**Pharmacophore**

ISSN-2229-5402

Journal home page: http://www.pharmacophorejournal.com

# From Toxicity Scores to Defensible Safety Judgments: A Conceptual Framework for Integrating Exposure, Biological Mechanisms, Species Differences, and Uncertainty in Artificial Intelligence–Supported Toxicology

**Ana Silva[1]*, Mohamed Al-Farsi[2], Luca Moretti[3], Grace Okafor[4]**

1. ***Department of Computational Toxicology and AI-Based Safety Assessment, University of Lisbon, Lisbon, Portugal.***
2. ***Department of Toxicological Sciences and Risk Assessment, Sultan Qaboos University, Muscat, Oman.***
3. ***Department of Pharmaceutical Toxicology and Mechanistic Safety, University of Milan, Milan, Italy.***
4. ***Department of Regulatory Toxicology and Safety Evaluation, University of Ibadan, Ibadan, Nigeria.***

**ARTICLE INFO**

**Received:**
08 June 2026
**Received** in revised form:
17 October 2026
**Accepted:**
20 October 2026
**Available online:**
28 October 2026

***Keywords:*** Artificial intelligence, Toxicology, Safety assessment, Exposure, Species differences, Uncertainty

**ABSTRACT**

Artificial intelligence is increasingly used to predict toxicological endpoints, prioritize chemicals, integrate high-dimensional biological data, and support alternatives to conventional testing. Yet a toxicity score is not itself a safety judgment. A predictive output usually characterizes the probability, class, or magnitude of a hazard-related endpoint under a particular model and data domain, whereas safety assessment requires an inference about biological significance under specified exposure conditions, mechanistic plausibility, species relevance, population variability, uncertainty, and decision context. This article develops a conceptual account of how those evidentiary dimensions should be combined without converting predictive performance into unwarranted claims of safety or harm. The analysis distinguishes hazard prediction from exposure-conditioned interpretation, examines how mechanistic evidence and cross-species biology can strengthen or weaken transfer, and treats uncertainty as a structured property of the evidence rather than a residual confidence score. The resulting approach positions AI output as one evidentiary input within an iterative safety judgment in which concordant evidence may increase confidence, discordant evidence may redirect interpretation, and unresolved limitations may block downstream inference. The central implication is that defensible AI-supported toxicology depends less on whether a model produces an accurate toxicity score in isolation than on whether the complete evidentiary configuration justifies the decision being made.

*This is an **open-access** article distributed under the terms of the Creative Commons Attribution-Non Commercial-Share Alike 4.0 License, which allows others to remix, and build upon the work non commercially.*

**To Cite This Article:** Silva A, Al-Farsi M, Moretti L, Okafor G. From Toxicity Scores to Defensible Safety Judgments: A Conceptual Framework for Integrating Exposure, Biological Mechanisms, Species Differences, and Uncertainty in Artificial Intelligence–Supported Toxicology. Pharmacophore. 2026;17(5):151-8. https://doi.org/10.51847/94Dy3t5G9p

## Introduction

Artificial intelligence is reshaping toxicology by extending the scale at which chemical, biological, textual, imaging, omics, and other heterogeneous data can be analyzed. Contemporary applications include endpoint prediction, chemical prioritization, mechanistic pattern recognition, data integration, and support for new approach methodologies (NAMs). These capabilities may reduce dependence on slow or poorly scalable testing strategies and may reveal relationships that are difficult to detect through conventional analyses. Their usefulness, however, remains conditioned by data quality, transparency, interpretability, bias, applicability, and the evidentiary meaning of the predicted endpoint itself [1].

The category "AI-supported toxicology" also contains materially different analytical tasks. Machine-learning systems may predict categorical toxicity labels, continuous biological responses, pharmacokinetic parameters, molecular interactions, histopathological features, or latent representations derived from high-dimensional data. Model architectures, training datasets,

**Corresponding Author:** Ana Silva; Department of Computational Toxicology and AI-Based Safety Assessment, University of Lisbon, Lisbon, Portugal. E-mail: ana.silva@ulisboa.pt.

validation strategies, and intended uses vary accordingly [2]. A prediction therefore cannot be interpreted solely from the numerical form of its output. Its evidentiary value depends on what was measured, how the endpoint relates to adverse biology, which population or experimental system is represented, and whether the model is being used for screening, prioritization, mechanistic interpretation, quantitative risk assessment, or another decision.

Recent examination of machine-learning practice in toxicology reinforces this distinction. Reporting remains heterogeneous, and important details concerning data processing, interpretability, reproducibility, validation, and model implementation are not consistently documented [3]. Such gaps do not mean that machine learning lacks toxicological value. They mean that confidence in an output must incorporate the conditions under which the output was produced and the limits on what it can support.

