Eurasia Specialized Veterinary Publication

International Journal of Veterinary Research and Allied Science

ISSN:3062-357X

2026, Volume 6, Issue 1, Page No: 12-23
Copyright CC BY-NC-SA 4.0
Available online at: www.esvpub.com/

Veterinary Artificial Intelligence Should Know When Not to Decide: A Clinical Governance Architecture for Diagnostic Uncertainty, Distribution Shift, Model Abstention, Explainability, Human Oversight, Safe Escalation, and Post-Deployment Monitoring

Krzysztof Wiśniewski[1]*, Małgorzata Nowak[1], Tomasz Adamczyk[2]

[1]Department of Veterinary AI and Clinical Governance, Faculty of Veterinary Medicine, Warsaw University of Life Sciences, Warsaw, Poland.
[2]Department of Veterinary Diagnostic Informatics and Model Validation, Faculty of Veterinary Medicine, Jagiellonian University, Krakow, Poland.

***E-mail** ✉ krzysztof.wisniewski@sggw.edu.pl

ABSTRACT

Artificial intelligence is moving from experimental veterinary applications toward clinical decision support, yet predictive performance alone does not establish whether an output is safe to use for a particular animal, task, clinician, or setting. This article develops an original non-empirical governance architecture for veterinary AI centered on a simple requirement: a clinically responsible system should be able to recognize conditions in which it should not control the decision. The analysis separates intended use, training-data representativeness, reference-standard quality, external validity, distribution shift, out-of-distribution inputs, calibration, predictive uncertainty, abstention, explainability, human oversight, escalation, post-deployment monitoring, updating, and incident learning. The proposed architecture treats decision authority as conditional rather than fixed. AI output may remain actionable only while the case and operating environment remain within a validated-use envelope and uncertainty is compatible with the consequence of error; otherwise, authority should shift toward veterinary review, additional testing, or non-use. This framing also distinguishes low confidence from unfamiliar inputs, explanation from correctness, nominal human presence from effective oversight, detected distribution shift from demonstrated performance degradation, and model updating from validated improvement. The architecture is intended as a governance synthesis rather than a validated clinical intervention. Much of the empirical safety literature still comes from human medical AI, while veterinary evidence is concentrated in diagnostic imaging and commercial-product transparency. Prospective multispecies validation, veterinarian–AI interaction studies, workload-sensitive abstention testing, and longitudinal post-deployment evaluation are therefore necessary before clinical effectiveness or welfare benefit can be inferred.

Keywords: Veterinary artificial intelligence, Clinical governance, Model abstention, Predictive uncertainty, Human oversight, Distribution shift

Received: 27 October 2025
Revised: 27 January 2026
Accepted: 02 February 2026

How to Cite This Article: Wiśniewski K, Nowak M, Adamczyk T. Veterinary Artificial Intelligence Should Know When Not to Decide: A Clinical Governance Architecture for Diagnostic Uncertainty, Distribution Shift, Model Abstention, Explainability, Human Oversight, Safe Escalation, and Post-Deployment Monitoring. Int J Vet Res Allied Sci. 2026;6(1):12-23. https://doi.org/10.51847/nln521tXTp

Introduction

Veterinary AI is increasingly visible in diagnostic imaging and commercial decision-support products, but the evidence does not justify treating “AI performance” as a single, transferable property. Recent veterinary literature shows expanding technical capability, distinct error profiles between commercial AI and veterinary radiologists, and substantial gaps in publicly available information about training populations, validation, out-of-distribution

© 2026 International Journal of Veterinary Research and Allied Science

handling, and monitoring [1–3]. These observations create a governance problem rather than merely a benchmarking problem: clinicians must know not only whether a model can classify accurately, but under which conditions its output should influence care.

Veterinary radiology scholarship has therefore emphasized independent evaluation and the continuing role of professional judgment [4]. This article extends that logic by proposing that safe veterinary AI requires conditional decision authority. The core distinction is between a model producing an output and a clinical system being justified in acting on that output. The analysis consequently links task definition, evidence quality, external validity, uncertainty, abstention, oversight, escalation, and lifecycle monitoring while keeping veterinary-specific evidence separate from principles imported from human clinical AI.

From predictive performance to clinical safety

Diagnostic accuracy, sensitivity, specificity, discrimination, or other performance measures remain necessary, but none alone defines clinical safety. In canine and feline radiography, commercial AI and veterinary radiologists can achieve similar aggregate performance while exhibiting different error trade-offs [2]. A model can therefore appear competitive overall yet fail differently on findings that matter disproportionately to clinical action. Veterinary evaluation guidance likewise treats algorithms as decision-support tools requiring appraisal of specifications, evidence, and intended use rather than automatic equivalence with specialist interpretation [4].

The proposed safety concept is consequently relational: safety emerges from the interaction between model, patient, clinical task, user, workflow, and consequence of error. This does not make predictive performance secondary; it changes its interpretation. A high-performing model operating outside the population, acquisition process, or decision context in which it was validated may warrant less authority than a modestly performing model used within a narrowly defined and well-characterized task. Clinical safety is therefore treated here as a governed use condition, not an intrinsic model label.

Intended use and clinical task definition

Clinical translation fails when development metrics are interpreted without specifying what decision the system is meant to support. Work in clinical AI has emphasized that strong retrospective performance does not itself establish clinical impact, particularly when workflow, population, or implementation context changes [5]. For veterinary AI, intended use should therefore identify the target species or population, input type, clinical question, expected user, output, decision consequence, and setting in which the output may influence care.

