International Journal of Advanced Engineering, Management and Science (IJAEMS)
Peer-Reviewed Journal
ISSN: 2454-1311 | Vol-12, Issue-2; Mar-Apr, 2026
Journal Home Page: https://ijaems.com/
DOI: https://dx.doi.org/10.22161/ijaems.122.14

Methodological Aspects of the Transition from Accuracy Metrics to Risk Modelling in the Design of Hybrid Intelligent Systems

Bezhentsev Yurii

Software Developer, Houston, Texas, United States

Received: 20 Feb 2026; Received in revised form: 22 Mar 2026; Accepted: 26 Mar 2026; Available online: 01 Apr 2026

Abstract *— The article examines the methodological transition from the use of standard accuracy metrics to risk-oriented modelling in the design of hybrid intelligent systems embedded in a management and control loop. The relevance of the approach stems from the fact that, in real processes, the quality of a solution is determined by the probability of undesired events, the magnitude of their consequences, and the speed of error detection and reversibility in the operational environment. The aim of the work is to translate the evaluation of intelligent models from the plane of numerical indicators into the language of systemic risk associated with people, infrastructure, and organisational procedures. The scientific novelty lies in integrating cost-sensitive error assessment, class imbalance analysis, and data shifts with an architectural description of hybrid (neuro-symbolic) systems, in which risk is distributed along the entire chain: data – model – rules – human – action. Accuracy metrics are proposed to be treated as particular input characteristics within a more general scheme for managing undesired events, defined by loss functions, barrier architecture, traceability, explainability, and controlled degradation modes. It is shown that, under class imbalance, label defects, and drifting data, the choice of metric and thresholds becomes a methodological decision that directly influences actual damage rather than a technical detail of the experiment. A conclusion is formulated on the necessity of shifting acceptance criteria from maximising aggregated metrics to constraining expected loss and ensuring risk controllability at the level of the hybrid system as a whole. The article is intended for researchers and engineers developing and deploying risk-sensitive intelligent systems in safety-critical and regulated domains.*

Keywords *— **hybrid intelligent systems, risk-oriented evaluation, accuracy metrics, risk modelling, cost-sensitive classification***

I. INTRODUCTION

The transition from evaluating intelligent solutions via accuracy metrics to risk modelling is associated with changes in the tasks for which such solutions are deployed. An intelligent system increasingly functions as part of a management and control loop, where the outcome is determined by how that prediction is translated into action, how quickly an error is detected, and the consequences for people, infrastructure, and the organisation [1]. Under these conditions, a risk-oriented methodological frame defines a different object of analysis. The focus shifts to undesired events and their consequences, as well as to prevention, containment, and recovery mechanisms that can be designed and empirically verified.

Accuracy metrics remain useful; however, their interpretation in real processes is unstable and depends on factors that are rarely captured in a laboratory formulation of the task. In practice, metric values may change substantially with shifts in class prevalence and with errors in the reference

annotation, leading to systematic misjudgements of a model's suitability for decision-making. This effect has been demonstrated in modelling studies, where even metrics commonly considered robust to imbalance vary significantly with changes in prevalence and ground-truth quality, and where high aggregate scores can arise alongside low actual utility of a classifier in the applied process [2].

In practice, asymmetry in consequences further amplifies this effect, since the costs of a false alarm and a miss rarely coincide. The target criterion is then naturally displaced towards expected losses and resource savings, which is formalised through error cost functions and enables comparison of solutions in terms of their impact on damage [3].

Hybrid intelligent systems integrate statistical models, formal rules, expert knowledge, and human-in-the-loop procedures; as a result, risk is distributed across the entire transformation chain from data to action. An error may be induced by data, algorithms, rule sets, integration logic, or usage protocols, and then amplified by organisational delays and by improper distribution of trust between automation and human operator [4]. Within such an architecture, a risk-oriented perspective enables responsibility to be decomposed across components and controlled operating regimes to be specified, including constraints, input-condition checks, routing of difficult cases, and traceability requirements. A survey of research on neuro-symbolic approaches emphasises that combining data-driven learning with symbolic reasoning improves interpretability and reliability in decision-making tasks, where the robustness of system behaviour in complex contexts is of primary importance [5].