This problem becomes more consequential when AI moves toward regulatory or other high-stakes applications. Regulatory science requires evidence that can be interrogated, reproduced, qualified for its intended use, and integrated with established scientific reasoning. Trust therefore depends on more than technical sophistication or predictive discrimination; it also depends on testability, transparency, appropriate human expertise, and the capacity to revise judgments as evidence changes [4]. These requirements are especially important when a model prediction is several inferential steps removed from the biological or regulatory proposition ultimately at issue.

Problem formulation provides the first control on that inferential distance. In-silico methods can support risk assessment only when the decision problem, relevant endpoint, intended population, exposure context, uncertainties, and consequences of error are made explicit [5]. The same predicted signal may be useful for prioritizing chemicals for additional testing while being insufficient to establish a health-protective exposure conclusion. Conversely, a modest hazard signal may become highly consequential when exposure is substantial, the mechanism is biologically plausible, and relevant human variability is poorly bounded.

Scientific confidence in NAMs is therefore better understood as context dependent than as a universal property attached to a method [6]. For the present analysis, a safety judgment is a bounded inference about whether available evidence supports a decision concerning potential harm under specified biological, exposure, population, and decision conditions. This definition separates the object of prediction from the object of judgment. The central question is consequently not whether an AI system can predict toxicity, but when and how a predicted toxicity signal can legitimately contribute to a defensible conclusion about safety.

*Why Toxicity Prediction Is Not Equivalent to Safety Assessment*

Toxicity prediction usually estimates an endpoint defined by the training data: for example, the likelihood of drug-induced liver injury, cardiotoxicity, mutagenicity, or another adverse response. Improvements in chemical representation and knowledge integration can increase predictive performance. Quantitative knowledge–activity approaches, for instance, broaden the informational basis available to toxicity models beyond conventional structure-centred representations [7]. Such advances improve the quality of prediction within the modeled task, but the modeled task remains narrower than the safety question unless exposure, biological interpretation, translation, and uncertainty have also been addressed.

The first missing inference concerns exposure. A chemical may possess biological activity at concentrations that are irrelevant to realistic human exposure, while a weaker intrinsic hazard may be important when internal exposure is sustained or concentrated in a sensitive tissue. Comparisons between in-vitro bioactivity and in-vivo adverse-effect levels illustrate why bioactivity may provide useful screening-level information without independently specifying a safe or unsafe human exposure [8]. A toxicity score that is detached from plausible dose therefore describes hazard-related potential more readily than actual risk.

Integrated NAM programmes have begun to make that distinction operational. In data-poor chemical assessment, high-throughput biological activity can be interpreted together with exposure estimates and other evidence to support prioritization and identify situations requiring greater scrutiny [9]. The informational gain comes from combining evidence streams rather than treating a single biological response as a final verdict. The resulting judgment remains purpose-specific: evidence adequate to prioritize a substance may remain inadequate for a regulatory conclusion with greater consequences.

Internal exposure sharpens the same principle. Systemic safety approaches that compare modeled internal exposure with concentrations associated with biological activity can help identify whether the margin between exposure and bioactivity is reassuring or concerning [10]. Yet such ratios inherit uncertainty from both sides of the comparison. Assay coverage may omit relevant biology; kinetic models may incompletely represent absorption, distribution, metabolism, or elimination; and biological activity may not correspond directly to an adverse outcome. The relationship is therefore interpretive rather than mechanically determinative.

A broader computational-toxicology architecture reaches the same conclusion from a different direction. Regulatory assessment increasingly combines chemical characterization, hazard information, exposure estimates, toxicokinetics, population variability, mechanistic evidence, and explicit treatment of uncertainty [11]. Within that architecture, prediction has an important but delimited role. It can fill data gaps, identify patterns, estimate endpoints, or direct additional investigation. It cannot determine by itself whether a predicted effect will occur at a relevant exposure, whether the modeled mechanism applies to humans, or whether uncertainty is acceptable for the decision being made.

These inferential gaps can be organized according to what a toxicity score establishes, what remains unresolved, and what additional evidence can legitimately change the safety interpretation.

