

Anticipated Regulatory Issues for Code-level Surrogate Models Replacing Nuclear Safety Analysis Codes

Sihyeong Yu^a, Chaeun Jang^b, Joongoo Jeon^{a*}

^a*Division of Advanced Nuclear Engineering, POSTECH, Pohang, Republic of Korea*

^b*Department of Nuclear Engineering, Hanyang University, Seoul, Republic of Korea*

**Corresponding author: jgjeon41@postech.ac.kr*

***Keywords:** nuclear safety analysis code, code-level surrogate model, artificial intelligence, regulatory gap, verification and validation

1. Introduction

With artificial intelligence (AI) technologies being adopted across nearly every domain of daily life, numerous attempts have also been made in science and engineering to employ AI to compensate for scarce experimental data and to reduce the substantial computational cost of numerical analyses.

In Korea, the Framework Act on the Development of Artificial Intelligence and the Creation of a Foundation for Trust (the AI Basic Act) was enacted to govern the safe use of AI and has been in force since January 2026 [1]. The Act enumerates AI utilized for the safe management and operation of nuclear materials and nuclear facilities as one domain of high-impact AI (Article 2, Subparagraph 4 (e)), and imposes operator obligations such as risk management plans, explanation measures, and human oversight (Article 34). AI applications in the operational stage, such as operator support or radiation safety tasks, can therefore be captured by the Act.

For AI element technologies applied to nuclear safety analysis codes, however, the coverage of the Act is questionable. Nuclear safety analysis codes are the computational codes that analyze the accident and transient behavior of a nuclear power plant to evaluate its safety at the design and licensing stage. First, it is unclear whether the use of such codes constitutes the 'management and operation' of a facility at all. Second, even if this use were captured by the AI Basic Act, the obligations under Article 34 consist of establishing and operating plans for risk management and user protection, preparing explanation measures to the extent technically feasible, ensuring human oversight, and retaining documentation. None of these prescribes how the adequacy of an AI model for safety analysis is to be demonstrated, and they provide none of the technical acceptance criteria such as verification and validation (V&V), extrapolation capability, and numerical stability. This gap is not limited to Korea or to the level of general legislation. In the United States, where the nuclear regulatory framework itself was screened, the U.S. Nuclear Regulatory Commission (NRC) identified 71 potential regulatory gaps in its 2024 regulatory framework gap assessment of regulatory guides and regulations [2]. A regulatory methodology for safety analysis codes incorporating AI element technologies therefore needs to be developed to close this regulatory gap.

AI element technology, however, is not a single regulatory object. Depending on the scope of substitution, a spectrum exists, ranging from model-level surrogates, which replace individual constitutive models or correlations inside a code, to code-level surrogates, which replace the entire input–output relationship of the code. The nature and magnitude of the regulatory gap vary along this spectrum. For model-level surrogates, the governing equations and the numerical framework of the code are preserved, so the existing review framework for constitutive models can in principle be extended, and error propagation remains partially constrained by the conservation structure of the code. For code-level surrogates, in contrast, the very objects that the current review guidelines presuppose do not exist inside the model, so the premises of the existing framework no longer hold.

This study therefore takes the code-level surrogate model, in which the regulatory gap appears in its most fundamental form, as its primary object and identifies the anticipated regulatory issues. The assessment concepts established here can be applied, with relaxation, to types with a smaller substitution scope. Model-level surrogates are beyond the scope of this paper.

2. Code-level Surrogate Models for Safety Analysis Codes

The code-level surrogate model considered in this paper is defined as follows. A code-level surrogate model is an AI model that accepts the same form of input as the reference code (e.g., MELCOR/MAAP), namely the nodalization and the initial and boundary conditions, and approximates the complete set of state-variable time histories that the code would produce by numerically integrating its governing equations, constitutive relations, and auxiliary correlations (the pressures, temperatures, and water levels of all control volumes, the flow rates of all flow paths, etc.). That is, rather than replacing any individual equation or correlation, it learns and approximates the code's input–output relationship as a whole. Figures of merit (FOMs) such as the peak cladding temperature or the vessel failure time are not predicted directly but are derived by post-processing the predicted time series.