II. MATERIALS AND METHODOLOGY

The materials are based on seven sources that present complementary lines of argument for why the evaluation of intelligent solutions should rest on risk and systemic consequences rather than solely on agreement with a reference annotation. The theoretical framework is provided by work on risk-oriented AI and ML, which emphasises that quality manifests through the probability of undesired events, the severity of damage, and the controllability of system behaviour within the operational loop, including error detectability and the effectiveness of response procedures [1].

The empirical basis for the challenges of interpreting accuracy metrics under real-world conditions comprises studies demonstrating the sensitivity of metrics to class prevalence and to defects in reference annotation, as well as the possibility of high aggregate scores coexisting with a classifier's low factual usefulness in the applied process [2]. For the formalisation of applied utility and the transition to criteria of expected losses, results on cost-sensitive classification and error cost functions are employed, enabling the linkage of threshold choice to the profile of consequences and resource constraints [3].

Materials on the architectural specificity of hybrid intelligent systems include studies of neuro-symbolic models in risk-sensitive domains and a survey of dependable neuro-symbolic approaches, highlighting interpretability, traceability, and robustness as engineering properties spanning the entire data–decision–action chain [4, 5]. Additionally, data are used on how class imbalance affects the consistency of conclusions when comparing models under different metrics, which is important for the methodological choice of criteria in tasks characterised by rare events and asymmetry of damage [6]. Practice-oriented measurement of the gap between out-of-deployment evaluation and in-deployment behaviour is supported by findings on harmful data shifts in clinical applications, where limited availability of true labels amplifies the importance of monitoring and protective mechanisms [7].

The research methodology is constructed as an analytical integration of several interrelated procedures aimed at translating model quality into the language of systemic risk for hybrid intelligent architectures.

First, a conceptual analysis of evaluation objects is performed, in which accuracy metrics are interpreted as partial descriptors of agreement with annotation under a fixed decision-making logic, while the central object becomes the risk of undesired events and its controllability within the application loop [1, 2]. Second, a methodological reconstruction of the relationship between metrics and consequences is carried out through a cost model of errors, where the criterion for comparing solutions is posed in terms of

expected losses and allows for threshold tuning that accounts for asymmetry of damage, contextual constraints, and instance-dependent costs [3].

Third, a comparative analysis of uncertainty regimes in hybrid systems is undertaken, in which sources of risk are distributed across components and integration interfaces, including data, model, rules, organisational delays, and human participation, while requirements for barriers, traceability, and robust degradation are identified in a manner consistent with the neuro-symbolic logic of reliability enhancement [4, 5]. Fourth, an empirically motivated analysis examines the influence of imbalance and distributional shifts on the validity of conclusions about model suitability. This analysis substantiates the need for risk segmentation, prevalence control, and monitoring of degradation indicators in operation [6, 7].

As a result, the chosen methodology fixes the transition from evaluation by numbers to a designed scheme of undesired-event management, in which accuracy metrics serve as input characteristics, while acceptance and validation criteria are formulated in terms of damage, detectability, reversibility, and response procedures at the level of the hybrid system as a whole [1–3, 5, 7].

III. RESULTS AND DISCUSSION

Basic accuracy metrics describe agreement between predictions and reference annotations under a fixed decision-making procedure [6]. The Accuracy indicator reflects the proportion of correct responses among all instances and depends on the selected threshold when the system outputs probabilities or scores. Precision and Recall capture different aspects of quality for a rare target event and are based on the structure of type I and type II errors in the confusion matrix. F1 aggregates precision and recall into a single value, which is convenient for comparing alternatives, but it obscures the underlying trade-off between error types and conveys limited information about system behaviour under threshold variation. ROC-AUC evaluates ranking performance across the entire threshold range and is often used to compare models; however, the resulting scalar value may have different practical interpretations depending on event prevalence and error-consequence profiles.

These differences are critical for hybrid intelligent systems because, further along the decision loop, the model score is converted into action via rules, constraints, and control procedures; consequently, the meaning of a metric is determined by how exactly the system will be used in the process.