**Table 1** summarizes inferential gaps between a toxicity-score output and a defensible safety judgment.

**Table 1.** Inferential gaps between a toxicity-score output and a defensible safety judgment

| Toxicity-score proposition | What the output can establish | Unresolved safety question | Additional evidence that can change interpretation | Consequence for judgment |
|---|---|---|---|---|
| **Predicted probability or class of a toxic endpoint** | Model-specific expectation for the represented endpoint | Whether the endpoint is biologically relevant under the intended use | Endpoint definition, assay relevance, mechanistic evidence | Retain as hazard evidence; do not elevate automatically to a safety conclusion |
| **Strong predicted biological activity** | Potential for perturbation within the modeled domain | Whether relevant internal concentrations are reached | External exposure, toxicokinetics, internal-dose modeling | Concern may strengthen, weaken, or become negligible depending on exposure |
| **High-performing model result** | Agreement with a defined validation dataset | Whether the model generalizes to the present chemical, population, species, or context | Applicability-domain analysis, external validation, domain-specific evidence | Confidence is bounded by transportability rather than headline performance |
| **Concordant signals across predictive methods** | Cross-model agreement | Whether the shared signal reflects a relevant adverse mechanism | Orthogonal mechanistic and biological evidence | Concordance can increase confidence but cannot substitute for biological interpretation |
| **Predicted hazard in a nonhuman system** | Potential concern in the represented species or model | Whether the response transfers to humans | Species-relevant mechanism, comparative biology, metabolism and response-capacity evidence | Translation may be supported, qualified, redirected, or blocked |
| **Apparently low predicted hazard** | Low signal within measured/modelled endpoints | Whether relevant hazards are absent or simply unrepresented | Assay coverage, alternative mechanisms, uncertainty and bias assessment | Absence of predicted signal is not equivalent to demonstrated safety |
| **Quantitative model output** | Numerical estimate produced by a specified model | Whether uncertainty is acceptable for the consequence of the decision | Parameter uncertainty, model uncertainty, variability and evidence quality | Numerical precision must not be confused with evidentiary certainty |

The analytical implication is that prediction and safety assessment answer different questions. Prediction asks what a model expects for a defined endpoint under its representation of the problem. Safety assessment asks whether the available evidence supports a consequential inference about harm under specified conditions. Moving from the first question to the second requires additional evidence, and the direction of that evidence cannot be assumed in advance.

*Integrating Exposure, Mechanism, Species and Uncertainty*

Exposure is the first major modifier because biological hazard becomes decision-relevant through dose, route, duration, timing, distribution, and internal concentration. Physiologically based pharmacokinetic modeling provides a mechanistic means of connecting administered exposure to concentrations in tissues over time, while machine-learning methods can support parameter estimation and model development [12]. These tools are valuable precisely because nominal exposure and biologically effective exposure are not interchangeable. Their outputs nevertheless remain model-based and inherit uncertainty from physiological assumptions, parameter quality, population representation, and chemical domain.

In-vitro to in-vivo extrapolation extends this reasoning by translating biological responses observed in experimental systems toward concentrations or doses relevant to intact organisms. AI-supported IVIVE can use complex toxicogenomic information to improve cross-domain translation [13]. Empirical applications combining toxicogenomic dose-response information with pregnancy PBPK modeling further illustrate how biological activity can be interpreted in relation to human exposure rather than considered in isolation [14]. The broader lesson is not that IVIVE yields a universal safety threshold; it is that predicted hazard gains or loses significance when the exposure required to produce the response is compared with plausible internal exposure in the population of interest.

Mechanistic evidence provides a second modifier. Toxicogenomic signals become more informative when they can be situated within biologically coherent processes, including adverse outcome pathway structures that connect molecular initiating events with downstream biological changes [15]. Mechanistic concordance can strengthen a toxicity interpretation when predicted activity, observed pathway perturbation, and the adverse phenotype are mutually compatible. Discordance can have the opposite effect. A statistical prediction that lacks a plausible route from initiating interaction to relevant adverse outcome should carry a different evidentiary status from a prediction supported by orthogonal mechanistic evidence.

Mechanistic reasoning is also central to species translation. Cross-species differences are not adequately represented by treating “species” as a fixed correction factor. Experimental systems show that comparable toxic insult can generate substantially different hepatic responses because species differ in stress-response capacity, compensatory biology, metabolism, and downstream signaling [16]. More specific mechanistic work demonstrates that species differences in constitutive androstane receptor biology can arise from differential interaction with YAP, altering proliferative responses between rodent and human systems [17]. Likewise, species-resolved liver spheroids and transcriptomic analyses have been used to investigate a dog-specific hepatotoxicity signal and to distinguish mechanisms that were not reproduced in the human system [18]. These

studies do not establish a universal rule that animal findings lack human relevance. They show why the relevance of an animal signal depends on the biological mechanism that generates it.

Computational modeling can respond to that problem by representing biological species information explicitly. Cross-species toxicity models that combine chemical descriptors with biological information demonstrate that predictive performance can benefit when organismal differences are treated as explanatory features rather than ignored variation [19]. Although evidence from ecotoxicological modeling cannot be transferred directly to human pharmaceutical safety, the computational principle is pertinent: species relevance should be modeled, investigated, or justified rather than assumed.

Human variability introduces a further layer after species translation. Population-based human in-vitro systems can provide empirical estimates of inter-individual variation in biological response [20]. Such evidence is important because a prediction calibrated to an average response may be insufficient for decisions intended to protect heterogeneous populations. Variability may reflect genetic background, baseline physiology, metabolic capacity, disease state, life stage, co-exposures, or other modifiers. The evidentiary problem is therefore not solved simply by moving from animals to a nominally "human" system.