Code-level surrogates are classified, by how physical knowledge enters the training, into data-driven and physics-informed models. Data-driven models learn the input–output relationship solely from code-generated data, whereas physics-informed models include the

residuals of the governing equations in the loss function [3]. Data-driven models are further divided, by their temporal scheme, into auto-regressive (AR) models, which sequentially predict the next state from previous states, and non-autoregressive (non-AR) models, which predict the entire time series at once.

Each type exhibits distinct characteristics from a regulatory point of view. Auto-regressive models share a structure similar to the time-marching of the reference code and are advantageous for predicting rapid transients (peaks); however, since the prediction error at each step is fed back as the input of the next step, errors can accumulate and amplify [4], making long-horizon rollout stability a concern. Conversely, non-autoregressive models are free from error accumulation but tend to smooth rapid transients and underestimate peaks. Physics-informed models have the advantage that a 'mathematical model describing the phenomena' literally exists, since the governing equations appear in the loss function, and that the equation residuals can be quantified. The residual term, however, is only a soft constraint that does not structurally guarantee the satisfaction of physical laws [5], and no criterion has been established for quantitatively evaluating its enforcement strength. Table I summarizes these characteristics.

Table I: Characteristics of code-level surrogate model types

Category	AR	Non-AR	Physics-informed
Time scheme	Sequential prediction (rollout)	Prediction of the entire time series at once	Varies with formulation (e.g., continuous space-time solution)
Strength	Superior prediction of rapid transients (peaks)	No error accumulation	Mathematical model literally present; residuals quantifiable
Weakness	Error accumulation and amplification; long-horizon rollout stability	Smoothing of rapid transients (peak underestimation)	Physics satisfaction not guaranteed (soft constraint); no criterion for enforcement strength

Meanwhile, three characteristics are shared by code-level surrogates regardless of type. First, since the training data are the products of the reference code, the model is a proxy not of the actual system behavior but of the reference code, and the reference for validation is therefore the code rather than experiments. The validation of the code against experiments is the responsibility of the existing code V&V framework, and the evaluation of a code-level surrogate concentrates on its fidelity in reproducing the reference code. Second, with physics-informed models as a partial exception, no phenomenon-wise physical model or numerical scheme exists inside the model, so the very evaluation objects presupposed by the existing review framework do not hold. Third, since the output is the time series of the same

full set of state variables as the reference code, direct comparison with the code across all time points and channels is possible. The information available for validation is, if anything, richer than for shortcut-type models that output only a single FOM.

These characteristics differ fundamentally from the evaluation objects presupposed by the existing review framework for safety analysis codes. The next section analyzes whether the domestic and foreign regulatory frameworks can accommodate them.

3. Regulatory Gap Analysis

3.1 Korean Regulatory Framework

Two regulatory layers can apply to code-level surrogates in Korea: the AI Basic Act and the nuclear-specific regulatory framework.

The AI Basic Act enumerates AI utilized for the safe management and operation of nuclear materials and nuclear facilities under the Act on Physical Protection and Radiological Emergency as one domain of high-impact AI (Article 2, Subparagraph 4 (e)) [1]. The provision is directed at the utilization of AI in the operation of a facility. A safety analysis code, in contrast, is an analysis tool used at the design and licensing stage, and whether AI element technologies embedded in such a tool constitute the management and operation of a facility is not clear from the statutory text. The determination of high-impact status therefore remains an open question (a gap in applicability). Furthermore, as described in Section 1, even if captured as high-impact AI, the obligations under Article 34 are procedural and provide no technical acceptance criteria for safety analysis (a gap in regulatory content).