Error asymmetry and the costs of false-positive and false-negative decisions shift evaluation from the domain of universal indicators into the domain of applied utility. When the cost of missing a high-risk event exceeds the cost of a false alarm, the optimal decision threshold shifts, and maximising standard metrics no longer coincides with minimising expected loss. Under class imbalance, the problem of average values is exacerbated: high Accuracy may be obtained by systematically ignoring a rare class, while F1 and related metrics yield different model rankings under identical base data and noise levels, making metric choice a component of methodology rather than a technical detail. Empirical studies demonstrate that, as the imbalance grows, the agreement between popular metrics decreases, and the selection of an indicator begins to alter which model is deemed best [6].

The gap between pre-deployment evaluation and in-deployment behaviour arises from changes in the context of use, shifts in data flows, and the influence of the procedures into which the solution is embedded. Data may drift over time, across sites, and across user groups, as well as through changes in measurement and recording practices. This leads to quality degradation and to shifts in probabilistic estimates. In clinical applications, situations arise in which data shifts can degrade performance and increase the risk of harmful interventions, while timely evaluation against true labels is often infeasible; consequently, monitoring shifts and designing protective mechanisms at the system level are required [7]. Model Evaluation Metrics are summarised in Table 1.

Table 1. Model Evaluation Metrics

Metric/idea	Meaning	Pitfall/limit	What matters in a system
Accuracy	share correct	class imbalance, threshold	don't rely on it alone
Precision	alert quality	threshold/prevalence sensitive	when FP are costly
Recall	don't miss	can inflate FP	when FN are costly
F1	P+R in one	hides trade-off/threshold behavior	add threshold-based view
ROC-AUC	ranking ability	interpretation shifts with prevalence/costs	tie to how the score is used
FP/FN cost	utility	metrics ≠ expected loss	set threshold by expected loss
Imbalance	rare class	metrics disagree	metric choice is methodological
Offline≠Online	deployment behavior	data shift, labels unavailable	monitoring + safeguards

The risk-oriented paradigm for evaluating intelligent systems rests on the notion that solution quality manifests in the probability of undesired events, the severity of their consequences, and the system's ability to detect hazardous regimes and keep them under control. Risk is conveniently conceptualised as the product of propensity to error, the cost of that error, and the degree of controllability, which is determined by detectability, reversibility, and the presence of response procedures. This definition renders the evaluation application-oriented and enables the linking of modelling results to safety, reliability, and resilience requirements in real operation. In hybrid intelligent systems, risk is not localised in a single component, so its assessment requires a description of the entire architecture and of how information and decisions propagate through rules, constraints, and human interaction.

A typology of risks is based on sources of uncertainty and mechanisms that amplify them. Model risks include behavioural instability at distributional boundaries, erroneous overconfidence, and degradation under changing conditions. Data-related risks include incomplete or biased data collection, distributional drift (shift in distribution), and data transformations before it is input to the model. Rules- and knowledge-related risks include contradictory rules, incomplete case coverage, and changes in regulations that such systems must comply with. Risks include operational risks of delays, failures of infrastructure, and breaks in integration and version incompatibilities; and human-factor risks of improper forms of trust allocation, ignoring processes, and misinterpretation of recommendations and assumptions. Regulatory risks manifest as non-compliance with established procedures, insufficient decision traceability, and a lack of audit readiness.

Risk as a systemic characteristic requires end-to-end consideration of the decision-making pipeline. At the input, measurements and data are collected; these are followed by cleaning and transformations, then by state estimation via the model, then by rule-based logic and constraints, after which the decision is converted into action and generates new feedback data. Errors may arise and accumulate at any stage, and their effect depends on the application context. A critical feature is that an identical model error occurring at different points of the pipeline leads to different consequences. Therefore, evaluation must account for escalation paths, stopping conditions, manual review procedures, and recovery speed, since these factors determine controllability and detectability of risk.