This distinction also changes how uncertainty should be handled. Conventional uncertainty factors provide protection when direct evidence is incomplete, but emerging NAM data create opportunities to replace some default assumptions with empirically informed estimates when the evidence is sufficiently robust [21]. The appropriate conceptual distinction is between variability, which may sometimes be characterized with data, and uncertainty, which represents incomplete knowledge about the model, mechanism, parameters, relevance, or unmeasured alternatives. Combining the two into a single confidence score would conceal rather than resolve the safety problem.

The interaction among these dimensions determines whether a predicted toxicity signal becomes more or less credible as evidence moves toward a human safety decision. Exposure determines whether relevant biological concentrations are plausible; mechanism determines whether the predicted effect has a coherent path to adversity; species evidence determines whether that path transfers across biological systems; and uncertainty determines how strongly any conclusion should be stated. Their roles are interdependent rather than additive. Strong mechanistic evidence cannot compensate for an exposure scenario that is biologically implausible, while apparently favorable exposure margins may be less reassuring if the kinetic model excludes an important metabolite or susceptible subgroup.

The resulting interpretive conditions can be differentiated without assigning unsupported numerical weights or universal thresholds.

**Table 2** summarizes conditions that alter translation of AI-supported toxicity evidence across exposure, mechanism, species and uncertainty.

**Table 2.** Conditions that alter translation of AI-supported toxicity evidence across exposure, mechanism, species and uncertainty

| Interpretive condition | Evidence question | Finding that can strengthen transfer | Finding that can weaken or redirect transfer | Residual uncertainty that must remain visible | Safety-judgment implication |
|---|---|---|---|---|---|
| **Exposure plausibility** | Can the predicted toxic process occur at plausible internal exposure? | Modeled or measured internal concentrations overlap biologically active ranges under the relevant scenario | Required concentrations are implausible under credible exposure assumptions | Kinetic parameters, route, duration, metabolites, tissue distribution | Hazard signal becomes more or less decision-relevant according to exposure |
| **Mechanistic coherence** | Is there a biologically plausible route from predicted activity to adversity? | Orthogonal evidence supports linked molecular, cellular and functional events | Predicted endpoint lacks mechanistic concordance or competing mechanisms better explain the signal | Incomplete pathway coverage and causal uncertainty | Mechanistic evidence modifies confidence rather than replacing outcome evidence |
| **Species relevance** | Does the mechanism operate similarly in the target species? | Conserved target, metabolism, signaling and response capacity | Species-specific metabolism, receptor interaction, compensation or stress response changes the pathway | Incomplete comparative biology | Translation may be supported, qualified or blocked |
| **Human-system relevance** | Does the evidence represent meaningful human biology? | Human-derived systems reproduce relevant exposure–response and mechanism | Model lacks required cell types, physiology, chronicity or systemic interactions | Model fidelity and population representativeness | Human-derived evidence strengthens relevance but does not automatically establish clinical effect |
| **Population variability** | Could susceptible individuals experience materially different internal exposure or response? | Population-based data characterize chemical-specific response distributions | Evidence is derived from a narrow biological background or omits major susceptibility factors | Unmeasured genetics, disease, age, co-medication and co-exposure | Protective interpretation must account for heterogeneity |
| **Model applicability** | Is the chemical and biological problem | Target case is well represented by training | Extrapolation extends beyond chemical, | Unknown extrapolation error | Prediction should be downgraded when |

| | | | | | |
|---|---|---|---|---|---|
| | represented within the model's learned domain? | chemistry, biology and endpoint definition | mechanistic or population domain | | transportability is uncertain |
| **Evidence concordance** | Do independent evidence streams point toward the same interpretation? | Exposure, mechanism, species relevance and biological response converge | Evidence streams conflict or support competing explanations | Dependence among evidence streams and missing alternatives | Concordance may increase confidence; discordance triggers reassessment |
| **Uncertainty sufficiency** | Is remaining uncertainty acceptable for the consequence of the decision? | Key uncertainty sources are characterized and decision consequences are limited | Important mechanisms, exposures, populations or validation domains remain unresolved | Irreducible and epistemic uncertainty | Greater decision consequence requires stronger evidentiary support |

The core implication of this integration is that safety evidence should be interpreted configurationally. The same toxicity prediction can justifiably lead to different levels of concern under different exposure, mechanistic, species, and uncertainty conditions. Those differences do not reflect inconsistency in the assessment; they reflect the fact that safety is a contextual judgment about the relationship between hazard and the circumstances under which harm may occur. The next analytical task is therefore to translate these dimensions into a decision structure that specifies when AI-supported evidence can support progression, when it requires qualification or additional investigation, and when transfer toward a safety conclusion should be blocked.

