Causal Machine Learning for Risk Informed Decision Making in Probabilistic Safety Assessment

Karen D'Souza[a], and Wen-Chi Cheng[b]
[a] Idaho National Laboratory, Idaho Falls, USA, karen.dsouza@inl.gov
[b] Idaho National Laboratory, Idaho Falls, USA, wenchi.cheng @inl.gov

Abstract: Probabilistic Risk Assessment (PRA) provides a well-established framework for evaluating risk in safety-critical systems by modeling accident scenarios, estimating their likelihood, and assessing potential consequences. Central to PRA is the representation of causal relationships among system components, human actions, and external factors, typically captured through structured logical models and expert judgment. Traditional PRA approaches are largely static, while Dynamic PRA extends risk assessment capabilities by incorporating time dependent system behavior. However, neither approach explicitly focuses on causal discovery, intervention analysis, or counterfactual reasoning. Causal Machine Learning (ML) offers a complementary paradigm by enabling formal causal inference and supporting the estimation of causal effects from operational and simulated data. This paper explores the potential integration of causal ML within the PRA framework. Opportunities include improved modeling of complex dependencies, enhanced capability to evaluate the impact of operational or design changes, and the incorporation of data driven insights into risk assessments. A conceptual framework is proposed in which causal ML augments traditional PRA through causal discovery, causal inference, and expert validation while maintaining the central role of domain expertise. An illustrative use case is presented to demonstrate how causal intervention analysis can distinguish observational associations from causal effects when estimating dependencies used in PRA models. Significant challenges remain, including data limitations, causal assumptions, verification and validation requirements, interpretability, and alignment with established PRA practices. The paper concludes by outlining research directions for integrating causal methods into safety assessment workflows and emphasizes the importance of combining data driven approaches with expert judgment to ensure reliability and trust in high-risk applications.

1. INTRODUCTION

Probabilistic Risk Assessment (PRA) is the cornerstone of risk informed decision making in critical industries such as nuclear energy, aerospace, healthcare, and transportation [1]. By systematically identifying accident scenarios, quantifying their likelihood, and evaluating the potential consequences, PRA provides a structured framework for understanding and managing risk. PRA supports regulatory decision making using logical models such as fault trees and event trees combined with expert judgement and operational experience [2, 3].

A fundamental strength of PRA is its ability to represent causal relationships among system components and events. These relationships are captured through structured logical models that describe combinations of failures and undesirable outcomes. However, there is a growing demand for methods that can support more adaptive and responsive approaches to risk assessment [4].

Artificial Intelligence (AI) and Machine Learning (ML) algorithms can extract information from operational data to identify patterns associated with system performance [5]. However, ML methods are primarily designed for prediction rather than causal reasoning. In safety-critical applications, understanding cause-and-effect relationships is often more important than prediction alone. These insights help decision makers to make accurate evaluations of the overall system risk.

Causal Machine Learning (Causal ML) extends machine learning by incorporating formal methods for causal inference. In Causal ML, the emphasis is on identification and reasoning of causal relationships rather than statistical associations alone [6]. Using Causal ML, the decision makers can address questions such as what would be the impact of a failure on system risk? What would be the effect of improving maintenance quality? Such questions closely align with the objectives of PRA and suggest opportunities for integrating causal methods into existing risk assessment frameworks.

This paper explores the potential role of causal ML within a PRA framework. A conceptual workflow is proposed in which operational data, simulation data, and expert knowledge are combined to identify, evaluate, and validate causal relationships prior to incorporation into PRA models. The paper also presents an illustrative example demonstrating how causal intervention analysis may complement traditional PRA quantification by distinguishing observational associations from causal effects. The remainder of the paper is organized as follows. Section 2 provides background information on PRA, predictive ML, and causal ML. Section 3 discusses the motivation for integrating causal ML with PRA. Section 4 presents a conceptual framework for integration, the role of expert judgment, potential integration pathways, and practical challenges. Finally, conclusions and future research directions are presented in Section 5.

2. BACKGROUND

Traditional PRA is often referred to as static PRA because the event trees and fault trees that are developed are generally fixed. The dependencies are predefined by experts. The component states can be represented as functioning or failed. The risk metrics such as core damage frequency are computed based on the assumed models and probabilities.