The methodology for transitioning to risk management begins with identifying undesired events at the process level, where an event is described in terms of the disrupted objective, context, and potential damage. Causal models are then constructed that connect events to combinations of conditions, data defects, model errors, rule conflicts, and

execution failures. Such analysis defines control points at which risk can be reduced through input-condition checks, action constraints, and switching to safe modes.

Subsequently, the loss function is formalised via an error cost matrix and an outcome utility, so that comparisons of alternatives reflect expected damage in process terms. Accuracy metrics are embedded into this frame through probability calibration, threshold tuning, and stratified evaluation across segments, since risk is unevenly distributed and concentrated in specific regimes. Demonstrable risk reduction is achieved through a barrier architecture that includes constraints, routing of difficult cases, abstention from decision-making under low reliability, control over action execution, and monitoring of degradation signals and feedback mechanisms that keep the system within an acceptable behavioural region. Figure 1 illustrates Risk Management in Intelligent Systems.

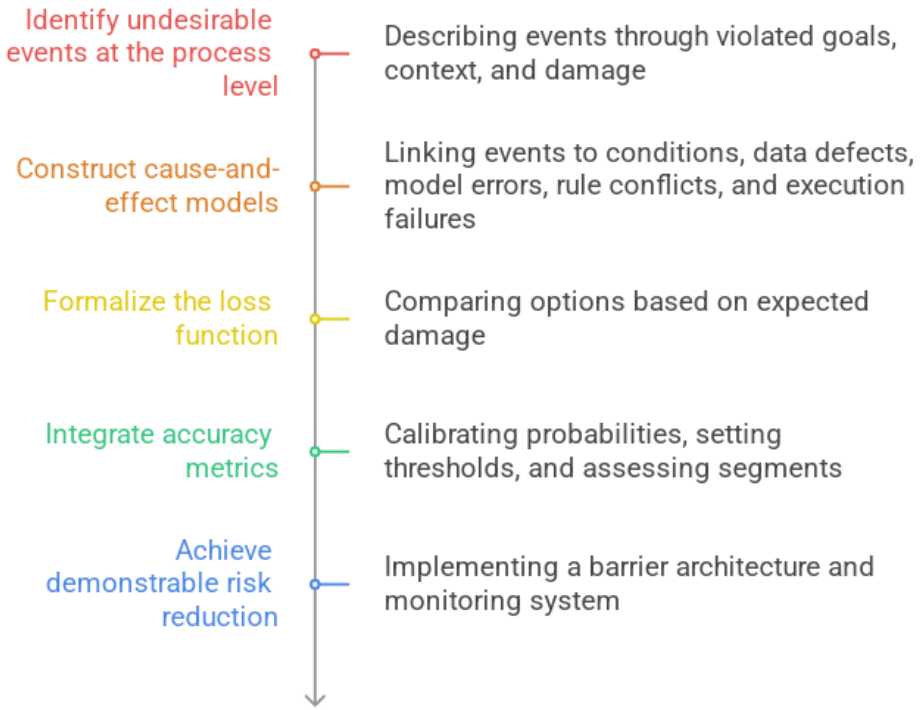

Fig. 1. Risk Management in Intelligent Systems

In risk modelling, a hybrid intelligent system is a linkage of heterogeneous components, each contributing its own uncertainty and mechanisms for constraining it. The statistical model generates estimates and rankings; rules define admissible decision regions and ensure compliance with regulations; expert knowledge determines the semantics of features and the prioritisation of consequences. The human operator makes decisions in situations where automatic action is hazardous or poorly specified. The distribution of responsibility becomes both a technical and an organisational object of design, specifying which component makes the decision, which component constrains the action, which component explains the reasons, and which component bears the duty to stop and escalate. Such decomposition links risk to architecture and prevents situations in which responsibility is blurred among model, rules, and operator.

Hybridisation mechanisms determine where exactly a control signal arises and how it is transformed into action. In a model-to-rule scheme, the model outputs an estimate, after which a set of constraints determines whether action is permissible and in what mode. Sometimes rules specify which cases are admissible, so the model can assume a smaller, more stable domain. For example, an ensemble or cascade may split risk across levels, with simple classifiers filtering out simple cases and more complex classifiers activating only in the presence of more complex cases. In this configuration, risk is shifted to interfaces, where alignment of confidence scales, a shared

interpretation of context, and predictable behaviour in the face of signal conflicts become crucial. In a hybrid system, value is provided by controlled degradation: failure of one layer should trigger a safe mode rather than uncontrolled action.

