evaluation dimensions for XAI in safety-critical ATR.

Dimension	Relevance for safety-critical ATR
Stakeholder-specific interpretability	Provides task-appropriate information for operators, domain experts, engineers, and auditors rather than assuming a single notion of interpretability.
Robustness	Stable under sensor noise, occlusion, aspect-angle variation, and domain shift.
Faithfulness	Reflects decision-relevant model behavior rather than visual artifacts.
Manipulation resistance	Difficult to alter independently of the underlying prediction.
V&V suitability	Supports verification, validation, auditing, or assurance workflows.
Real-time feasibility	Compatible with operational latency constraints.
Uncertainty awareness	Linked to confidence, calibration, and uncertainty estimates.
Domain plausibility	Consistent with sensing physics and domain knowledge.

The dimensions in Table 1 are not intended as an exhaustive taxonomy of XAI quality. Rather, they capture properties that are particularly relevant when explanations are used as assurance evidence in safety-critical ATR. Interpretability is inherently stakeholder- and task-dependent: operators may require concise information supporting an immediate decision, engineers may require detailed diagnostic evidence for failure analysis, domain experts may assess consistency with sensor physics, and auditors or VV personnel may require reproducible and documented evidence. Robustness, faithfulness, and manipulation resistance concern the reliability of the explanation itself, whereas VV suitability, real-time feasibility, uncertainty awareness, and domain plausibility concern its use within an operational assurance process. Here, domain plausibility denotes consistency with established knowledge about the sensing modality, target characteristics, and operational environment; it does not by itself imply causal validity. This perspective is consistent with recent XAI evaluation work emphasizing faithfulness, robustness, human-centered usefulness, task-specific evaluation protocols, and the need for evaluation metrics that avoid misleading perturbation artifacts [12, 14, 15, 16, 17].

4. An XAI-Enabled Assurance Pipeline for ATR

To operationalize this perspective, we propose an XAI-enabled assurance pipeline for safety-critical ATR systems. The pipeline reframes XAI as part of a structured assurance workflow rather than as an isolated post-hoc explanation step.

The proposed pipeline is not intended as a certified safety process or a replacement for established verification and validation procedures. Rather, it provides a conceptual reference model for organizing XAI outputs as assurance-relevant evidence within safety-critical ATR workflows. The ordering of the pipeline reflects dependencies between different forms of assurance evidence: input quality should be assessed before model outputs are interpreted, predictive uncertainty contextualizes the detection, and an explanation must be generated before its quality and domain plausibility can be assessed. At the same time, the framework is modular rather than monolithic. Individual assurance functions and their implementation may vary with sensor modality, model architecture, operational context, and runtime constraints.