*Proposed AI-Supported Safety-Judgment Framework*

The preceding analysis implies that an AI-supported safety judgment should begin with a predicted toxicological signal but should not preserve that signal unchanged as evidence moves toward a decision. Instead, the prediction must be repeatedly reinterpreted as additional information becomes available. Exposure asks whether biologically relevant concentrations are plausible. Mechanistic evidence asks whether the predicted response connects coherently to adversity. Species evidence asks whether that mechanism is transferable to the target population. Variability and uncertainty determine how confidently any resulting inference can be generalized.

Evidence convergence is especially important when translation spans multiple levels of biological organization. Analyses of central nervous system safety illustrate how mechanistic information, exposure data, refined experimental systems, conventional toxicology, and AI-supported signals can be combined to interpret findings that would remain ambiguous if any single evidence stream were treated as decisive [22]. The appropriate unit of judgment is therefore the evidentiary configuration, not the toxicity score.

This configuration should also determine whether evidence needs to be escalated. Tiered next-generation risk assessment provides an example of how toxicokinetic and toxicodynamic evidence can be introduced progressively as unresolved uncertainty increases [23]. The present framework extends that logic conceptually: additional evidence becomes necessary when disagreement among exposure, mechanism, species relevance, or uncertainty materially affects the consequence of a safety decision.

Regulatory purpose provides the final boundary. Frameworks developed for integrating NAMs into chemical safety assessment show that evidence sufficiency depends on the question being asked and the regulatory use for which the evidence is intended [24]. Consequently, AI-supported evidence should not yield an unqualified binary label of "safe" or "unsafe." It should instead support a bounded decision state reflecting how much of the inferential path from predicted hazard to decision-relevant safety has been justified.

Four qualitative states follow from this logic: supported progression, in which multiple relevant evidence streams converge; qualified progression, in which the direction of the judgment is supportable but important limitations remain; reassessment, in which discordant evidence or alternative explanations require reinterpretation; and blocked transfer, in which missing or contradictory evidence prevents the toxicity prediction from licensing the downstream safety claim.

The relationship among these states can be expressed without assigning numerical weights or universal thresholds.

**Table 3** summarizes safety-decision states generated by different configurations of hazard, exposure, mechanism, species relevance and uncertainty.

**Table 3.** Safety-decision states generated by different configurations of hazard, exposure, mechanism, species relevance and uncertainty

| Safety-decision state | Hazard prediction | Exposure interpretation | Mechanistic evidence | Species/human relevance | Uncertainty condition | Permitted decision interpretation |
|---|---|---|---|---|---|---|
| **Supported progression** | Relevant and technically credible signal | Plausible internal exposure is characterized | Mechanism is coherent and independently supported | Transfer is biologically plausible | Major uncertainties are characterized and proportionate to the decision | Prediction may contribute materially to the safety judgment |

| | | | | | | |
|---|---|---|---|---|---|---|
| **Qualified progression** | Directionally informative signal | Exposure relationship is plausible but incompletely resolved | Mechanism is partly supported | Translation is plausible but incompletely demonstrated | Residual uncertainty remains consequential but bounded | Judgment may proceed only with explicit qualification |
| **Reassessment** | Signal exists but competes with discordant evidence | Exposure assumptions are unstable or scenario-dependent | Alternative mechanisms remain plausible | Cross-species evidence is inconsistent | Important uncertainty changes interpretation | Prediction should trigger additional analysis rather than immediate progression |
| **Blocked transfer** | Signal may be technically valid within its model | Relevant exposure cannot be established or model is inapplicable | Mechanistic connection to adversity is unsupported | Human/species relevance is unsubstantiated or contradicted | Uncertainty is too large for the decision consequence | Prediction must not be used to support the downstream safety claim |

The framework therefore treats safety judgment as a sequence of conditional evidentiary transfers rather than a conversion of model output into a categorical conclusion. The essential distinction is whether each transfer is supported strongly enough for the decision context.