In the nuclear-specific regulatory framework, the most concretely codified requirements on evaluation models are Nuclear Safety and Security Commission (NSSC) Notice No. 2017-23 [6] and, supporting it, Section 15.0.2 (Transient and Accident Analysis Methods) of the Safety Review Guidelines for Light Water Reactors [7] together with KINS Regulatory Guideline 7.1.1. Article 3 of the Notice sets the acceptance criteria, and Article 4 describes the requirements on evaluation models only functionally: 'a model whose conservatism has been demonstrated and whose use has been approved', or 'realistic simulation of the behavior, comparison with experimental data, and uncertainty evaluation'. No provision such as 10 CFR Part 50 Appendix K of the United States, which enumerates acceptable models item by item, exists in Korea, so AI models are not formally excluded. However, no detailed criteria for judging how a code-level surrogate satisfies those functional requirements have been prepared anywhere in the review guidelines either (a gap in criteria). Meanwhile, MELCOR/MAAP, the main targets of code-level surrogates, are severe accident codes handled in the accident management program evaluation framework, whose requirements on evaluation models do not reach the specificity of 15.0.2. This paper therefore takes 15.0.2, whose requirements are the most specific, as the

benchmark for examination. If the judgment is impossible even against this benchmark, the gap in the less specific frameworks can only be expected to be larger.

In summary, whether nuclear safety analysis codes fall within the scope of the AI Basic Act is unclear, and the nuclear-specific regulatory framework provides no criteria for judgment because its evaluation-model requirements remain functional descriptions. The code-level surrogate is thus judged to sit in the gap between the two frameworks.

3.2 U.S. NRC Regulatory Framework Gap Assessment

The U.S. NRC published a regulatory framework gap assessment report on the use of AI in nuclear applications in October 2024 [2]. By identifying AI use cases, screening 517 regulatory guides (RGs), and back-tracking to the 10 CFR regulations, the report identified 71 potential regulatory gaps and classified them into eight types. Of these, the ones directly relevant to code-level surrogates are Special Computations (Gap 2) and Software for Safety-Related Applications (Gap 6).

The most notable finding is the confirmation of an explicit exclusion. Because 10 CFR 50.46 and Part 50 Appendix K require physics-based models item by item for emergency core cooling system (ECCS) evaluation models, the report judged that the use of AI models within the scope of Appendix K is “practically not allowed”.

Even within the same accident analysis domain, the judgments diverge. RG 1.203 permits phenomenological models and correlations in the position of the calculational devices that constitute an evaluation model, and was therefore judged to allow limited AI applications at the separate-effect level, while stating that applications to complex processes at the integral-effect level would invite a high level of scrutiny. RG 1.157, in contrast, was judged to be a gap, having no flexibility for the use of AI. In other words, the treatment of AI is not consistent even within the U.S. framework, and only the principle that the larger the substitution scope, the higher the intensity of review, is confirmed.

Gaps in standards and tools were also pointed out. The report assessed that no AI standard mature enough for NRC endorsement exists. It found that the overall structure of IEEE Std 1012 remains valid but requires three supplements: the quality and quantity of machine learning data, AI V&V best practices, and systematic fail-safe design. It also noted that even the data quality standards (the ISO 8000 series) provide no guidance on data variety and sufficiency, and that no tools exist for judging predictive reliability other than the train-validation split.

The AI applications covered by the report, however, extend only to the level of metamodels of high-resolution physics models. The code-level surrogate, which replaces the entire input-output relationship of the reference code, was not identified as an independent type. The derivation of issues specific to code-level surrogates remains an open gap.

3.3 Comparative Observations

Table II compares the Korean and U.S. regulatory frameworks for accident analysis evaluation models. In the United States, enumerative provisions such as Appendix K create a formal conflict with AI models, which is an exclusion-type gap. In Korea, with no enumerative provision, no formal conflict arises. This, however, does not mean that the domestic framework is open to AI, but that no criteria for judging acceptability exist at all. The domestic gap is of the criteria-absence type, which translates directly into the unpredictability of review. Since neither framework is prepared to review code-level surrogates, the identification of the issues arising from their technical characteristics must precede such review.