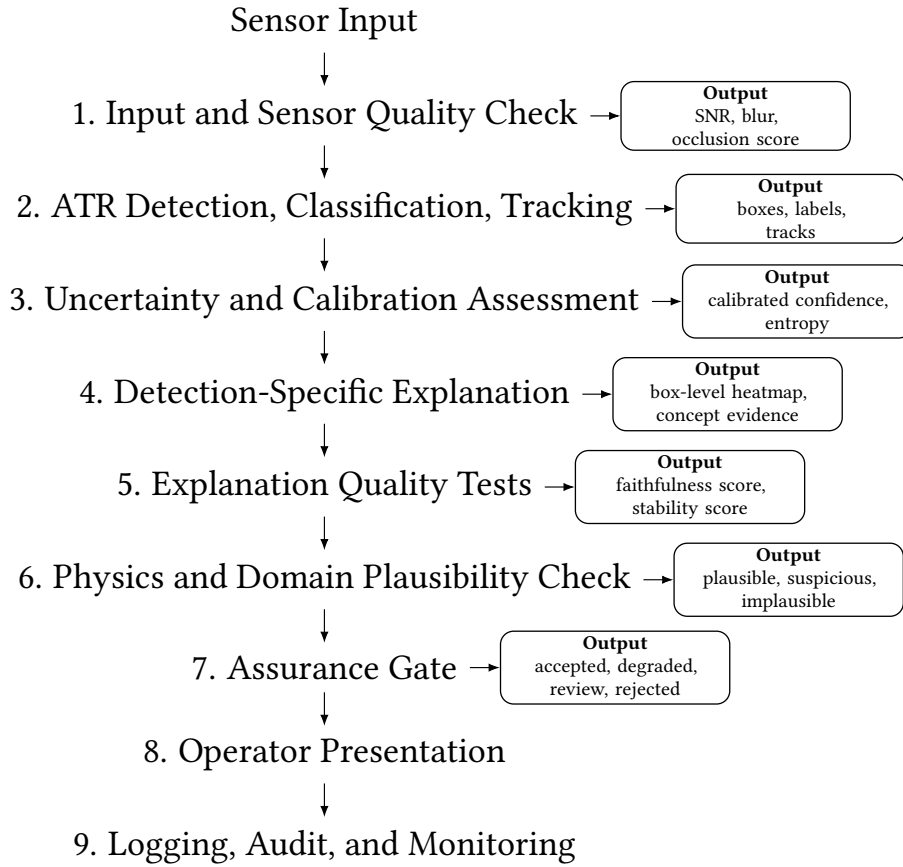


Figure 1: Conceptual XAI-enabled assurance pipeline for safety-critical ATR systems, including representative outputs of the main processing stages.

The pipeline starts with input and sensor quality assessment. Before interpreting a detection, the system should evaluate whether the input data are sufficiently reliable. This includes checks for noise, missing data, occlusion, sensor degradation, or adverse environmental conditions.

The ATR model then performs detection, classification, or tracking. The resulting output should be complemented by uncertainty and calibration assessment, because high confidence does not necessarily imply correctness, especially under distribution shift or unfamiliar target configurations.

A detection-specific explanation is then generated. Unlike generic saliency maps, this explanation should be tied to object-level outputs such as bounding boxes, class labels, detection scores, or tracking decisions. However, the mere existence of an explanation is insufficient. Explanation quality must be tested with respect to faithfulness, robustness, localization quality, and consistency under perturbations.

The next step is a physics or domain plausibility check. In radar, SAR, thermal, or multisensor ATR, explanations should be consistent with known sensing principles and domain expectations. An explanation that highlights background clutter rather than physically meaningful target structures may indicate unreliable model behavior.

The assurance gate integrates the available evidence from the preceding stages but is not intended as a universal weighted assurance score. Instead, it represents a policy-based evidence aggregation mechanism in which application-specific criteria determine the resulting assurance status. For example, an output may be accepted when sensor quality is sufficient, predictive uncertainty is low, required explanation-quality checks are satisfied, and no relevant domain-plausibility violation is detected. Non-critical deviations may result in a degraded status, while elevated uncertainty, insufficient explanation evidence, or questionable domain plausibility may route the output to human review. Critical violations may lead to rejection according to the operational policy.

Importantly, failure of an explanation-quality test does not necessarily imply that the underlying ATR prediction is incorrect. Instead, the corresponding explanation should no longer be considered valid assurance evidence. The assurance gate therefore assesses whether sufficient evidence supports operational reliance on the prediction rather than making or replacing the ATR prediction itself.

Whereas Table 1 defines the conceptual dimensions of assurance-relevant XAI, Table 2 illustrates how a subset of these dimensions can be operationalized through measurable checks and corresponding assurance actions.

Table 2: Example operationalization of assurance checks in the proposed ATR reference framework.

Check	Metric	ATR example	Action
Sensor quality	Blur, SNR, occlusion score	Low thermal contrast or partial occlusion	Degrade / review
Uncertainty	ECE, entropy, ensemble disagreement	High uncertainty for vehicle class	Review
Explanation faithfulness	Deletion, insertion, ROAD	Heatmap not decision-relevant	Invalidate evidence
Explanation robustness	Similarity under perturbation	Unstable under small blur	Degrade
Domain plausibility	Expert / rule-based check	Heatmap on shadow or clutter only	Flag / review
Runtime	ms/frame	Explanation too slow for online use	Offline-only

The checks and metrics in Table 2 are examples rather than fixed thresholds. Acceptable values depend on sensor modality, operational context, mission risk, and required human oversight. Calibration metrics such as ECE assess whether confidence reflects empirical correctness [18]. Faithfulness can be evaluated with deletion and insertion tests [19], while ROAD and F-Fidelity address limitations of perturbation-based evaluation [17, 20].

As an illustrative end-to-end example, consider a thermal image with low contrast and a partially occluded vehicle. The input-quality check identifies degraded conditions; the ATR model nevertheless returns a high-confidence vehicle detection, while predictive uncertainty is elevated. A detection-specific attribution is then found to be unstable under small perturbations and largely focused on a nearby shadow. The explanation therefore fails the required quality and domain-plausibility checks. Under a policy requiring acceptable sensor quality, uncertainty, and explanation evidence for automatic acceptance, the assurance gate assigns the detection to the review state rather than accepting it automatically. The operator receives the detection together with the elevated uncertainty and an indication that the explanation evidence is unreliable, and these states are logged for subsequent analysis. This example illustrates evidence aggregation rather than validated thresholds.