The human-in-the-loop contour enhances reliability when it is designed as an engineering mechanism rather than merely a formal signature in documentation. Escalation zones are defined according to uncertainty level, potential damage magnitude, and decision type. Trust control includes operator training, interface design, comprehensible rationales for recommendations, and explicit constraints that prevent blindly following system advice. Automation bias arises when an operator comes to perceive automated outputs as guaranteed correct and ceases to verify applicability conditions. Its prevention requires modes that allow the system to communicate its limitations, offer alternative actions, and demand confirmation in high-risk situations. In such settings, the human acts as part of the barrier architecture and assumes the role of active diagnostician rather than passive executor.

Traceability and explainability become instruments of risk management because they support verifiability of decisions and accelerate recovery after incidents. Traceability links a decision to input data, component versions, applied rules, context, and human actions. This enables localising the cause of an undesired event and updating a specific segment of the pipeline without dismantling the entire system. Explainability is important to ensure that escalation rules operate correctly and that operators understand which features and constraints led to an action. In hybrid systems, explanations must remain consistent with rules and data. Otherwise, they become decorative and increase risk by fostering a false sense of control. Risk management requires explanations that clearly delineate applicability boundaries and reveal the reasons for doubt.

Risk-oriented design relies on practices that keep the system within controlled regimes throughout its life cycle. Handling uncertainty includes confidence calibration, abstaining from decision-making when reliability is insufficient, and detecting cases that do not resemble known patterns. Scenario-based testing is built around stress conditions, boundary feature combinations, and rare events that account for the bulk of damage. Risk segmentation leads to differentiated thresholds and policies for different contexts, because identical errors incur different costs in different segments. Operational monitoring tracks data drift, performance degradation, incident growth, and changes in confidence distributions, after which feedback and component-update procedures are triggered. Risk management is reinforced through audits, decision logs, response procedures, and periodic revisions to the registry of undesired events, which transform acceptance criteria and the development life cycle itself. Requirements are formulated in terms of damage constraints and escalation rules. Design establishes barriers and control points. Validation examines risk scenarios and the system’s capacity for safe degradation. The operation comprises continuous monitoring, incident analysis, and policy and model updates to keep risk controllable in a changing environment. Figure 2 illustrates Hybrid Intelligent Systems.

When implementing a risk-oriented approach, a common distortion of meaning arises, treating risk as a cosmetic complication of familiar metrics. A team may introduce additional indicators, apply class weighting, and complicate averaging schemes, thereby creating an appearance of progress, while the connection to process-level consequences remains unspecified. As a result, the solution is optimised with respect to formal numbers, whereas undesired events continue to occur because their generating mechanism lies in the application's logic, in escalation rules, and in the limits of admissible actions. The absence of a formalised list of undesired events aggravates the problem, as discussion shifts toward convenient quality indicators while error costs remain implicit and each project participant interprets them idiosyncratically. Consequently, system requirements become blurred, and acceptance degrades into comparing models using metrics that are only weakly related to damage and controllability.