Early-stage clinical evaluation guidance further requires attention to live workflow and human–AI interaction rather than model behavior in isolation [6]. Trustworthy-AI consensus recommendations similarly frame safety, robustness, usability, traceability, fairness, and explainability as lifecycle properties [7]. Adapted to veterinary care, these principles support a proposed “task contract”: the AI system should be governed according to the exact decision it is authorized to assist. A model validated for case prioritization, for example, should not silently acquire authority for diagnosis, prognosis, or treatment selection without separate evidence.

Training-data representativeness

Training-data representativeness is not a binary property. For veterinary AI, relevant dimensions may include species, breed, age, sex, disease spectrum, referral status, geography, prevalence, equipment, image-acquisition protocol, practice type, and time period, although the importance of each dimension is task dependent. The commercial veterinary AI audit found that training-population and subgroup information was often not publicly disclosed [3]. Such non-disclosure prevents independent assessment of representativeness; it does not prove that a product is biased or poorly trained.

The governance implication is to separate dataset composition from clinical transportability. A dataset can be internally diverse yet still fail to represent the population or acquisition conditions of a new practice. Conversely, a deliberately narrow dataset may be appropriate for a tightly constrained intended use. Veterinary evaluation frameworks therefore support examining whether development data match the declared clinical task rather than rewarding diversity as an end in itself [4]. The proposed architecture treats representativeness as evidence about the validated-use envelope, not as evidence of universal generalization.

Label quality and reference standards

AI systems learn and are evaluated against labels that may themselves contain error. A contemporary survey of medical-image learning under label noise shows that incorrect or inconsistent labels can alter model training and evaluation [8]. The governance issue is especially important when the model's apparent correctness depends on a reference that is treated as unquestionable.

Radiology research demonstrates that annotation schemas and expert readers can disagree, meaning that a "gold" label may partly reflect how a reference standard was constructed [9]. Veterinary imaging reviews identify related constraints in dataset quality and ground-truth definition [10]. These findings do not imply that valid reference standards are impossible. They require that reference-standard provenance, expertise, adjudication, repeatability, and known ambiguity be considered when interpreting model errors. The proposed architecture therefore distinguishes model uncertainty from reference-standard uncertainty. A discordant AI–clinician result may represent model failure, clinician error, label error, biological ambiguity, or a combination; governance should not collapse those explanations into a single accuracy statistic.

External validity

External validity concerns whether evidence generated in one development or validation context supports use in another. Clinical AI translation literature cautions that performance in curated or internal datasets does not establish impact in routine settings [5]. In veterinary AI, the problem is sharpened when the characteristics of training data, external testing, or product updates are incompletely disclosed [3]. External validity should therefore be judged against the declared intended use rather than treated as a permanent certificate attached to a model.

This article proposes the term validated-use envelope for the combination of species or population, clinical task, acquisition conditions, workflow, model version, and decision context for which supporting evidence is reasonably applicable. The envelope is conditional: evidence may transfer strongly to a closely matched practice yet weakly to another species, device, referral population, or time period. External validation can enlarge confidence in that envelope, but a single external study cannot establish universal transportability.

Because these states change both interpretation and allowable action, they are better adjudicated explicitly than compressed into a binary judgment. **Table 1** presents deployment conditions inside, at the edge of, and outside a veterinary AI validation envelope.

Table 1. Deployment conditions inside, at the edge of, and outside a veterinary AI validation envelope

Validation-envelope dimension	Closely matched use	Partial-context match	Envelope boundary	Outside envelope	Temporal mismatch
Clinical meaning	Evidence comparatively transportable	Applicability weakened	Evidence sparse	Transfer unsupported	Earlier evidence may have decayed
Population, task, device, or workflow match	Same species, task, device, workflow	Prevalence, breed mix, referral status, or device differs	Atypical phenotype or acquisition	New species, task, workflow, or input process	Case mix, hardware, software, or model version changed
Validation evidence	Independent testing matches deployment	Only some deployment dimensions tested	Few comparable validated cases	No directly applicable validation	Current use differs from prior validation
AI authority allowed	Planned support role may continue	Reduce reliance; targeted review	Require veterinary corroboration	Withhold expanded AI authority	Heighten surveillance
Residual transport problem	Residual sampling/time uncertainty	Effect direction uncertain	Boundary membership uncertain	Lack of evidence is not proof of failure	Shift may be benign
Required assurance	Routine monitoring	Local/subgroup validation	Track repeated boundary cases	Revalidate before use expansion	Reassess calibration and performance
Basis	Evidence-informed	Proposed	Proposed	Proposed	Proposed

Distribution shift

Distribution shift occurs when deployment data differ from the data on which model behavior was established. Postmarket research shows that shifts can sometimes be detected statistically before complete outcome labels are

available; uncertainty research emphasizes that point predictions can obscure limits of knowledge; and multisource medical-AI studies show that hidden acquisition shortcuts can impair generalization [11–13]. None of these findings means that every detected shift is clinically harmful. A shift is a warning state whose significance depends on whether performance, calibration, or actionability changes.

In veterinary settings, plausible sources include species or breed mix, disease prevalence, referral patterns, imaging equipment, acquisition protocols, laboratory platforms, seasonal conditions, workflow changes, and software updates. These categories are proposed risk dimensions rather than empirically validated veterinary thresholds. Governance should therefore distinguish shift detection from shift adjudication: the first asks whether the input environment has changed; the second asks whether the change alters clinically relevant model behavior and whether decision authority should be reduced, suspended, or revalidated.