5. Discussion: Assurance, Human Oversight, and Trust Calibration

The proposed pipeline connects technical XAI methods with broader requirements for trustworthy AI. It supports human oversight by providing operators with explanations that are not only visually intuitive but also contextualized by uncertainty, robustness, and assurance status. However, meaningful human oversight requires more than the formal presence of a human operator; it also requires technical and organizational conditions that enable effective supervision, intervention, and accountability [21].

This is particularly important because XAI should not be designed to maximize user trust unconditionally. In safety-critical ATR, the goal is calibrated trust: users should rely on the system when evidence is strong and remain cautious when uncertainty, explanation instability, or domain implausibility indicates potential failure.

The pipeline also supports governance and risk management. Current frameworks such as the EU AI Act, the NIST AI Risk Management Framework, the NIST Generative AI Profile, and ISO/IEC 23894 emphasize transparency, human oversight, robustness, documentation, risk management, and monitoring [22, 23, 24, 25]. XAI can contribute to these objectives, but only as complementary evidence within a broader assurance process. It cannot replace verification, validation, certification, or safety-case argumentation.

Existing ML assurance approaches such as AMLAS structure safety claims and the evidence used to support them across the ML lifecycle [26]. The proposed framework is complementary rather than an alternative to such approaches: its narrower role is to structure how XAI-related ATR outputs are evaluated and transformed into assurance-relevant evidence. Faithfulness, robustness, uncertainty context, and domain plausibility can thus contribute evidence to broader assurance arguments, while the assurance gate indicates whether the currently available evidence is sufficient for an operational reliance decision. This perspective may also extend to autonomous and agentic systems, where explanations must similarly be connected to monitoring, intervention, and auditability [23, 24].

6. Conclusion

This position paper argued that XAI for safety-critical Automatic Target Recognition should move beyond post-hoc visualization toward assurance-oriented integration. Post-hoc explanations provide useful diagnostic information, but they are insufficient as standalone evidence of correctness, robustness, or safety [10].

The proposed pipeline is not a certified assurance process and has not yet been empirically validated on representative ATR benchmarks. Its contribution is therefore conceptual: it formulates a testable reference framework rather than providing validated thresholds or a comparison of individual explanation methods. The checks in Table 2 provide an initial operationalization for future evaluation across sensor modalities and operational conditions, particularly to investigate whether combinations of uncertainty, explanation quality, and domain plausibility correspond to ATR failure modes and support calibrated human decisions.

In summary, trustworthy XAI for ATR requires explanations that are faithful, robust, uncertainty-aware, domain-grounded, and embedded within human-centered governance and validation processes.

Declaration on Generative AI

During the preparation of this work, the author(s) used ChatGPT in order to support text structuring, language refinement, LaTeX drafting, and summarizing selected literature. After using this tool/service, the author(s) reviewed and edited the content as needed and take full responsibility for the publication's content.

