Reliability Engineering and System Safety 256 (2025) 110758

Contents lists available at ScienceDirect

Reliability Engineering and System Safety

journal homepage: www.elsevier.com/locate/ress

Structural causal modeling and STPA for the risk analysis of a rail system powered by H2 fuel

L. Riccardi [a,b], M. Compare [a,*], R. Mascherona [a], E. Zio [a,b,c]

[a] Aramix s.r.l., Milano, Italy
[b] Dipartimento di Energia, Politecnico di Milano, Milano, Italy
[c] Mines-Paris, PSL University, CRC, Sophia Antipolis, France

ARTICLE INFO

Keywords:
Hydrogen
STPA
Causal modeling
Counterfactuals
Confoundings

ABSTRACT

Hydrogen fuel is being considered for rail transport applications. As a new technology, it poses risks. We propose to integrate System-Theoretic Process Analysis (STPA) with Structural Causal Models (SCMs) to analyze the risks of new technology systems. The integration allows leveraging the STPA capability of identifying hazardous scenarios for a system, also due to the socio-technical environment in which the system is operated, and the SCM capability of assisting experts in the understanding of risks and in their evaluation. The integration provides a flexible framework that is here applied for the analysis of the hazards and risks emerging from the introduction of hydrogen as a fuel for the rail industry.

1. Introduction

Novel technologies are being introduced to replace traditional fuels (i.e., oil and natural gas) with more sustainable alternatives, in an effort to mitigate climate change. Hydrogen, in particular, is considered a very valuable alternative fuel for transportation (e.g., Energy Policy Act of 1992, [1]), due to the high efficiency of the fuel cells powering zero-emission vehicles (a fuel cell coupled with an electric motor is two to three times more efficient than an internal combustion engine powered by gasoline [2]), the potential for domestic production and the fast filling time. Furthermore, when used in combination with other technologies, such as renewable power and biofuels, hydrogen fuel has the potential to decarbonize some of the worst greenhouse gas emitters and, thus, it could contribute to more than 20% of annual global emissions reductions by 2050 [3]. Given this potential, investments in H2 technologies are rapidly rising: in May 2023, more than 1000 large-scale hydrogen projects have been announced globally, amounting to 320B$ in direct investments. In Europe, where 117B$ have been invested in hydrogen projects, hydrogen is expected to play a significant role in meeting decarbonization targets [3].

On the other hand, novel technologies like hydrogen-powered systems may bring new risks, due to new and unknown functional relations, failure mechanisms and hazards, etc. [4,5]. Incomplete knowledge, lack of evidence and experience result in the absence of mature guidelines and normative frameworks to guide the design and operation of systems adopting these technologies with definite assignment of liability for the specific tasks. These risks can be assessed and managed through a quantitative systematic procedure of risk analysis called Probabilistic Risk Assessment (PRA) studies, founded on well-known techniques like Hazard and Operability (HAZOP) study, Failure Modes and Effects Analysis (FMEA), Phenomena Identification and Ranking Table (PIRT), Fault Tree Analysis (FTA), Event Tree Analysis (ETA), to cite a few (e.g., see [6–8] for an introduction). PRA has proven valuable and practical for the analysis of many safety-critical industrial systems and a pillar of the Safety I paradigm [9]. Nonetheless, it suffers from some limitations ([9–12]:

(*i*) PRA stands on deep knowledge of the system functioning conditions [13]. This is questionable for new technological systems, for which unknown unsafe situations can arise. Furthermore, in the case of alternative fuels, due to the lack of experience, the risk analysis might be biased by the analysts' mindset, as their experience is based on mature practices for traditional fuels, possibly not fully applicable to identify, evaluate, prevent and mitigate the risks of the new technologies.

(*ii*) PRA assumes that accidents are caused by direct failures or causal chains of events (e.g., according to the Domino [14] or the Swiss cheese [15] models). This renders it difficult to consider important dynamic features related to context variables, control feedbacks, etc. [10].