Out-of-distribution inputs

An out-of-distribution input is a stronger challenge than ordinary variation within the validated-use envelope. The practical concern is not simply that a case is unusual, but that the model may be processing an input for which its learned representation is poorly supported. Distribution-shift methods can identify some changes in input structure, but detection is imperfect and does not itself establish clinical consequences [11]. Uncertainty estimates may also help identify cases requiring additional scrutiny, yet uncertainty methods are not uniformly reliable across models or settings [12].

The proposed architecture therefore keeps OOD status and low confidence separate. A familiar in-distribution case can yield low confidence because the disease pattern is ambiguous, while an unfamiliar input can occasionally produce high confidence despite poor support. Either state may justify escalation, but for different reasons. OOD handling should be specified before deployment, tested with task-relevant challenge cases, and connected to an actionable response such as repeat acquisition, alternative testing, specialist review, or refusal of automated interpretation.

Calibration and predictive uncertainty

Predictive confidence becomes clinically meaningful only if it is sufficiently related to correctness or observed outcome frequency. Medical-imaging work shows that label noise can distort calibration and that calibration methods themselves require evaluation under imperfect labels [14]. Calibration is also distinct from discrimination: it concerns agreement between predicted probabilities and observed outcomes, and multiple complementary assessments may be needed [15]. A model may rank cases well while assigning probabilities that are systematically too high or too low [16].

For veterinary governance, this matters because decision thresholds often depend on the credibility of confidence, not merely on rank order. The proposed architecture therefore treats calibration as a property that may vary by site, subgroup, prevalence, time, and model version. Predictive uncertainty, meanwhile, should be interpreted as evidence about how cautiously an output is used, not as a universal measure of clinical danger.

The clinically important states are summarized as a response ledger because equal confidence values should not imply equal authority across unlike validation conditions. **Table 2** presents confidence meanings created by the interaction of calibration and input familiarity.

Table 2. Confidence meanings created by the interaction of calibration and input familiarity

Calibration–uncertainty configuration	Input and deployment condition	Meaning of confidence	AI authority allowed	How confidence can fail	Validation required
Calibrated, low uncertainty	**Condition:** Stable population and model version **Observed:** Confidence aligns with observed outcomes	Confidence comparatively interpretable	Use within intended support role	Calibration can drift	**Recheck:** Continue surveillance **Basis:** Evidence-informed
Discriminative but miscalibrated	**Condition:** Changed prevalence/population **Observed:** Good separation with systematic probability error	Ranking useful; probability magnitude unreliable	Avoid probability-dependent action	Error direction may vary	**Recheck:** Recalibrate; externally verify **Basis:** Evidence-informed

High uncertainty, familiar input	**Condition:** Weak or overlapping clinical signal **Observed:** Low or unstable confidence on valid input	Ambiguity within known domain	Escalate or corroborate	May reflect intrinsic ambiguity	**Recheck:** Track outcomes and review burden **Basis:** Proposed
Confident, unfamiliar input	**Condition:** OOD or shifted acquisition **Observed:** Decisive score despite unsupported input context	Confidence may lack epistemic support	Reduce AI authority	OOD detection may fail	**Recheck:** Investigate distribution; revalidate **Basis:** Proposed
Calibration unknown	**Condition:** New site, subgroup, or version **Observed:** Suitable calibration evidence absent	Confidence lacks validated probabilistic meaning	Avoid unsupported thresholds	Absence of evidence is not miscalibration	**Recheck:** Collect local outcomes **Basis:** Proposed

Model abstention

Model abstention is the deliberate decision not to issue, complete, or operationalize an AI prediction under defined conditions. It should not be treated as another name for uncertainty. Uncertainty is information about confidence or knowledge limits; abstention is a governance action that changes who or what holds decision authority. This distinction matters because an uncertainty estimate can be displayed while the system still pressures the clinician toward a recommendation.

The proposed veterinary rule is therefore risk sensitive: abstention should be considered when the case lies outside the validated-use envelope, predictive confidence is not trustworthy, clinical consequences are high, required inputs are inadequate, or the system detects conditions for which its competence has not been established. Calibration evidence supports the broader point that a probability should not drive action merely because it is numerically precise [16].

Abstention also has costs. Excessive deferral can create workload, delay, or access problems, particularly where specialist review is limited. A safe abstention policy must therefore be evaluated jointly with the availability and timeliness of the veterinary pathway receiving deferred cases. No universal veterinary threshold is proposed here.

Explainability and interpretability

Explainability can support clinical scrutiny, but an explanation should not be mistaken for evidence that a prediction is correct. Trustworthy-AI guidance treats explainability as one component of a broader lifecycle system that also includes robustness, traceability, usability, and safety [7]. In veterinary practice, its value therefore depends on whether it helps a clinician interrogate an output, recognize limitations, identify relevant inputs, or decide whether further evaluation is necessary.

Interpretability requirements should also be proportionate to the decision being influenced. Early clinical AI evaluation emphasizes the importance of studying how information is presented and used within real workflows [6]. A low-consequence prioritization tool may require different explanatory support from an AI recommendation affecting surgery, euthanasia, antimicrobial treatment, or another consequential intervention. The proposed architecture consequently treats explanation as a support for challengeability, not a substitute for external validation, calibration, uncertainty assessment, or veterinary judgment. An intuitively persuasive explanation accompanying an incorrect prediction can itself become a safety problem if it increases unwarranted confidence.