References

- [1] J. Adebayo, J. Gilmer, M. Muelly, I. Goodfellow, M. Hardt, B. Kim, Sanity checks for saliency maps, in: *Advances in Neural Information Processing Systems*, 2018, pp. 9505–9515.
- [2] S. Jain, B. C. Wallace, Attention is not explanation, in: *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics*, 2019, pp. 3543–3556.
- [3] T. Laugel, M.-J. Lesot, C. Marsala, X. Renard, M. Detyniecki, The dangers of post-hoc interpretability: Unjustified counterfactual explanations, in: *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence*, 2019, pp. 2801–2807. doi:10.24963/ijcai.2019/388.
- [4] L. Longo, M. Brcic, F. Cabitza, J. Choi, R. Confalonieri, J. Del Ser, R. Guidotti, Y. Hayashi, F. Herrera, A. Holzinger, R. Jiang, H. Khosravi, F. Lecue, G. Malgieri, A. Páez, W. Samek, J. Schneider, T. Speith, S. Stumpf, Explainable artificial intelligence (XAI) 2.0: A manifesto of open challenges and interdisciplinary research directions, *Information Fusion* 106 (2024) 102301. doi:10.1016/j.inffus.2024.102301.
- [5] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, D. Batra, Grad-CAM: Visual explanations from deep networks via gradient-based localization, in: *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 618–626.
- [6] M. Sundararajan, A. Taly, Q. Yan, Axiomatic attribution for deep networks, in: *Proceedings of the International Conference on Machine Learning*, 2017, pp. 3319–3328.
- [7] M. T. Ribeiro, S. Singh, C. Guestrin, “Why Should I Trust You?” Explaining the Predictions of Any Classifier, in: *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2016, pp. 1135–1144. doi:10.1145/2939672.2939778.
- [8] S. M. Lundberg, S.-I. Lee, A unified approach to interpreting model predictions, in: *Advances in Neural Information Processing Systems*, 2017, pp. 4765–4774.
- [9] B. Kim, M. Wattenberg, J. Gilmer, C. Cai, J. Wexler, F. Viégas, R. Sayres, Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (TCAV), in: *Proceedings of the 35th International Conference on Machine Learning*, 2018, pp. 2668–2677.
- [10] C. Rudin, Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead, *Nature Machine Intelligence* 1 (2019) 206–215. doi:10.1038/s42256-019-0048-x.
- [11] Q. K. Nguyen, T. T. H. Nguyen, V. T. K. Nguyen, V. B. Truong, T. Phan, H. Cao, Efficient and concise explanations for object detection with Gaussian-Class Activation Mapping Explainer, *arXiv preprint arXiv:2404.13417* (2024).
- [12] L. P. T. Nguyen, H. T. T. Nguyen, H. Cao, ODExAI: A comprehensive object detection explainable ai evaluation, *arXiv preprint arXiv:2504.19249* (2025).
- [13] Z. Buçinca, M. B. Malaya, K. Z. Gajos, To trust or to think: Cognitive forcing functions can reduce overreliance on AI in AI-assisted decision-making, *Proceedings of the ACM on Human-Computer Interaction* 5 (2021) 1–21. doi:10.1145/3449287.
- [14] M. Nauta, J. Trienes, S. Pathak, E. Nguyen, M. Peters, Y. Schmitt, J. Schlötterer, M. van Keulen, C. Seifert, From anecdotal evidence to quantitative evaluation methods: A systematic review on evaluating explainable AI, *ACM Computing Surveys* 55 (2023) 1–42.
- [15] M. A. Kadir, A. Mosavi, D. Sonntag, Evaluation metrics for XAI: A review, taxonomy, and practical applications, in: *2023 IEEE 27th International Conference on Intelligent Engineering Systems (INES)*, IEEE, 2023, pp. 000111–000124.
- [16] X. Li, M. Du, J. Chen, Y. Chai, H. Lakkaraju, H. Xiong, M4: A unified XAI benchmark for faithfulness evaluation of feature attribution methods across metrics, modalities and models, in: *Advances in Neural Information Processing Systems*, 2023, pp. 000111–000124.
- [17] X. Zheng, F. Shirani, Z. Chen, C. Lin, W. Cheng, W. Guo, D. Luo, F-Fidelity: A robust framework for faithfulness evaluation of explainable AI, in: *Proceedings of the International Conference on Learning Representations*, 2025, pp. 1–14.
- [18] C. Guo, G. Pleiss, Y. Sun, K. Q. Weinberger, On calibration of modern neural networks, in: *Proceedings of the International Conference on Machine Learning*, 2017, pp. 1321–1330.

- [19] V. Petsiuk, A. Das, K. Saenko, RISE: Randomized input sampling for explanation of black-box models, in: *Proceedings of the British Machine Vision Conference*, BMVA Press, 2018, p. 151.
- [20] Y. Rong, T. Leemann, V. Borisov, G. Kasneci, E. Kasneci, A consistent and efficient evaluation strategy for attribution methods, in: *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, PMLR, 2022, pp. 18770–18795.
- [21] M. Ho-Dac, B. Martinez, Human oversight of artificial intelligence and technical standardisation, arXiv preprint arXiv:2407.17481 (2024).
- [22] European Parliament and Council of the European Union, Regulation (EU) 2024/1689 of the European Parliament and of the council of 13 june 2024 laying down harmonised rules on artificial intelligence and amending regulations and directives, Official Journal of the European Union, 2024. Artificial Intelligence Act.
- [23] National Institute of Standards and Technology, Artificial Intelligence Risk Management Framework (AI RMF 1.0), Technical Report, National Institute of Standards and Technology, 2023.
- [24] National Institute of Standards and Technology, Artificial Intelligence Risk Management Framework: Generative Artificial Intelligence Profile, Technical Report NIST AI 600-1, National Institute of Standards and Technology, 2024. doi:10.6028/NIST.AI.600-1.
- [25] International Organization for Standardization, ISO/IEC 23894:2023: Information Technology – Artificial Intelligence – Guidance on Risk Management, Technical Report, International Organization for Standardization, 2023.
- [26] R. Hawkins, C. Paterson, C. Picardi, Y. Jia, R. Calinescu, I. Habli, Guidance on the assurance of machine learning in autonomous systems (amlas), 2021. doi:10.48550/arXiv.2102.01564. arXiv:2102.01564.