(*iii*) The risks from new fuel technologies derive from facts and events of the entire socio-technical environment in which they are designed, manufactured, operated and maintained, with the

* Corresponding author.
E-mail address: michele.compare@aramix.ai (M. Compare).

https://doi.org/10.1016/j.ress.2024.110758
Received 8 January 2024; Received in revised form 1 November 2024; Accepted 12 December 2024
Available online 18 December 2024

involvement of multiple stakeholders, then including legislators, government agencies, industrial associations, insurance companies, companies internal organization and management, etc. Event-chain models such as those typically used in PRA are inadequate to integrate all these aspects, which can be risk sources [11].

The development of frameworks of risk analysis to give due account to the mentioned peculiar characteristics is crucial for the widespread application of new fuels. In this perspective, the Systems-Theoretic Accident Model and Processes (STAMP [11]) framework has been introduced as a complementary framework to PRA approaches for addressing limitations (*i*)–(*iii*) above [10], and the STAMP-based Systems-Theoretic Process Analysis (STPA, [16]) method has been proposed to analyze the potential causes of accidents, all the way to the development phase of a technological system, so that hazards can be identified and eliminated or controlled by design, norm or maintenance.

STAMP and STPA have been successfully applied to identify both safety and security hazards in different industrial systems (e.g., nuclear [17], maritime [18,19], aerospace [20], railroad [21], Oil&Gas [22], to cite a few). These applications highlight that STAMP offers a different approach to safety, whereby accidents are framed as the result of inadequate control or lack of enforcement by the entire socio-technical context over hazards and safety-related constraints. This is called Safety III perspective [9], where the focus is on understanding the hazard resulting from performing Unsafe Control Actions (UCAs), and is important for the risk assessment and management of technological systems [9], especially those based on new technologies.

Notice that Safety III has been introduced to extend the perspective of Safety II, which defines safety as the ability to succeed under varying conditions and, accordingly, frames safety management as the activity aimed at ensuring that 'as many things as possible go right' instead of ensuring that 'as few things as possible go wrong' [9]. The underlying idea of Safety II is that accidents are the result of surprising combinations of performance variability. The governing principle to understand accidents is resonance rather than causality, which means that the variability of two or more functions can coincide and either dampen or amplify each other to produce an outcome or output variability that is disproportionally large. [12]). As in Safety III, also Safety II rejects the Safety I perspective that models can be built to provide accurate representations of the actual system and activity, so that risk events can be controlled and, to a large extent, avoided. Nonetheless, as in Safety I, Safety III focuses on anomalous behaviors, rather than normal system working conditions. A detailed comparison of the different frameworks is proposed in [9].

On the other hand, probabilistic evaluations are not considered within the Safety III paradigm [9,10], under the claim that to prevent failures and surprises, what matters is to understand better and holistically the technological system, rather than to try to model the uncertainty in the occurrence of the causal sequence of events. In support of this claim, it has been shown that attempts to extend current PRA techniques to treat the risks emerging from software, new technologies, management, cognitively complex human control activities, etc., have been disappointing [16]. An in-depth analysis to compare STAMP/STPA and PRA is beyond the scope of the present work, and would not be an original contribution. Synthetically, Ref. [23] considers a series of works over the last decade of critical reviews of the STPA in different industrial sectors, also in comparison with PRA (e.g., [24,25]). This comparison, however, would not answer to the question of how the proposed approach can be verified to prove the benefit of its application, and justify its use in industrial practice. Nonetheless, it is evident that complementing STPA with a quantitative estimation of the hazard consequences yields insights on the most critical variables and events in the risk scenarios, which are fundamentals to reach the final aim of any risk analysis: improving the understanding of system risks for guiding further analysis and supporting decision-makers and other stakeholders in managing safety.

The main limitations of STPA reported in the comparison studies are:

- Complexity: STPA can be more complex and time-consuming to apply compared to traditional risk analysis methods, especially for large systems.
- Learning curve: It requires a different mindset and approach compared to traditional safety analysis techniques (challenging for teams used to other methods).
- Qualitative and subjective: STPA is primarily a qualitative method, which heavily depends on the analyst's subjective judgment. In particular, the lack of probabilistic models makes it challenging to prioritize the actions to mitigate and prevent risk.