Human oversight

Human oversight is meaningful only when the veterinarian can recognize, question, override, and appropriately escalate an AI output. Experimental clinical evidence demonstrates that incorrect computerized advice can reduce professional diagnostic accuracy, that machine-learning recommendations can influence treatment choices even when explanations are supplied, and that the effects of AI assistance vary substantially among clinicians and tasks [17–19]. These findings establish a human-factor risk in medicine, not a measured veterinary effect size.

The proposed governance architecture therefore distinguishes nominal oversight from effective oversight. Merely placing a veterinarian “in the loop” does not guarantee protection if time pressure, interface design, confidence presentation, workload, or institutional expectations make disagreement difficult. Effective oversight requires sufficient clinical information, opportunity to inspect discordant evidence, authority to reject the output, and

access to an alternative decision pathway. These conditions should be evaluated prospectively in veterinary workflows rather than inferred from professional qualification alone.

Automation bias

Automation bias becomes relevant when AI output disproportionately anchors subsequent interpretation or discourages independent evidence gathering. The problem is not that clinicians use decision support; appropriate reliance is the purpose of such systems. The safety concern is miscalibrated reliance—accepting AI when it should be challenged or rejecting it when it is informative. Evidence that explanations do not invariably protect clinicians from erroneous recommendations reinforces this distinction [18].

Expertise cannot be assumed to eliminate the problem. AI assistance has produced heterogeneous effects across radiologists and diagnostic tasks, including improvement in some circumstances and deterioration in others [19]. Veterinary susceptibility, however, remains insufficiently quantified. The architecture therefore treats automation bias as a possible decision state requiring observable indicators rather than as an assumed clinician trait.

Table 3 presents clinician–AI interaction patterns and the audits needed to distinguish reliance from automation bias.

Table 3. Clinician–AI interaction patterns and the audits needed to distinguish reliance from automation bias

Clinician–AI relationship	Source of agreement or disagreement	Clinical reading	Governance response	Audit question	Basis
Evidence-concordant reliance	**Context:** Familiar validated case **Observed:** Clinician reasoning remains recoverable	AI agrees with independently supported clinical evidence	AI support may continue	**Unresolved:** Agreement does not prove independence **Audit:** Periodic override review	Evidence-informed
Uncritical anchoring	**Context:** Time pressure or persuasive interface **Observed:** Differential diagnoses or checks prematurely stop	AI disproportionately narrows alternatives	Require deliberate reappraisal	**Unresolved:** Internal reasoning is difficult to observe **Audit:** Human-factor evaluation	Proposed
Appropriate disagreement	**Context:** Atypical presentation or new evidence **Observed:** Override linked to clinical findings	Veterinarian rejects AI for defensible case-specific reasons	Preserve veterinary authority	**Unresolved:** Clinician may also be wrong **Audit:** Outcome-linked review	Proposed
Constrained agreement	**Context:** Limited specialist access or workload **Observed:** Agreement persists despite unresolved concern	AI accepted because alternatives are unavailable	Escalation pathway may be inadequate	**Unresolved:** Behavior may reflect resources, not trust **Audit:** Assess capacity and access	Proposed
Recurrent discordance	**Context:** Site, subgroup, or workflow change **Observed:** Pattern of repeated disagreement	AI and clinicians systematically diverge	Investigate model and workflow	**Unresolved:** Either party may be responsible **Audit:** Joint model–human audit	Proposed

Safe escalation to veterinary judgment

Selective prediction demonstrates that an AI system can intentionally defer cases rather than automate every eligible input [20]. This creates a formal trade-off between coverage and error burden and provides a methodological basis for abstention as a governance action. Confidence-based pathology research further shows that low-confidence cases can, in a specifically validated setting, be preferentially directed toward review [21]. Such findings support the principle of selective review but do not supply veterinary confidence thresholds.

Operational safety work also shows that a decision space can include an intermediate region in which automated rule-in or rule-out is inappropriate and human review is required [22]. Veterinary adaptation should preserve that structure while validating task-specific boundaries locally. Escalation must additionally restore genuine professional authority rather than merely route a difficult case to a veterinarian who remains behaviorally subordinate to the algorithm. Veterinary scholarship on epistemic authority identifies this as a distinct emerging concern [23]. Safe escalation is therefore proposed as a transfer of decision authority, not simply a notification.

A proposed veterinary AI governance architecture

The proposed architecture organizes clinical AI authority around four interacting questions: Is the task authorized? Is the evidence applicable? Is the output epistemically trustworthy for this case? Is an adequate veterinary response available if uncertainty is unacceptable? These questions integrate intended use, reference-standard quality, external validity, distribution state, calibration, abstention, and effective oversight. They do not create a validated scoring system.

Decision authority is proposed to vary across a validated-use envelope. Within a well-supported use state, AI may contribute according to its defined role. At an evidence, distribution, calibration, or uncertainty boundary, authority should become conditional. Outside that envelope, or where consequences exceed the evidence supporting autonomous use, the architecture favors abstention, corroboration, alternative testing, or veterinary escalation. Operational gray-zone methodology supports separating such states rather than forcing every case into an automated binary decision [22].

Abstention and human oversight are therefore complementary controls: the first limits when the model should decide; the second determines whether a human can safely take over when it does not [20]. Veterinary epistemic authority remains the final professional safeguard rather than an ornamental human-in-the-loop label [23].

The architecture's distinctive logic is compressed visually because its safety function depends on reconciling several evidence streams and preserving their disagreements rather than averaging them into a single confidence score.