Dynamic PRA approaches have been developed to capture time dependent system behavior in accident scenarios. The approaches address the limitations of the traditional PRA models by including several features such as time dependencies, sequential operator actions, or simulation of event progressions. One major question that is addressed in dynamic PRA is how risk evolves as the system changes over time.

2.2 Machine Learning and Causal Machine Learning

Traditional ML methods are primarily designed for prediction and pattern recognition. These approaches identify statistical associations within data but generally do not determine whether a variable directly causes an outcome. In safety critical systems, however, decision making often requires understanding how operational changes, or component failures influence system behavior.

Causal ML extends conventional statistical learning by incorporating formal causal reasoning methods that support intervention analysis and counterfactual evaluation [8]. It is used in high impact, high stakes domains such as in education, healthcare, meteorology, economics, and aerospace.

Structural Causal Models (SCMs). A SCM is a mathematical framework to define the relationship between the cause and effect within a system [7]. Traditional statistical models capture correlations, however, these SCMs explicitly describe how one variable influences the other through the underlying causal mechanisms. The behavior of the system is represented by a set of structural equations. Each variable is generated as a function of its direct causes and external factors. The system behavior of the SCM is defined by the equation below.

$$X_i = f_i\,(PA_i, U_i) \qquad (1)$$

where X_i is the variable of interest

PA_i are the parent variables (or direct causes)

U_i are the exogenous or unobserved factors

f_i is the structural function describing the causal mechanism

The structural equations can be combined with the visual representation of the equations called causal directed acyclic graphs (DAGs) thus providing decision makers with tools to distinguish between causation and correlation.

SCMs provide a mechanism for incorporating causal reasoning into risk models. Since SCMs represent causal pathways leading to system outcomes, they share several similarities with PRA methods such as fault trees and event trees. They additionally support formal causal inference.

Causal Directed Acyclic Graphs (DAGs). Causal DAGs provide a graphical representation of the causal assumptions underlying a system [7]. The nodes represent variables and directed edges represent cause-and-effect relationships between those variables. The direction of the graph indicates the direction of influence. There is a level of interpretability and transparency due to the visual nature of the DAG.

Causal DAGs are conceptually similar to the logical structures in fault trees and event trees. This is because fault trees depict how combinations of basic events contribute to a top event. DAGs explicitly encode causal assumptions and can represent broader range of relationships among system components and external influences.

Interventions. Interventions are used to evaluate the effect of actively changing a system condition and observing the impact on other variables within a causal model. In SCMs, interventions are represented by the do-operator, denoted as do(X = x) [7].

An **observational or conditional probability** is represented as $P(\,Y \mid X\,)$ which describes the probability of observing outcome (Y) given that event (X) has occurred. However, this quantity may be influenced by confounding variables, which are factors that affect both X and Y.

Unlike observational probabilities, intervention probabilities estimate the outcome when a variable is externally fixed to a specific value.

An intervention probability is represented as $P(\,Y \mid do(X = x)\,)$ which describes the probability of outcome (Y) when variable (X) is deliberately set to a value (x). By intervening on (X), the effect of confounding variables can be accounted for, allowing the direct effect of (X) on (Y) to be estimated [7].

$$P\big(Y\big|do(X = x)\big) = \textstyle\sum_z (P(Y|X = x, Z = z)P(Z = z) \qquad (2)$$

where Z represents the confounding variables that influence both X on Y. Intervention analysis directly supports risk-informed decision making by estimating the effects of proposed actions before implementation.

Counterfactual Reasoning. Counterfactual reasoning is an advanced form of causal reasoning that evaluates hypothetical scenarios by examining how outcomes might have differed under alternative conditions. The difference is that while predictive analysis focuses on forecasting future events, counterfactual analysis investigates alternate outcomes for events that have already occurred. Counterfactual reasoning enables analysts to address questions such as: What would have happened if a component had remained operational? How would system risk have changed under different operating conditions?