**Figure 1** depicts when a predicted toxicological signal becomes decision-relevant.

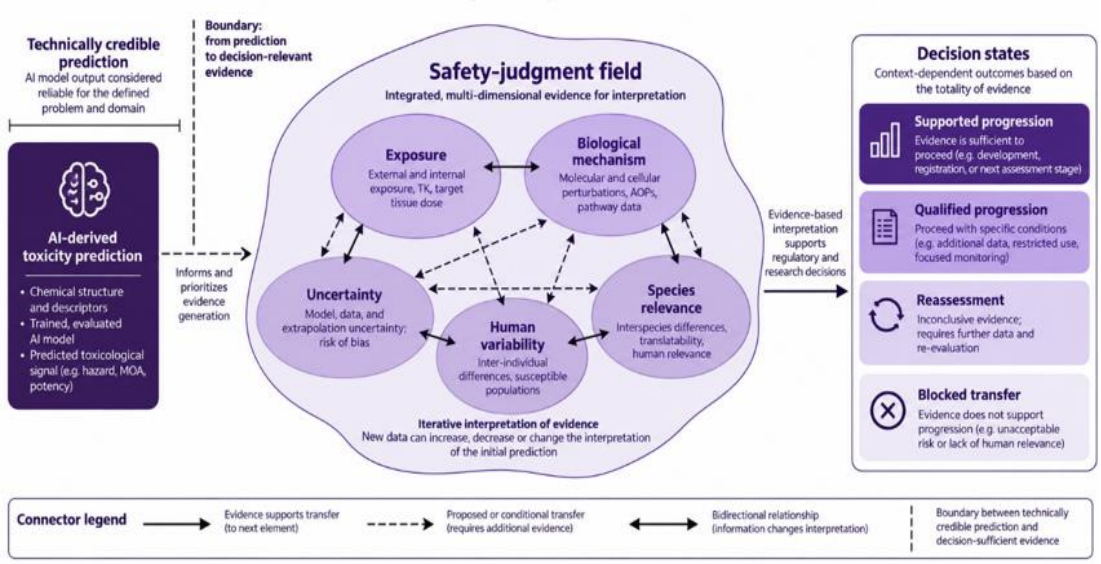



**Figure 1.** When a predicted toxicological signal becomes decision-relevant

*Validation and Regulatory Implications*

A conceptual safety-judgment framework is useful only if the evidence entering it can itself be trusted. Validation of AI-based NAMs therefore requires more than demonstration of discrimination or prediction accuracy. Data provenance, reference standards, model transparency, interpretability, applicability, generalizability, and the relationship between the model endpoint and the intended regulatory question all affect scientific confidence [25]. Validation should consequently be matched to the claim being made: a model validated for prioritization cannot be assumed to be validated for a higher-consequence safety decision.

Reproducibility introduces a separate requirement. Regulatory toxicology depends on evidence that survives changes in laboratory, operator, protocol implementation, and other conditions relevant to transfer. Interlaboratory ring trials remain important because strong performance in one setting does not establish robustness across settings [26]. For AI-supported approaches, analogous concerns extend to software versioning, preprocessing, training-data provenance, model updating, and the stability of prediction under reasonable analytical variation.

Evidence quality must also be separated from model performance. Risk of bias can arise in study design, data generation, selection, reporting, model construction, or evaluation, and sophisticated analytical methods do not remove those vulnerabilities [27]. A high-performing model trained on biased or poorly representative evidence may produce precise predictions that remain unsuitable for the intended safety inference.

Finally, implementation occurs within regulatory systems rather than in an abstract technical environment. Experience with NAM adoption indicates that scientific validity, institutional processes, expertise, incentives, communication, legal expectations, and organizational capacity interact to enable or inhibit regulatory use [28]. Scientific confidence and implementation readiness are therefore related but non-identical.

The regulatory consequence is that progression from predicted toxicity toward an actionable safety judgment requires several transfers to remain defensible at once: the model must be credible for its intended use; the biological interpretation must be supported; exposure and species relevance must be adequate for the target context; residual uncertainty must be proportionate to the consequence of error; and the evidence must remain reproducible and auditable. Failure at any one of these points may justify additional testing, reinterpretation, or refusal to transfer the prediction into the proposed decision.

**Figure 2** depicts where translational evidence supports, weakens or blocks movement toward regulatory safety judgment.