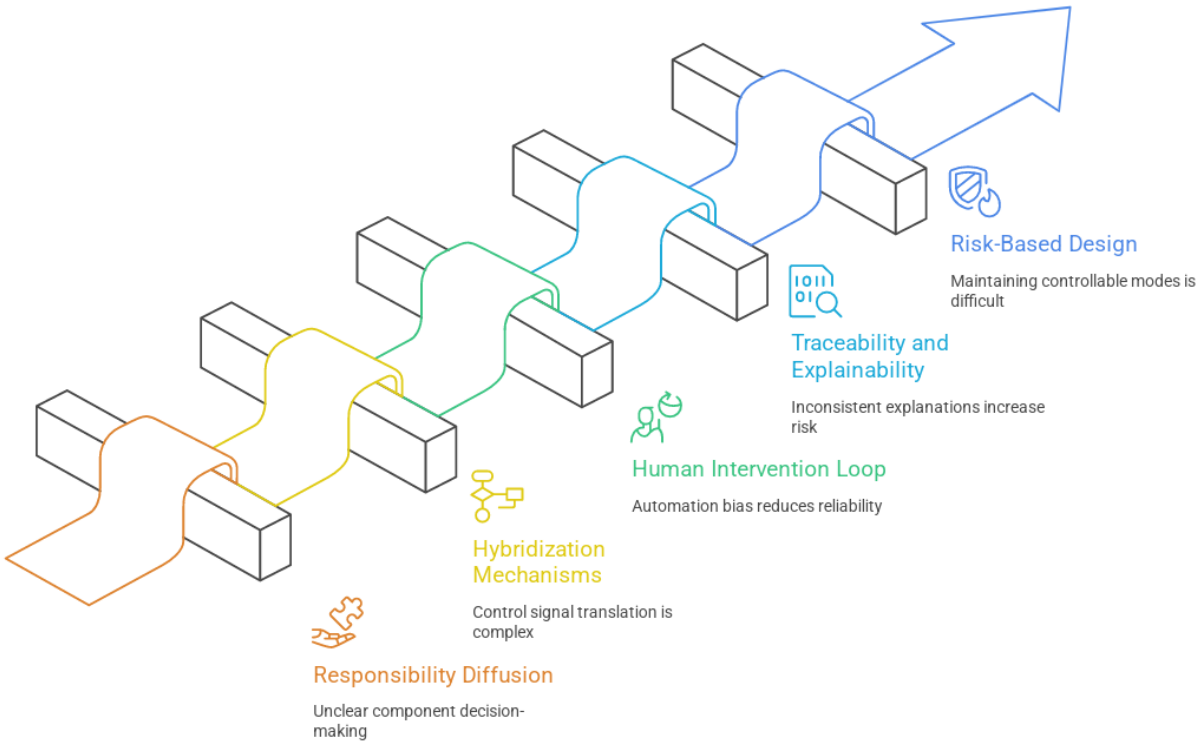

Fig. 2. Hybrid Intelligent Systems

Another persistent antipattern is associated with uncalibrated probabilities and arbitrary threshold management. Probabilistic estimates are perceived as reliable measures of confidence, although they may be systematically over- or underestimated, so thresholds begin to act as generators of latent risks. Threshold errors are particularly dangerous in hybrid architectures, where a threshold can trigger rule cascades, switch service modes, and influence human decisions. Neglect of the human and organisational loop entrenches these errors, because operators receive signals without clear applicability boundaries, and protocols lack procedures for stopping and verification. Over-complex modelling of damage completes this picture, when, instead of a minimally sufficient loss scheme, a cumbersome construct is introduced with numerous parameters that cannot be reliably estimated from data and cannot be maintained in a changing environment. In such circumstances, risk modelling loses practical footing, and the system reverts to intuitive decisions that are poorly reproducible and poorly controlled.

IV. CONCLUSION

The transition from evaluation via accuracy metrics to risk modelling naturally follows from the growing integration of intelligent systems into management and control loops, where significance attaches not only to agreement with a reference standard but also to how a prediction is converted into action, how rapidly an error is detected, and how extensive the consequences are for people, infrastructure, and the organisation. In this formulation, the object of analysis shifts to undesired events, their consequences, and designed mechanisms of prevention, containment, and recovery. Accuracy metrics retain their role, but their interpretation becomes context-dependent. In practice, metric values change substantially with shifts in class prevalence and with errors in reference annotation. High aggregate scores may coexist with low actual utility. Asymmetry of consequences renders threshold and criterion selection fundamental, because the cost of a false alarm and the cost of a miss generally differ. This shifts evaluation into the realm of expected losses and error cost functions, where comparisons of solutions are linked to impacts on damage and resource savings.