Table II: Comparison of Korean and U.S. regulatory frameworks for accident analysis evaluation models

Category	United States	Korea
Acceptance criteria	10 CFR 50.46(b)	NSSC Notice No. 2017-23, Article 3
Prescription of evaluation methods	Part 50 Appendix K: item-by-item enumeration of acceptable models	Notice Article 4: functional description only ('a model whose conservatism has been demonstrated and whose use has been approved' / 'realistic simulation + comparison with experimental data + uncertainty evaluation')
Detailed guidance	RG 1.157, RG 1.203, SRP 15.0.2	Safety Review Guidelines for LWRs 15.0.2 (KINS/RG-N07.01, N07.02) → KINS Regulatory Guideline 7.1.1
Formal conflict with AI	Present: exclusion-type gap (violation of the App. K enumeration)	Absent: criteria-absence-type gap (no enumerative provision exists)

The gaps identified in this section are framework-level gaps and are therefore common to AI element technologies at any substitution level. The next section separates, from these common gaps, the issues that arise specifically from the characteristics of code-level surrogates.

4. Anticipated Regulatory Issues Specific to Code-level Surrogates

4.1 Issues Identified in the Existing Literature

Issues directly relevant to code-level surrogates can be inherited from the NRC gap assessment report [2] and the NIST AI Risk Management Framework it references [8]. The first is the gap in training data criteria. The report points out that supplementary guidance is needed on the quality and quantity of machine learning data, and that even the existing data quality standards (the ISO 8000

series) provide no guidance on how well the data cover the various states of a system (variety) and how many data per state are sufficient (sufficiency). For code-level surrogates trained on reference-code products, this translates directly into the problem of demonstrating coverage over the full range of accident conditions. The second is the requirement for systematic fail-safe design. On the premise that the possibility of anomalous outputs always exists in AI systems, the report emphasizes three elements: the active identification of inputs that differ significantly from the training data, the active identification of anomalous outputs, and the mitigation and correction of errors and the prevention of their propagation. The third is the gap in evaluation tools. While training a neural network to produce predictions is not difficult in itself, there is no general guidance on how to evaluate such calculations and what level of supporting documentation is required, and no tools exist

for judging predictive reliability other than the train-validation split. The fourth is the reproducibility concern arising from the opacity of deep neural networks: the basis of a prediction is difficult to understand, and reproduction across independent systems is challenging.

These four issues concern AI technologies in general and apply to model-level surrogates as well. They are inherited here as the common baseline, and the remainder of this section examines what changes when the substitution reaches the code level.

4.2 Assessment against Safety Review Guideline 15.0.2

Whereas the issues in the existing literature target AI in general, this section directly substitutes the code-level surrogate into the current domestic requirements and examines their satisfiability. Table III summarizes the results, distinguishing data-driven and physics-informed models.

Table III: Satisfiability of the evaluation-model requirements of Safety Review Guideline 15.0.2 by code-level surrogates
(Δ : partially or conditionally satisfiable; X: not satisfiable under the current provisions)