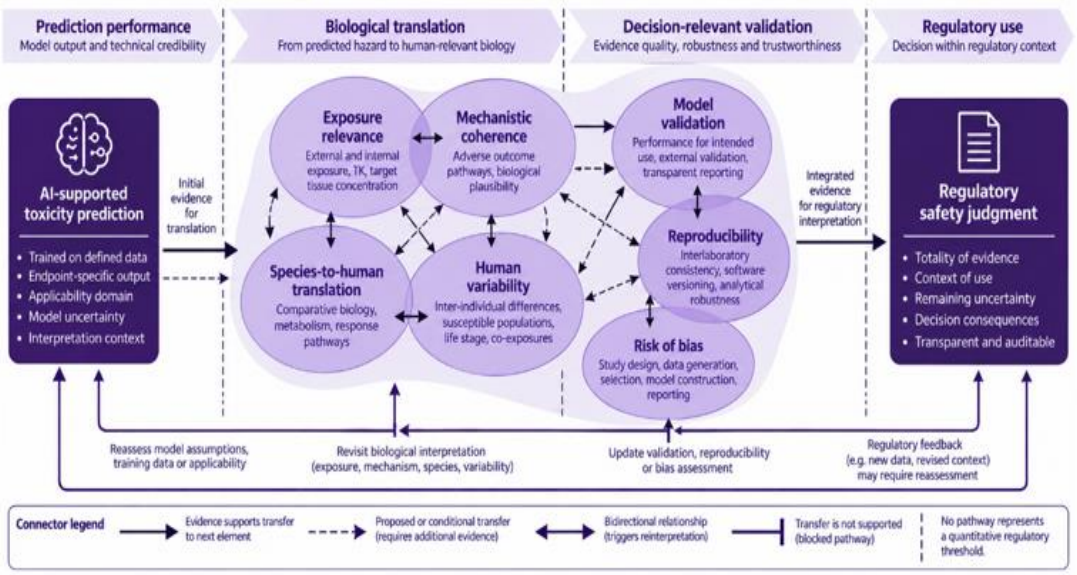



**Figure 2.** Where translational evidence supports, weakens or blocks movement toward regulatory safety judgment

**Conclusion**

AI-supported toxicology becomes scientifically defensible when toxicity prediction is treated as the beginning of a safety inference rather than its conclusion. A model may characterize hazard with considerable technical accuracy while leaving unresolved whether relevant exposure occurs, whether the predicted biology is mechanistically coherent, whether findings transfer across species, whether human variability has been represented, and whether remaining uncertainty is acceptable for the decision at hand. These dimensions determine whether a prediction can support progression, requires qualification or reassessment, or should be prevented from licensing a downstream safety claim.

The resulting standard is deliberately conditional. AI can strengthen safety assessment by integrating evidence and exposing relationships that conventional approaches may overlook, but the authority of its outputs must remain proportional to the evidence connecting prediction to biological and regulatory context. Defensible safety judgment therefore rests on transparent, reproducible and revisable evidentiary transfer—not on the toxicity score alone.

**Acknowledgments:** None

**Conflict of interest:** None

**Financial support:** None

**Ethics statement:** None

**References**

1. Kleinstreuer N, Hartung T. Artificial intelligence (AI)-it's the end of the tox as we know it (and I feel fine). Arch Toxicol. 2024;98(3):735-54. doi:10.1007/s00204-023-03666-2
2. Lin Z, Chou WC. Machine learning and artificial intelligence in toxicological sciences. Toxicol Sci. 2022;189(1):7-19. doi:10.1093/toxsci/kfac075
3. Kappenberg F, Stolte M, Sauer L, Duda JC, Lau M, Schürmeyer L, et al. A review of machine learning in toxicology: Current practices and reporting gaps. Arch Toxicol. 2026;100(8):3491-505. doi:10.1007/s00204-026-04393-0
4. Hartung T, Whelan M, Tong W, Califf RM. Is regulatory science ready for artificial intelligence? NPJ Digit Med. 2025;8(1):200. doi:10.1038/s41746-025-01596-0
5. Achar J, Cronin MTD, Firman JW, Öberg G. A problem formulation framework for the application of in silico toxicology methods in chemical risk assessment. Arch Toxicol. 2024;98(6):1727-40. doi:10.1007/s00204-024-03721-6