Within a PRA context, counterfactual reasoning can support accident investigations, operating experience reviews, and the evaluation of risk-reduction strategies. By examining alternative causal pathways, analysts can gain insight into the factors that contributed to an event and identify opportunities for improving system reliability and safety.

2.3 Comparison of PRA, Predictive ML, and Causal ML

Predictive ML focuses on identifying patterns in historical data to estimate future outcomes which could be classification or regression tasks. Common methods include linear regression models, decision trees, random forests, support vector machines, and deep neural networks. These methods have demonstrated success in anomaly detection, condition monitoring, predictive maintenance, and failure prediction. However, predictive models generally identify statistical associations and do not explicitly determine whether variables cause effects in another variable.

Causal ML extends traditional ML by incorporating formal causal reasoning. Representative approaches include SCMs, causal DAGs, causal discovery algorithms, and intervention analysis techniques. Causal methods explicitly seek to estimate the effect of actions or changes in variables within a system. They also support counterfactual reasoning regarding alternative outcomes.

PRA employs structured analytical techniques such as fault trees, event trees, human reliability analysis, and common cause failure models to evaluate accident scenarios and quantify risk. These methods represent causal pathways leading to undesirable outcomes and provide a transparent framework for risk-informed decision making. However, PRA differs from machine learning approaches because the model structure is primarily derived from system specifications, engineering knowledge and expert judgment rather than learned directly from data.

A summary of the objectives, representative methods, and example outputs of predictive ML, causal ML and PRA is provided in Table 1 below.

Table 1: Objective, methods and example outputs in Predictive Machine Learning (ML), Causal ML, and Probabilistic Risk Assessments (PRA)

Category	Objective	Methods	Example Output
Predictive Machine Learning	Predict future outcomes from patterns in data	Regression, Decision Trees, Random Forests, Gradient Boosting, Support Vector Machines, Neural Networks	Classification, prediction, anomaly detection
Causal Machine Learning	Identify and quantify the causes and effects in relationships	Structural Causal Models (SCMs), Causal DAGs, Bayesian Networks, Do-Calculus	Causal effects, interventions, counterfactuals
Probabilistic Risk Assessments (PRA)	Quantify risk	Fault Trees, Event Trees, Human Reliability Analysis, Common Cause Failure Models, Dynamic PRA	Risk metrics, failure sequences

3. MOTIVATION FOR INTEGRATING PRA AND CAUSAL ML

3.1. Limitations of Traditional PRA

In PRA, the analysts often require quantities such as, $P(Y \mid X)$
where:

- X= component failure, initiating event, operator action

- Y= subsequent system failure or accident outcome.

The problem is these probabilities are not known or difficult to estimate because there are limited failures and limited drastic operational events in safety critical systems. To fill this gap, PRA analysts must collect operating experience, perform simulation studies, make assumptions, or combine multiple sources of evidence and data. The data collection process to quantify the values of X and Y can be challenging. Questions such as what data is valid to support the analysis and are the assumptions valid arise. PRA models frequently incorporate expert judgement and rely on operating experience and simulations to characterize system behavior.

3.2 Limitations of Dynamic PRA

Dynamic PRA provides an improved representation of how risk evolves over time under changing operational conditions. Although dynamic PRA captures temporal dynamics, it does not focus on causal discovery, causal inferences, intervention analysis and counterfactual reasoning. Causal ML, on the other hand, provides complementary mechanisms for estimating, and validating conditional probabilities required by the PRA models.

A comparison of selected capabilities associated with static PRA, Dynamic PRA, and causal ML is provided in Table 2. Dynamic PRA is particularly effective for modelling time-dependent behavior, whereas causal ML provides additional capabilities for intervention analysis and counterfactual reasoning. Together, these approaches may support more comprehensive risk-informed decision-making.