Requirement of SRG 15.0.2 (gist)	Data-driven	Physics-informed
Presentation of mathematical models describing all important phenomena and components	X No phenomenon-wise model exists inside the model; documenting output channels mapped to phenomena identification and ranking table items is possible but does not correspond to the intent of the provision	Δ The governing equations are explicitly present in the loss function
Prohibition of unphysical predictions	Δ No structural guarantee, as physical constraints are not embedded in training; practical verification through comparison of all state variables is possible	Δ Approximate satisfaction only, but residual magnitudes can be quantified
Evaluation with a frozen version and control of changes	X A version is a composite of weights, dataset, preprocessing, hyperparameters, and execution environment, and training non-determinism can yield different results even for the same version	X Same
Determination of the uncertainty of individual physical models based on separate-effect tests (SETs)	X Not decomposable into individual physical models, so item-wise uncertainties cannot be produced; if the nodalization is an input, direct computation of SET conditions may be possible	Δ Error decomposition per equation and node is possible
Evaluation over the full range of accident conditions	X No established method for demonstrating full-range evaluation in a high-dimensional input space (no criteria for judging data variety and sufficiency)	Δ The applicable range can be declared at the training design stage
Code verification (comparison with analytical solutions)	X The equations being solved and the numerical scheme presupposed by the requirement do not exist	Δ Can be addressed with the method of manufactured solutions, comparison with analytical solutions, and residual convergence
Exclusion of compensating errors: comparison with all important variables	Δ The full state-variable time series is output, so the provision can be executed literally	Δ Equation constraints partially bind the intermediate variables

What the items marked X in Table III (the frozen version, separate-effect uncertainties, full-range evaluation, and code verification) have in common is not that the performance of the surrogate is insufficient, but that the objects presupposed by the requirements, namely a decomposable physical model structure and a

numerical scheme, do not exist inside the model. This is where the code-level case departs from the model-level case. For a model-level surrogate, the surrounding code retains its phenomenon-wise model structure and numerical scheme, so the replaced model can still be reviewed within the existing framework as a correlation.

For a code-level surrogate, in contrast, most of these presuppositions no longer hold. What is needed is therefore not an improvement of the current requirements but the establishment of new evaluation items corresponding to the characteristics of code-level surrogates, and to this end the issues specific to code-level surrogates are derived below.

4.3 Newly Identified Issues

In addition to the issues inherited and derived above, the following issues, not identified in the existing literature, are derived from the technical characteristics of code-level surrogates.

The first is the absence of a metric for the fidelity of reference-code reproduction. The evaluation target of a code-level surrogate is its fidelity in reproducing the reference code, and the availability of comparison across all time points and channels is advantageous for validation. Point-wise error metrics such as the root-mean-square error (RMSE), however, are vulnerable to time delays: for a prediction in which the shape of the accident progression is identical but the major events are delayed by several minutes, an error excessive relative to the actual discrepancy is computed. No fidelity metric robust to time delays has been established. Techniques such as dynamic time warping (DTW) [9] may be candidates, but their regulatory acceptability requires discussion. This issue presupposes that the evaluation target is the full state-variable trajectory and therefore arises only at the code level. For a model-level surrogate, direct separate-effect comparisons of the replaced model remain available.

The second is the problem of tolerances for the type-specific error characteristics and of their justification basis. Since non-autoregressive models underestimate peaks while auto-regressive models accumulate errors, which are opposite error characteristics, the tolerances must be defined separately for each type. Moreover, since the reproduction error can never be zero, any tolerance demands its own justification basis. A natural candidate is to show that the error is sufficiently small relative to the validation uncertainty of the reference code, but no such comparison procedure exists in the current framework either.

The third is the absence of means to confirm the performance at the deployment stage. In the validation stage, outputs of the reference code are available and the reproduction fidelity can be measured. In actual use, the reference code is not run, so the error of an individual prediction cannot be confirmed. The model's own uncertainty estimate can be utilized as a proxy indicator, but how to validate the reliability of that estimate becomes a problem in turn, and the active identification of inputs outside the training range is a requirement directly connected to the fail-safe elements presented in the gap assessment report.

Together with the judgments of Table III, these issues are synthesized into four categories: training data (the accuracy and consistency of reference-code products, and the criteria for judging coverage and sufficiency),

model evaluation (fidelity metrics, and tolerances and their basis), prevention of anomalous outputs (blocking unphysical values, and the active identification of out-of-range inputs), and management and configuration (the definition of the frozen version, non-determinism, and the control of changes upon retraining).