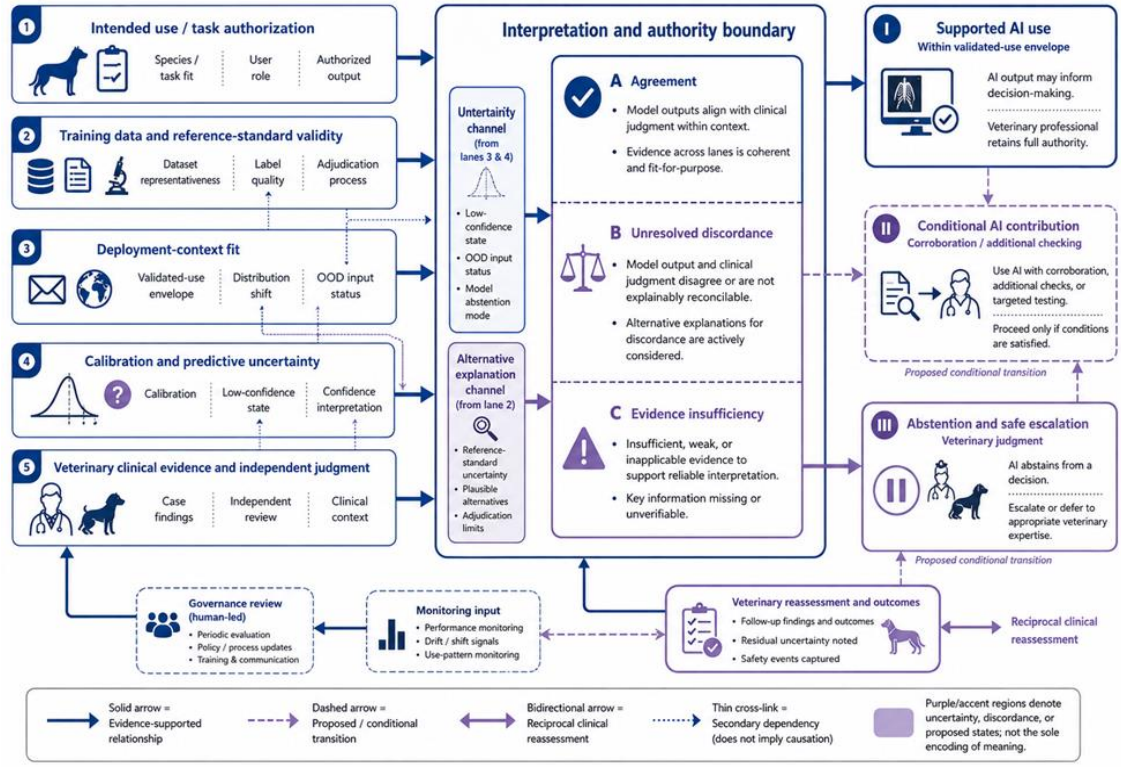

Figure 1. Reconciliation boundary for assigning veterinary AI decision authority

Post-deployment performance monitoring

Predeployment validation cannot characterize every condition encountered after implementation. Large-scale monitoring of deployed intracranial-hemorrhage AI identified context-specific false-positive patterns associated with technical and imaging circumstances, illustrating why error surveillance must remain active after deployment [24]. Veterinary monitoring should likewise examine not only headline performance but changing case mix, acquisition conditions, output distributions, overrides, abstentions, discordant cases, and clinically meaningful failures.

Post-deployment surveillance also requires organizational ownership. Operational frameworks emphasize that monitoring must connect signal detection to investigation and response rather than function as passive data collection [25]. The possibility of updating complicates this process because updating methods vary widely across

AI prediction models [26]. Longitudinal evidence shows that dynamic updating can maintain or improve performance in a specific prediction setting, but it also changes the model being governed [27].

The proposed monitoring states below therefore connect observed change to proportionate action without assuming that every signal represents model deterioration. **Table 4** presents post-deployment signals that change veterinary AI authorization, investigation, or revalidation.

Table 4. Post-deployment signals that change veterinary AI authorization, investigation, or revalidation

Post-deployment signal	Changed component	Operational evidence	Authorization response	Investigation question	Return-to-use criterion	Basis
Stable observed operation	Comparable case mix and workflow	Output/error patterns remain compatible with validation	Continue authorized use	**Meaning:** No material safety signal identified **Alternative:** Absence of signal is not proof of safety	Routine surveillance	Proposed
Input-distribution change	Population, device, protocol, prevalence, or workflow changes	Detectable input/case-mix difference	Intensify review	**Meaning:** Deployment environment has shifted **Alternative:** Shift may be benign	Test performance/calibration	Evidence-informed
Outcome-performance concern	Subgroup or temporal concentration	Error, calibration, or discordance pattern emerges	Restrict reliance if warranted	**Meaning:** Clinically relevant behavior may have changed **Alternative:** Reference standard may also change	Structured investigation	Proposed
Recurrent human–AI discordance	Clinician group or site-specific pattern	Overrides or disagreements cluster	Audit both model and workflow	**Meaning:** Model/workflow interaction requires review **Alternative:** Human judgment is not infallible	Outcome-linked adjudication	Proposed
Material model change	Retraining, recalibration, software/version change	New governed model state	Condition continued use on revalidation	**Meaning:** Prior validation may not fully transfer **Alternative:** Magnitude of change matters	Version-specific assurance	Proposed

Model updating and version control

Updating should not be treated as an automatic remedy for deteriorating performance. Reviews of AI prediction-model updating show heterogeneous methods, including recalibration, model revision, and retraining, with substantial differences in evaluation practice [26]. A dynamic updating pipeline can preserve performance over time in a particular clinical prediction context, but that evidence does not establish that every update improves every model [27].