Table 2: Comparison of capabilities in Static PRA, Dynamic PRA and Causal ML

Characteristic	**Static PRA**	**Dynamic PRA**	**Causal ML**
Time dependent behavior	Limited	Strong	Possible
Intervention Analysis	Limited	Limited	Strong
Counterfactual Reasoning	Limited	Limited	Strong

4. CONCEPTUAL FRAMEWORK

The conceptual workflow for integrating causal machine learning (ML) into probabilistic risk assessment (PRA) is shown in Figure 1. The data sources include system operational data, simulation data, and expert knowledge that are collected. The next step is causal discovery which is needed to study causal structures within the data. The most common methods are to use SCMs and Causal DAGs. These methods identify causal relationships. Causal inference methods estimate conditional dependencies and causal effects. These quantifications are then reviewed by domain experts to confirm assumptions and validate causal relationships. The next step is incorporation into PRA models that include fault trees, event trees and risk metrics for risk assessment. The process is iterative between subject matter experts as well as allows new data and insights to continuously refine model assumptions and risk estimates.

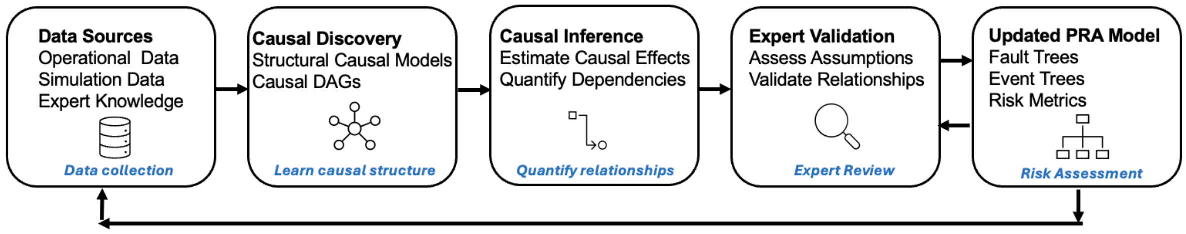

Figure 1. Conceptual workflow for integrating causal machine learning (ML) into probabilistic risk assessment (PRA)

4.1 Role of Expert Judgement

Expert validation remains a critical component of the proposed framework. Domain experts provide feedback and insights by evaluating whether the discovered relationships are physically meaningful and consistent with the system design. These experts also play an important role in assessing whether the data sources are appropriate for the system. They determine whether the extracted causal relationships should be integrated into the PRA models.

4.2 Integration Pathways of Causal ML into PRA models

Several opportunities exist for integrating causal ML into existing PRA workflows. One potential application is the addition of previously unrecognized relationships between components and operational conditions. This would support the development of an improved risk-informed model. A second opportunity involves updating conditional dependencies used within PRA models. Since direct observations of rare events are often limited, this makes it difficult to quantify conditional probabilities. Causal inference methods may improve the characterization of these dependencies and identify areas where additional data collection or simulation may be beneficial. Also, through intervention studies, by evaluating the effects of potential design modifications, analysts can investigate how specific actions may influence system risk before implementation. Such capabilities align closely with the objectives of risk-informed decision making. Finally, as additional data becomes available, causal relationships can be reassessed and validated through expert review, enabling a more adaptive approach to PRA.

4.3 Illustrative Pump Failure Use Case

A simplified example was developed using simulated system data to illustrate how causal machine learning may support the estimation of conditional dependencies used in PRA. The objective was to estimate the impact of pump failure on overall system failure and account for additional factors that influence system performance. The example system consists of five variables representing operational, maintenance, component, human, and system level factors.

Harsh operating conditions represent environmental or operational stresses that may increase the likelihood of component degradation and failure. Maintenance quality represents the effectiveness of maintenance activities and influences the reliability of system components. Pump failure represents the failure state of a critical component and serves as the primary event whose impact on system risk is being evaluated. Operator success represents the outcome of a human action that may mitigate or prevent system failure following an initiating event or equipment malfunction. Finally, system failure represents the undesirable system level outcome.

Within the causal analysis, pump failure is treated as the variable whose effect on system failure is being studied. System failure represents the outcome variable. Harsh operating conditions act as a confounding factor because they influence both pump failure and system failure. Consequently, observational data may overestimate the impact of pump failure if the effects of operating conditions are not properly accounted for. Maintenance quality and operator success provide additional information regarding system behavior and contribute to the estimation of system failure probability. Table 3 summarizes the variables used in the analysis.