5. Conclusions and Future Work

This study analyzed the gaps in the domestic and foreign regulatory frameworks and identified the anticipated regulatory issues for code-level surrogate models that replace the entire input–output relationship of nuclear safety analysis codes. The domestic framework exhibits a double gap. The coverage of the AI Basic Act for analysis-stage AI is unclear and its obligations provide no technical acceptance criteria, while the evaluation-model requirements of the nuclear-specific regulatory framework remain functional descriptions. In the United States, an exclusion-type gap arising from the enumerative requirements of Appendix K and the inconsistency of judgments across regulatory guides were confirmed. Neither framework recognizes the code-level surrogate as an independent object of review. The substitution into the requirements of Safety Review Guideline 15.0.2 confirmed that the very objects presupposed by the current requirements do not exist inside code-level surrogates, and three new issues were identified: the fidelity metric for reference-code reproduction, the justification basis of tolerances, and the confirmation of performance at the deployment stage.

The reasonable direction for developing the regulatory methodology is judged to be the extension of existing frameworks rather than the creation of a new one: grafting evaluation items such as the prior declaration of the applicability domain, rollout stability, conservation and the blocking of unphysical outputs, training data pedigree and coverage, and AI-specific configuration management onto the overall structure of the EMDAP (RG 1.203) and ASME V&V 20 [10] frameworks, together with a graded approach that differentiates the required level according to the safety significance and the use of the surrogate (screening and scoping calculations versus licensing-basis calculations).

Future work will include refining the basis for each identified issue, expanding the issue set through the further investigation of domestic and foreign regulatory documents, developing an evaluation-item matrix that maps the issues onto review criteria, and assessing candidate techniques for the identified issues, such as delay-robust fidelity metrics and uncertainty quantification methods for each model type.

ACKNOWLEDGEMENT

This work was supported by the Nuclear Safety Research Program through the Korea Foundation of Nuclear Safety (KoFONS) using the financial resource granted by the Nuclear Safety and Security Commission (NSSC) of the Republic of Korea (No. RS-2026-25538145) and the National Research Foundation of Korea (NRF) grant funded by the Korea government

(MSIT) (RS-2026-25475350).

REFERENCES

- [1] Framework Act on the Development of Artificial Intelligence and the Creation of a Foundation for Trust, Act No. 20676, Republic of Korea, 2025.
- [2] O. Pensado et al., Regulatory Framework Gap Assessment for the Use of Artificial Intelligence in Nuclear Applications, Southwest Research Institute, U.S. NRC ADAMS ML24290A059, Oct. 2024.
- [3] M. Raissi, P. Perdikaris, and G. E. Karniadakis, Physics-Informed Neural Networks: A Deep Learning Framework for Solving Forward and Inverse Problems Involving Nonlinear Partial Differential Equations, *Journal of Computational Physics*, Vol. 378, pp. 686-707, 2019.
- [4] A. Sanchez-Gonzalez et al., Learning to Simulate Complex Physics with Graph Networks, *Proceedings of the 37th International Conference on Machine Learning (ICML)*, 2020.
- [5] A. S. Krishnapriyan et al., Characterizing Possible Failure Modes in Physics-Informed Neural Networks, *Advances in Neural Information Processing Systems*, Vol. 34, 2021.
- [6] Nuclear Safety and Security Commission, NSSC Notice No. 2017-23, Acceptance Criteria for Emergency Core Cooling System Performance.
- [7] Korea Institute of Nuclear Safety, Safety Review Guidelines for Light Water Reactors, Section 15.0.2, Transient and Accident Analysis Methods.
- [8] NIST, Artificial Intelligence Risk Management Framework (AI RMF 1.0), NIST AI 100-1, 2023.
- [9] H. Sakoe and S. Chiba, Dynamic Programming Algorithm Optimization for Spoken Word Recognition, *IEEE Transactions on Acoustics, Speech, and Signal Processing*, Vol. 26, No. 1, pp. 43-49, 1978.
- [10] ASME, Standard for Verification and Validation in Computational Fluid Dynamics and Heat Transfer, ASME V&V 20-2009, 2009.