The governance implication is that a materially modified model becomes a new clinical object whose relationship to previous validation must be examined. This article proposes conditional validation inheritance: minor, well-characterized modifications may justify limited re-evaluation, whereas substantial changes in training data, architecture, intended use, calibration, preprocessing, or decision thresholds may require broader external and workflow validation.

Version control should link the deployed model identifier to its training or update data, calibration state, intended-use envelope, deployment date, known limitations, monitoring history, and relevant incidents. Rollback capability should be considered where technically feasible. Clinical learning, software modification, and formal model updating must remain separate concepts; observation of a failure should trigger investigation, not uncontrolled adaptation.

Incident and failure reporting

Veterinary AI failures require a learning process that can distinguish model error from data-quality problems, reference-standard uncertainty, workflow failure, inappropriate use, human overreliance, and combinations of these factors. Veterinary ethical and legal analysis supports transparency, professional accountability, and root-cause examination when AI contributes to error [28]. Incident reporting should therefore preserve competing explanations until evidence allows stronger attribution.

Experience with regulated human medical AI devices illustrates why postmarket failure cannot be dismissed as anomalous noise. Recall analyses have identified substantial post-clearance events and associations between

limited reported clinical validation and recall status, although observational data do not establish that missing validation caused individual recalls [29]. Public benefit-risk information for AI-enabled devices can also be incomplete [30], while newer analyses of clinical evidence and recalls retain uncertainty around several associations [31]. These studies provide safety analogues rather than veterinary regulatory rules.

The incident pathway must therefore update confidence cautiously rather than convert every adverse event into a deterministic verdict.

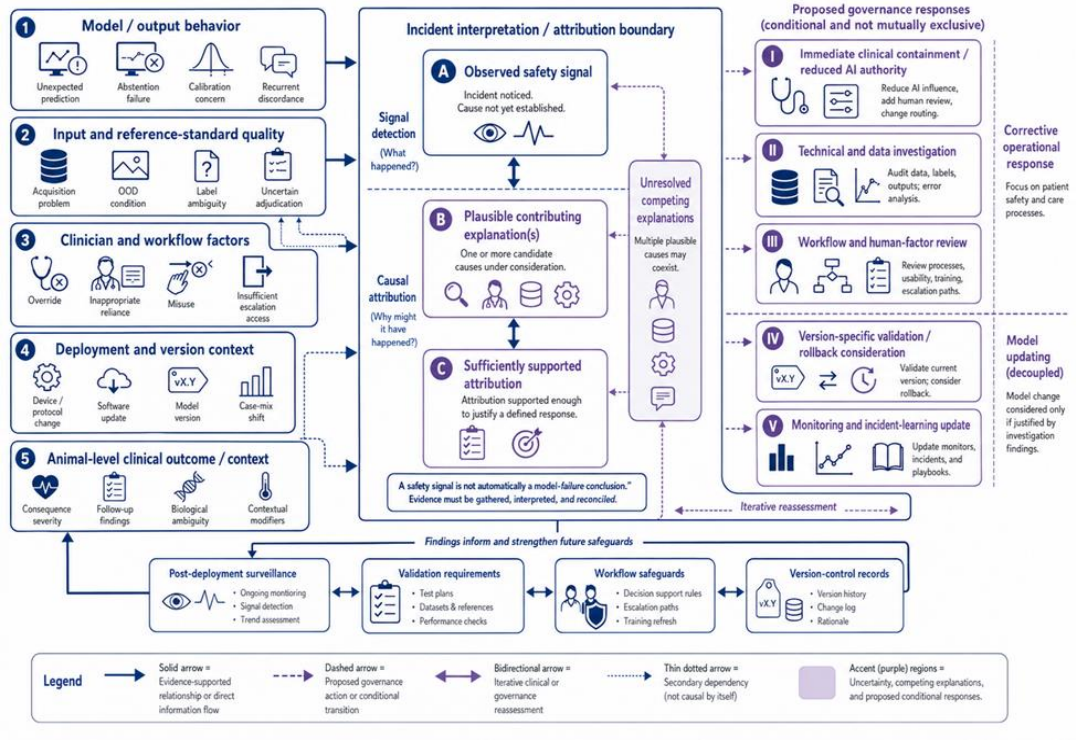

Figure 2. Incident evidence reconciliation for veterinary AI failure attribution and learning

Regulatory and ethical considerations

Veterinary AI governance differs from human medical AI because the patient cannot provide consent, clinical decisions are mediated through owners or keepers, welfare consequences may include euthanasia or major resource decisions, and professional responsibility remains with veterinary actors operating within jurisdiction-specific legal systems. Veterinary scholarship has raised concerns about epistemic authority and the possibility that AI may narrow independent reasoning when its recommendations are treated as privileged knowledge [23]. Veterinary ethical and legal analysis similarly emphasizes accountability, transparency, professional judgment, and attention to harm [28].

Human medical-device experience is informative primarily as a governance analogue. Recall and reporting studies demonstrate the value of traceable versions, documented evidence, postmarket surveillance, and structured response to failure [31]. They do not establish veterinary regulatory obligations. The proposed architecture therefore treats regulatory compliance as a minimum external constraint where applicable, while clinical governance asks the additional question of whether use remains justifiable for the particular animal, task, setting, and model version.

Validation requirements

Validation should test the conditions under which veterinary AI is expected to influence decisions, not merely reproduce a development benchmark. TRIPOD+AI requires transparent reporting of data, model development, evaluation, and intended use for prediction models incorporating AI [32]. PROBAST+AI adds structured assessment of risk of bias and applicability, reinforcing that reported performance cannot be interpreted independently of study design and target context [33].