To holistically identify, evaluate and manage the risks from new technologies, it seems appropriate to integrate Safety I and III techniques, with the final aim of improving the understanding of system risks, for supporting decision-makers and other stakeholders in managing safety.

In this work, we propose to leverage causal models [26] to equip STPA with the capability of managing uncertainties in safety modeling and analysis. To do this, we start from the consideration that in the STPA framework, a hazard is defined as a system state, which is mapped onto possible losses together with a particular set of worst-case environmental conditions favoring those losses [16]. This mapping of hazards onto losses is defined qualitatively only, being the STPA focus on the hazard understanding. It is precisely the definition of hazard as an event and its roughly sketched connection to the losses that make the integration of Structural Causal Models (SCMs) and STPA particularly useful: the hazard is considered as an initiating event occurring in the identified worst-case environmental conditions, which can lead to consequences of different severity levels, under different uncertain conditions. Then, SCMs are used to further investigate the hazard consequences, only roughly sketched by STPA. These investigations can guide deeper simulation and modeling, usually performed for quantitative risk analysis in PRA. Fig. 1 shows a pictorial view of the main benefits coming from the proposed combination of STPA and SCM, building on the BowTie paradigm.

Notice that the hazard definition in STPA is different from that in PRA, where an hazard is "a source with a potential to cause injury and ill health, or even the circumstances that could lead to injury and ill health" [27]. To wit, high-pressure energy is considered a hazard in PRA, which is activated, for example, by a failure event, whereas in STAMP the corresponding hazard is the well-defined event "unleash of high-pressure fluids" due to an unsafe control action (e.g., lack of control on the device integrity).

The proposed framework allows leveraging:

- the augmented capabilities of STPA over PRA techniques, such as FTA, ETA, FMECA and HAZOP, in identifying accident scenarios, as demonstrated by comparisons of PRA and STPA in which the latter found all causal scenarios found by traditional PRA analyses but also other more scenarios that the traditional PRA did not identify [16].
- The capabilities of PRA techniques of accounting for variability and causality, thus allowing us to look at safety from a quantitative risk analysis perspective [9].
- The capabilities of SCMs, which encompass Bayesian Networks (see Table 1), to offer a generalization of FTA/ETA [28,29] capable of explicitly encoding both the model uncertainty (i.e., go beyond the assumption of failures due to chains of events, item *ii*) and other non-linear dependencies.
- The capability of Modeling and Simulation (M&S) approaches to identify the relevant, possibly unknown features that affect the evolution of the hazards into losses. In this respect, SCM can be seen as a coarse M&S approach that allows exploring the potential accidents in preliminary studies, to pave the way for further investigations to be performed by more refined simulation approaches.

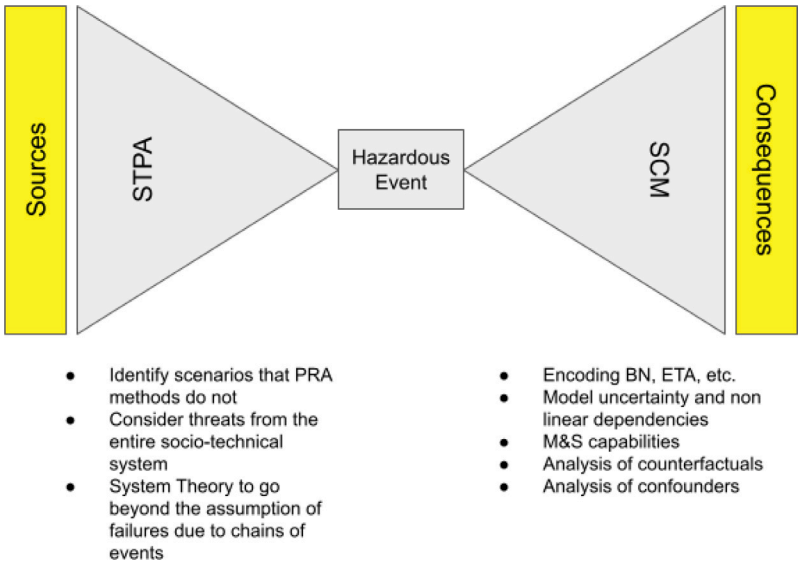


Fig. 1. Integration of STPA and SCM as two parts of a BowTie.