Table 3: Variables used in the Pump Failure Use Case

Variable	Type	Description	Role in Analysis
Maintenance Quality	Binary	0 = Poor maintenance quality 1 = Good maintenance quality	Predictor
Operator Success	Binary	0 = Operator action unsuccessful 1 = Operator action successful	Predictor
Harsh Operating Conditions	Binary	0 = Normal operating conditions 1 = Harsh operating conditions	Confounder
Pump Failure	Binary	0 = Pump operational 1 = Pump failed	Treatment variable
System Failure	Binary	0 = System operates successfully 1 = System failure	Outcome variable

Figure 2 presents a simplified fault tree for system failure. The top event occurs if either pump failure or operator failure occurs. Pump failure is modeled as an OR combination of harsh operating conditions and poor maintenance, while operator failure is modeled as an AND combination of operator error and inadequate training. These four basic events are included in the simulated dataset so that the causal analysis is consistent with the PRA structure.

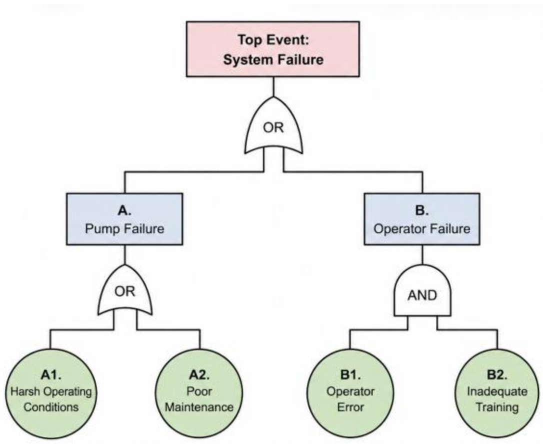

Figure 2: Fault Tree for the Pump Failure Use Case

Using Python, the system was simulated to observe the PRA and causal probabilities. The analysis begins by estimating observational conditional probabilities that a PRA analyst might calculate directly from available data. The observed conditional probability of system failure given pump failure was estimated as: $P(A) = P(System\ Failure\ |Pump\ Failure) = 0.4195$

The probability of system failure when the pump remained operational was 0.0513. These results indicate that approximately **41.95%** of observed pump failures are associated with a system failure.

However, before interpreting that this relationship is causal, it is necessary to evaluate whether this quantity $P(A)$ includes the influence of other factors called confounding variables that affect both system failure and pump failure. Table 5 presents probabilities associated with harsh operating conditions. The results indicate that harsh operating conditions increase both the probability of system failure and the probability of pump failure. Because harsh operating conditions influence both the event of interest and the outcome, they act as a confounding variable that may bias the observational estimates.

Table 4: Observational Probabilities

Quantity	Value
P(System Failure | Pump Failure)	**0.4195**
P(System Failure | No Pump Failure)	0.0513

Table 5: Confounding Analysis

Quantity	Harsh Conditions	No Harsh Conditions
P(System Failure)	0.2720	0.0188
P(Pump Failure)	0.3753	0.0271

To account for the identified confounding effects, intervention analysis was performed using the do-operator. The causal intervention probability $P(B)$ is given below.

$$P(B) = P(System\ Failure \mid do(Pump\ Failure) = 1) = 0.2123$$

It produced an estimated value of 0.2123. This quantity represents the probability of system failure if the pump were deliberately forced into a failed state while accounting for confounding variables. The probability of system failure when the pump is forced to remain operational is calculated below and represented by $P(C)$.

$$P(C) = P(System\ Failure \mid do(Pump\ Failure) = 0) = 0.0637$$

The estimated causal effect is then calculated as the difference between P(B) and P(C) which is 0.1486. This result indicates that pump failure increases system failure risk by approximately 14.9 percentage points after accounting for the effects of harsh operating conditions, maintenance quality, and operator success.

Table 6: Causal Intervention Analysis

Quantity	Value
P(System Failure | do(Pump Failure = 1)	**0.2123**
P(System Failure | do(Pump Failure = 0)	0.0637
Causal Effect = 0.2123 – 0.0637	**0.1486**

Comparing Tables 4 through 6 demonstrates the distinction between observational associations and causal effects. While the observational estimate suggests that system failure occurs in approximately 42% of observations involving pump failure, intervention analysis reduces this estimate to approximately 21% after accounting for confounding factors.