6. van der Zalm AJ, Barroso J, Browne P, Casey W, Gordon J, Henry TR, et al. A framework for establishing scientific confidence in new approach methodologies. Arch Toxicol. 2022;96(11):2865-79. doi:10.1007/s00204-022-03365-4
7. Li T, Qu Y, Chen A, Thakkar S, Li D, Tong W. Beyond QSARs: Quantitative Knowledge–Activity Relationships (QKARs) for enhanced drug toxicity prediction. Toxicol Sci. 2025;208(2):269-78. doi:10.1093/toxsci/kfaf135
8. Paul Friedman K, Gagne M, Loo LH, Karamertzanis P, Netzeva T, Sobanski T, et al. Utility of in vitro bioactivity as a lower bound estimate of in vivo adverse effect levels and in risk-based prioritization. Toxicol Sci. 2020;173(1):202-25. doi:10.1093/toxsci/kfz201
9. Paul Friedman K, Thomas RS, Wambaugh JF, Harrill JA, Judson RS, Shafer TJ, et al. Integration of new approach methods for the assessment of data-poor chemicals. Toxicol Sci. 2025;205(1):74-105. doi:10.1093/toxsci/kfaf019
10. Cable S, Baltazar MT, Bunglawala F, Carmichael PL, Contreas L, Dent MP, et al. Advancing systemic toxicity risk assessment: Evaluation of a NAM-based toolbox approach. Toxicol Sci. 2025;204(1):79-95. doi:10.1093/toxsci/kfae159
11. Thomas RS, Bahadori T, Buckley TJ, Cowden J, Deisenroth C, Dionisio KL, et al. The next generation blueprint of computational toxicology at the U.S. Environmental Protection Agency. Toxicol Sci. 2019;169(2):317-32. doi:10.1093/toxsci/kfz058
12. Chou WC, Lin Z. Machine learning and artificial intelligence in physiologically based pharmacokinetic modeling. Toxicol Sci. 2023;191(1):1-14. doi:10.1093/toxsci/kfac101
13. Chandra M, Li T, Tong W. AIVIVE: A novel AI framework for enhanced in vitro to in vivo extrapolation (IVIVE) of toxicogenomics data. Toxicol Sci. 2025;207(2):361-71. doi:10.1093/toxsci/kfaf100
14. Ni M, Deepika D, Li X, Xiong W, Zhang L, Chen J, et al. IVIVE-PBPK based new approach methodology for addressing early life toxicity induced by Bisphenol A. Environ Res. 2024;240(Pt 1):117343. doi:10.1016/j.envres.2023.117343
15. Saarimäki LA, Morikka J, Pavel A, Korpilähde S, del Giudice G, Federico A, et al. Toxicogenomics data for chemical safety assessment and development of new approach methodologies: An adverse outcome pathway-based approach. Adv Sci (Weinh). 2023;10(2):2203984. doi:10.1002/advs.202203984
16. Russomanno G, Sison-Young R, Livoti LA, Coghlan H, Jenkins RE, Kunnen SJ, et al. A systems approach reveals species differences in hepatic stress response capacity. Toxicol Sci. 2023;196(1):112-25. doi:10.1093/toxsci/kfad085
17. Shizu R, Makida N, Sobe K, Ishimura M, Takeshita A, Hosaka T, et al. Interaction with YAP underlies the species differences between humans and rodents in CAR-dependent hepatocyte proliferation. Toxicol Sci. 2024;198(1):101-12. doi:10.1093/toxsci/kfad129
18. Bajaj P, Brennan RJ, Laurent S, Sauzeat S, Dufault M, Richards B, et al. Transcriptomic analysis in liver spheroids identifies a dog-specific mechanism of hepatotoxicity for amcenestrant. Toxicol Sci. 2025;204(2):228-41. doi:10.1093/toxsci/kfaf012
19. Zubrod JP, Galic N, Vaugeois M, Dreier DA. Bio-QSARs 2.0: Unlocking a new level of predictive power for machine learning-based ecotoxicity predictions by exploiting chemical and biological information. Environ Int. 2024;186:108607. doi:10.1016/j.envint.2024.108607
20. Alshammari I, Ford LC, Tsai HHD, Lin HC, Negi CK, Dickey AN, et al. Quantitative estimates of inter-individual variability for new approach methodologies-based systemic safety toolbox using a population-based human in vitro model. Toxicol Sci. 2026;209(4):kfag038. doi:10.1093/toxsci/kfag038
21. Rusyn I, Chiu WA. Decision-making with new approach methodologies: Time to replace default uncertainty factors with data. Toxicol Sci. 2022;189(1):148-9. doi:10.1093/toxsci/kfac033
22. Behl M, Daly FS, Hogberg HT, Berridge BR, D'Agostino J, McDuffie JE, et al. Integrating new approach methodologies and artificial intelligence to advance central nervous system toxicity prediction: Lessons from preclinical and clinical case studies. Toxicol Sci. 2026;209(9):kfag100. doi:10.1093/toxsci/kfag100
23. Fernandez-Agudo A, Tarazona JV. A tiered next-generation risk assessment framework integrating toxicokinetics and NAM-based toxicodynamics: “Proof of concept” case study using pyrethroids. Arch Toxicol. 2025;99(7):2759-81. doi:10.1007/s00204-025-04045-9
24. Ball N, Bars R, Botham PA, Cuciureanu A, Cronin MTD, Doe JE, et al. A framework for chemical safety assessment incorporating new approach methodologies within REACH. Arch Toxicol. 2022;96(3):743-66. doi:10.1007/s00204-021-03215-9
25. Hartung T, Kleinstreuer N. Challenges and opportunities for validation of AI-based new approach methods. ALTEX. 2025;42(1):3-21. doi:10.14573/altex.2412291
26. Jacobs MN, Hoffmann S, Hollnagel HM, Kern P, Kolle SN, Natsch A, et al. Avoiding a reproducibility crisis in regulatory toxicology—on the fundamental role of ring trials. Arch Toxicol. 2024;98(7):2047-63. doi:10.1007/s00204-024-03736-z
27. Hartung T, Hoffmann S, Whaley P. Assessing risk of bias in toxicological studies in the era of artificial intelligence. Arch Toxicol. 2025;99(8):3065-90. doi:10.1007/s00204-025-03978-5
28. Bearth A, Kopainsky B, Jones LB, Vist GE, Husøy T, Svendsen C, et al. Exploring experiences of the regulatory toxicology system: System-level promoters and inhibitors of new approach methodologies. Arch Toxicol. 2025;99(12):4909-30. doi:10.1007/s00204-025-04168-z