The article demonstrates that classical accuracy metrics describe agreement between predictions and

reference annotations under a fixed decision logic and therefore lose determinacy when the model score is used as an input to rules, constraints, and control procedures. Accuracy is sensitive to imbalance and threshold. Precision and Recall capture different facets of quality and depend directly on prevalence and the chosen threshold. F1 is convenient for comparison but obscures the trade-off between error types and provides limited insight into behaviour under threshold variation. ROC-AUC reflects ranking ability, yet the practical interpretation of its scalar value changes with event prevalence and consequence profiles. Against this background, the gap between pre-deployment evaluation and in-deployment behaviour becomes central, exacerbated by temporal and cross-site data shifts, changes in measurement practices, and limited availability of true labels. In clinical scenarios, it is particularly evident that drift can lead to degradation and to the risk of harmful interventions, necessitating shift monitoring and protective mechanisms at the system level. These conclusions are reinforced by the observation that, as imbalance increases, agreement among popular metrics diminishes, and the choice of indicator begins to change model-comparison outcomes.

A risk-oriented framework formalises quality in terms of the probability of undesired events, the severity of consequences, and controllability, which is determined by detectability, reversibility, and the presence of response procedures. In hybrid intelligent systems, risk emerges along the entire chain from data to action, as errors may arise in data, models, rules, integration logic, and usage protocols, and may be amplified by organisational delays and misallocation of trust between automation and human operators. Consequently, evaluation requires architectural description and mapping of decision paths. It must account for control points, stopping conditions, manual review, recovery speed, and the system's barrier organisation, in which constraints, routing of difficult cases, abstention from decisions under low reliability, and degradation monitoring keep behaviour within acceptable bounds. Traceability and explainability play a key role, as they provide verifiability, accelerate incident analysis, and help operators understand applicability limits.

Antipatterns are explicitly identified as: risk reduced to mere complication of familiar metrics without linkage to consequences; thresholds set arbitrarily under uncalibrated probabilities; the human and organisational loop remaining outside design; and damage models excessively parameterised and losing practical grounding. Against this backdrop, the methodological conclusion is that acceptance criteria for hybrid intelligent systems should shift from maximising aggregated accuracy metrics to constraining expected damage and ensuring risk controllability at the system level.

REFERENCES

[1] X. Zhang, F. T. S. Chan, C. Yan, and I. Bose, "Towards risk-aware artificial intelligence and machine learning systems: An overview," *Decision Support Systems*, vol. 159, p. 113800, May 2022, doi: https://doi.org/10.1016/j.dss.2022.113800.

[2] G. M. Foody, "Challenges in the real world use of classification accuracy metrics: From recall and precision to the Matthews correlation coefficient," *PLOS ONE*, vol. 18, no. 10, p. e0291908, Oct. 2023, doi: https://doi.org/10.1371/journal.pone.0291908.

[3] S. De Vos, T. Vanderschueren, T. Verdonck, and W. Verbeke, "Robust instance-dependent cost-sensitive classification," *Advances in Data Analysis and Classification*, vol. 17, pp. 1057–1079, Jan. 2023, doi: https://doi.org/10.1007/s11634-022-00533-3.

[4] C. K. Kolli, "Hybrid Neuro-Symbolic Models for Ethical AI in Risk-Sensitive Domains," *arXiv*, 2025, doi: https://doi.org/10.48550/arXiv.2511.17644.

[5] Z. Lu, I. Afridi, H. J. Kang, I. Ruchkin, and X. Zheng, "Surveying neuro-symbolic approaches for reliable artificial intelligence of things," *Journal of Reliable Intelligent Environments*, vol. 10, pp. 257–279, Jul. 2024, doi: https://doi.org/10.1007/s40860-024-00231-1.

[6] J.-G. Gaudreault and P. Branco, "Empirical analysis of performance assessment for imbalanced classification," *Machine Learning*, vol. 113, pp. 5533–5575, Jan. 2024, doi: https://doi.org/10.1007/s10994-023-06497-5.

[7] V. Subasri *et al.*, "Detecting and Remediating Harmful Data Shifts for the Responsible Deployment of Clinical AI Models," *JAMA Network Open*, vol. 8, no. 6, p. e2513685, Jun. 2025, doi: https://doi.org/10.1001/jamanetworkopen.2025.13685.