For diagnostic systems, STARD-AI extends diagnostic-accuracy reporting to AI-specific issues including data handling, evaluation, and model information [34]. Imaging-specific CLAIM guidance likewise emphasizes

transparent intended use, reference standards, and external testing [35]. These standards support validation discipline but do not themselves prove clinical utility or safety.

For veterinary deployment, the proposed validation dossier should therefore address reference-standard quality, representative development data, external and local performance, calibration, subgroup behavior where justified, OOD and abstention behavior, human–AI interaction, escalation feasibility, monitoring procedures, and version-specific revalidation. Prospective multispecies and multicentre studies are required to determine whether the architecture improves clinical or welfare outcomes.

Conclusion

Veterinary AI should not be governed as though producing a prediction automatically confers authority to influence care. The contribution of this article is a proposed architecture in which decision authority depends on intended use, evidence applicability, calibration, uncertainty, distribution state, meaningful human oversight, and the availability of safe escalation. Abstention is therefore a clinical governance function rather than a technical defect. Monitoring, incident learning, and version control extend the same principle across the model lifecycle. The architecture remains unvalidated as an integrated intervention, and much supporting safety evidence originates outside veterinary medicine. Its value should therefore be tested prospectively across species, clinical settings, AI modalities, practitioner groups, and resource conditions before effectiveness, welfare benefit, or implementation readiness is inferred.

Acknowledgments: None

Conflict of Interest: None

Financial Support: None

Ethics Statement: None

References

1. Barillaro G, Minniti S, Mezzatesta S, Macrì F, Torino C, Vilasi AD. Artificial intelligence in veterinary diagnostic imaging: A narrative review of the state of the art and its applications. Vet J. 2026;318:106762. doi:10.1016/j.tvjl.2026.106762
2. Ndiaye YS, Cramton P, Chernev C, Ockenfels A, Schwarz T. Comparison of radiological interpretation made by veterinary radiologists and state-of-the-art commercial AI software for canine and feline radiographic studies. Front Vet Sci. 2025;12:1502790. doi:10.3389/fvets.2025.1502790
3. Brundage D. A systematic audit of transparency and validation disclosure in commercial veterinary artificial intelligence. Front Vet Sci. 2026;13:1761038. doi:10.3389/fvets.2026.1761038
4. Joslyn S, Alexander K. Evaluating artificial intelligence algorithms for use in veterinary radiology. Vet Radiol Ultrasound. 2022;63 Suppl 1:871-9. doi:10.1111/vru.13159
5. Kelly CJ, Karthikesalingam A, Suleyman M, Corrado G, King D. Key challenges for delivering clinical impact with artificial intelligence. BMC Med. 2019;17(1):195. doi:10.1186/s12916-019-1426-2
6. Vasey B, Nagendran M, Campbell B, Clifton DA, Collins GS, Denaxas S, et al. Reporting guideline for the early-stage clinical evaluation of decision support systems driven by artificial intelligence: DECIDE-AI. Nat Med. 2022;28(5):924-33. doi:10.1038/s41591-022-01772-9
7. Lekadir K, Frangi AF, Porras AR, Glocker B, Cintas C, Langlotz CP, et al. FUTURE-AI: International consensus guideline for trustworthy and deployable artificial intelligence in healthcare. BMJ. 2025;388:e081554. doi:10.1136/bmj-2024-081554
8. Shi J, Zhang K, Guo C, Yang Y, Xu Y, Wu J. A survey of label-noise deep learning for medical image analysis. Med Image Anal. 2024;95:103166. doi:10.1016/j.media.2024.103166
9. Abdalla M, Fine B. Hurdles to artificial intelligence deployment: Noise in schemas and “gold” labels. Radiol Artif Intell. 2023;5(2):e220056. doi:10.1148/ryai.220056
10. Hennessey E, DiFazio M, Hennessey R, Cassel N. Artificial intelligence in veterinary diagnostic imaging: A literature review. Vet Radiol Ultrasound. 2022;63 Suppl 1:851-70. doi:10.1111/vru.13163