The objective of the illustrative example is not to replace or reconstruct a PRA model, but to illustrate how causal ML can provide information that supports PRA quantification. Specifically, the example demonstrates how causal inference can estimate and validate conditional dependencies that may subsequently be incorporated into fault tree, event tree, or Bayesian PRA models.

Within a PRA context, the estimated causal effect would not replace the fault tree or event tree structure. Instead, the resulting causal relationships could be used to inform the conditional probabilities and dependencies represented within the PRA model. For example, if pump failure is identified as having a significant causal influence on system failure, the estimated intervention probability may be incorporated into the quantification of fault tree dependencies or used to support the refinement of existing model assumptions. In this manner, causal machine learning serves as a complementary source of evidence that assists analysts in estimating and validating PRA parameters while preserving the underlying PRA framework.

4.4 Challenges and Practical Considerations

Several challenges must be addressed before the integration of causal ML with PRA is possible. One of the primary challenges is the availability and validity of data used to support causal analysis. Safety-critical systems are characterized by rare events, limited failure observations, and operational data that may not fully represent accident conditions. Simulation data and expert judgment often play a significant role in model development. Ensuring that these data sources are representative of the underlying system remains essential for producing reliable causal conclusions.

Causal ML does not eliminate uncertainty arising from limited or unrepresentative data. The reliability of causal conclusions remains dependent on the quality and validity of the underlying data sources. Causal models rely on assumptions regarding dependencies, and potential confounding factors that cannot always be verified directly from data. Furthermore, verification and validation procedures are necessary to ensure that learned causal relationships are physically meaningful and consistent with established engineering knowledge. Addressing these challenges will require close collaboration between causal ML researchers, PRA analysts, and domain experts to ensure transparency, interpretability, and trust in the resulting models.

5. CONCLUSION

PRA remains a foundational methodology for evaluating risk in safety-critical systems. However, the increasing availability of operational data and the growing complexity of engineered systems motivate the exploration of complementary analytical approaches. This paper examined the potential integration of causal ML within the PRA framework and discussed how causal discovery, causal inference, intervention analysis, and counterfactual reasoning may enhance risk-informed decision-making. A conceptual workflow was proposed in which operational data, simulation data, and expert knowledge are combined to identify and validate causal relationships prior to incorporation into PRA models. Several challenges in integration of causal ML to PRA models remain related to data validity, causal assumptions, verification and validation, and regulatory acceptance. Causal ML should be viewed as a complementary capability that augments rather than replaces traditional PRA. Future research directions will explore the development of practical integration strategies, validation methodologies, and case studies that demonstrate how causal methods can be reliably applied within safety assessment workflows while maintaining the central role of expert judgement.

References

[1] Kaplan, S., & Garrick, B. J. (1981). On the quantitative definition of risk. *Risk analysis*, *1*(1), 11-27.

[2] Modarres, M. (2006). *Risk analysis in engineering: techniques, tools, and trends*. CRC press.

[3] Guide, P. P. (1983). NUREG/CR-2300. *US Nuclear Regulatory Commission*, *1*, 6-43.

[4] Aldemir, T. (2013). A survey of dynamic methodologies for probabilistic safety assessment of nuclear power plants. *Annals of Nuclear Energy*, *52*, 113-124.
[5] Guo, R., Cheng, L., Li, J., Hahn, P. R., & Liu, H. (2020). A survey of learning causality with data: Problems and methods. *ACM Computing Surveys (CSUR)*, *53*(4), 1-37.
[6] Kaddour, J., Lynch, A., Liu, Q., Kusner, M. J., & Ricardo, S. (2025). Causal machine learning: A survey and open problems. *Foundations and Trends in Optimization*, *9*(1-2), 1-247.
[7] Pearl, J. (2009). *Causality*. Cambridge university press.
[8] Lamsaf, A., Carrilho, R., Neves, J. C., & Proença, H. (2025). Causality, machine learning, and feature selection: a survey. *Sensors*, *25*(8), 2373.