11. Koch LM, Baumgartner CF, Berens P. Distribution shift detection for the postmarket surveillance of medical AI algorithms: A retrospective simulation study. NPJ Digit Med. 2024;7(1):120. doi:10.1038/s41746-024-01085-w
12. Kompa B, Snoek J, Beam AL. Second opinion needed: Communicating uncertainty in medical machine learning. NPJ Digit Med. 2021;4(1):4. doi:10.1038/s41746-020-00367-3
13. Ong Ly C, Unnikrishnan B, Tadic T, Patel T, Duhamel J, Kandel S, et al. Shortcut learning in medical AI hinders generalization: Method for estimating AI model generalization without external data. NPJ Digit Med. 2024;7(1):124. doi:10.1038/s41746-024-01118-4
14. Penso C, Frenkel L, Goldberger J. Confidence calibration of a medical imaging classification system that is robust to label noise. IEEE Trans Med Imaging. 2024;43(6):2050-60. doi:10.1109/TMI.2024.3353762
15. Huang Y, Li W, Macheret F, Gabriel RA, Ohno-Machado L. A tutorial on calibration measurements and calibration models for clinical prediction models. J Am Med Inform Assoc. 2020;27(4):621-33. doi:10.1093/jamia/ocz228
16. Van Calster B, McLernon DJ, van Smeden M, Wynants L, Steyerberg EW; Topic Group 'Evaluating diagnostic tests and prediction models' of the STRATOS initiative. Calibration: The Achilles heel of predictive analytics. BMC Med. 2019;17(1):230. doi:10.1186/s12916-019-1466-7
17. Gaube S, Suresh H, Raue M, Merritt A, Berkowitz SJ, Lermer E, et al. Do as AI say: Susceptibility in deployment of clinical decision-aids. NPJ Digit Med. 2021;4(1):31. doi:10.1038/s41746-021-00385-9
18. Jacobs M, Pradier MF, McCoy TH Jr, Perlis RH, Doshi-Velez F, Gajos KZ. How machine-learning recommendations influence clinician treatment selections: The example of antidepressant selection. Transl Psychiatry. 2021;11(1):108. doi:10.1038/s41398-021-01224-x
19. Yu F, Moehring A, Banerjee O, Salz T, Agarwal N, Rajpurkar P. Heterogeneity and predictors of the effects of AI assistance on radiologists. Nat Med. 2024;30(3):837-49. doi:10.1038/s41591-024-02850-w
20. Swaminathan A, Lopez I, Wang W, Srivastava U, Tran E, Bhargava-Shah A, et al. Selective prediction for extracting unstructured clinical data. J Am Med Inform Assoc. 2024;31(1):188-97. doi:10.1093/jamia/ocad182
21. Cheng Y, Azouzi N, Laurent-Bellue A, Guo Z, Chung T, Zeng Q, et al. A confidence-based, artificial intelligence pathology model for diagnosis of intrahepatic cholangiocarcinoma. Ann Oncol. 2026;37(7):974-85. doi:10.1016/j.annonc.2026.02.018
22. Kim YT, Kim H, Bahl M, Lev MH, González RG, Gee MS, et al. Defining operational safety in clinical artificial intelligence systems. NPJ Digit Med. 2026;9(1):281. doi:10.1038/s41746-026-02450-7
23. Eryol A. Artificial intelligence, epistemic authority, and emerging risks in veterinary clinical decision-making. Front Vet Sci. 2026;13:1861384. doi:10.3389/fvets.2026.1861384
24. Rohren E, Ahmadzade M, Colella S, Kottler N, Krishnan S, Poff J, et al. Post-deployment monitoring of AI performance in intracranial hemorrhage detection by ChatGPT. Acad Radiol. 2025;32(10):6104-13. doi:10.1016/j.acra.2025.07.055
25. Perdomo Lampignano JA, Kale A, Kumar S, Shah D, Reddy B, Lowe DJ. Beyond validation: Operationalising post-deployment surveillance of AI medical devices in clinical practice. J Med Syst. 2026;50(1):98. doi:10.1007/s10916-026-02426-w
26. Meijerink LM, Dunias ZS, Leeuwenberg AM, de Hond AAH, Jenkins DA, Martin GP, et al. Updating methods for artificial intelligence-based clinical prediction models: A scoping review. J Clin Epidemiol. 2025;178:111636. doi:10.1016/j.jclinepi.2024.111636
27. Tanner KT, Diaz-Ordaz K, Keogh RH. Implementation of a dynamic model updating pipeline provides a systematic process for maintaining performance of prediction models. J Clin Epidemiol. 2024;175:111531. doi:10.1016/j.jclinepi.2024.111531
28. Cohen EB, Gordon IK. First, do no harm. Ethical and legal issues of artificial intelligence and machine learning in veterinary radiology and radiation oncology. Vet Radiol Ultrasound. 2022;63 Suppl 1:840-50. doi:10.1111/vru.13171
29. Lee B, Kramer P, Sandri S, Chanda R, Favorito C, Nasef O, et al. Early recalls and clinical validation gaps in artificial intelligence-enabled medical devices. JAMA Health Forum. 2025;6(8):e253172. doi:10.1001/jamahealthforum.2025.3172

30. Lin JC, Jain B, Iyer JM, Rola I, Srinivasan AR, Kang C, et al. Benefit-risk reporting for FDA-cleared artificial intelligence-enabled medical devices. JAMA Health Forum. 2025;6(9):e253351. doi:10.1001/jamahealthforum.2025.3351
31. Ren Y, Zheng Y, Windecker D, Fraser AG, Siontis GCM, Caiani EG. Clinical evidence and FDA recalls of artificial intelligence-enabled medical devices. JAMA Netw Open. 2026;9(6):e2617920. doi:10.1001/jamanetworkopen.2026.17920
32. Collins GS, Moons KGM, Dhiman P, Riley RD, Beam AL, Van Calster B, et al. TRIPOD+AI statement: Updated guidance for reporting clinical prediction models that use regression or machine learning methods. BMJ. 2024;385:e078378. doi:10.1136/bmj-2023-078378
33. Moons KGM, Damen JAA, Kaul T, Hooft L, Andaur Navarro C, Dhiman P, et al. PROBAST+AI: An updated quality, risk of bias, and applicability assessment tool for prediction models using regression or artificial intelligence methods. BMJ. 2025;388:e082505. doi:10.1136/bmj-2024-082505
34. Sounderajah V, Guni A, Liu X, Collins GS, Karthikesalingam A, Markar SR, et al. The STARD-AI reporting guideline for diagnostic accuracy studies using artificial intelligence. Nat Med. 2025;31(10):3283-9. doi:10.1038/s41591-025-03953-8
35. Tejani AS, Klontzas ME, Gatti AA, Mongan JT, Moy L, Park SH, et al. Checklist for Artificial Intelligence in Medical Imaging (CLAIM): 2024 update. Radiol Artif Intell. 2024;6(4):e240300. doi:10.1148/ryai.240300