- the SCM capability of analyzing the confounders (i.e., variables that cause spurious associations), which helps identify further unconsidered conditions that can impact the final losses of the accident scenarios found by STPA. To do this, we rely on the method proposed by [30].
- The SCM capability of analyzing the counterfactuals (i.e., hypothetical or unobserved consequences that would have occurred under conditions different from those which actually lead to the consequence [31]) to investigate accidents and improve the understanding of the system risks.

Whilst SCMs have been recently proposed for accident investigation [32], their positioning as a tool to complement STAMP/STPA, and more generally risk analysis, is novel as well as their use for seeking unconsidered variables impacting risk assessment.

We apply the novel framework to analyze the hazards and losses emerging from the introduction of hydrogen as a fuel in the rail industry while giving due account to the complexity of the overall system, which is made up of an H2-powered train interacting with the rail infrastructure (including passenger stations, bridges, tunnels, signaling systems, etc.), maintenance depots, refueling stations, etc. The framework is developed to support the project "H2iseO Hydrogen Valley", a pioneer project for the introduction of Hydrogen-powered trains in Italy. It concerns the non-electrified Brescia-Iseo-Edolo railway line, a gateway to the Milan-Cortina 2026 Winter Olympics.

The paper is organized as follows. Section 2 briefly summarizes the main features of STAMP/STPA. Section 3 introduces the main theoretical aspects of SCM, whereas Section 4 shows its application within the STAMP/STPA framework. In Section 5, the proposed framework is applied to the H2-powered train case study. Section 6 concludes the work.

2. STAMP/STPA

STAMP/STPA is a qualitative framework for accident causality modeling, and is now being broadly applied in various industrial sectors. A relatively rich literature on STAMP/STPA theory and practice exists, and in this work we propose the integration with SCM, without going through the technical details of the STAMP/STPA methodology, which is only briefly summarized.

The STAMP approach is based on the following concepts [11,33]:

Table 1
Taxonomy of DAGs.

	Casual edges assumption	Nodes relation	Interventions & Counterfactuals	Bayesian inference
Bayesian networks	✗	Probability tables	✗	✓
Casual Bayesian networks	✓	Probability tables	✓	✓
Influence diagrams	✓	Probability tables	On leaf nodes only	✓
Structural equation models	✓	Linear	✓	✓
Structural causal models	✓	Stochastic or Functional	✓	✓

[*] All can be visually represented using Direct Acyclic Graph (DAG).

- The notion of accident goes beyond that of failure event to generalize as "*an unplanned and undesired loss event concerning something of value to stakeholders*" [16]. Accordingly, hazards are analyzed in terms of why the safety controls in place did not prevent or detect the threat and why these controls were inadequate to enforce safety constraints.
- The focus is on the system taken as a whole, rather than on its single parts taken separately. According to the system theory perspective of STAMP, some system properties emerge from the interactions of its different elements or arise from the relations established among individual components. These properties can be treated adequately only if one looks at the system as a whole [33]. If a system is hierarchically decomposed, then each level imposes constraints on the activity of the level beneath it, and two different levels communicate by means of effective Control and Feedback channels, as shown in Fig. 2. Constraints or lack of constraints at a higher level allow or control lower-level behavior. Control processes operate across multiple levels of the hierarchy to enforce safety constraints, whereby accidents occur when inadequate control is provided, violating safety constraints within the behavior of the lower-level components.

Fig. 2 gives a snapshot of the STPA method, whose main steps are [16]:

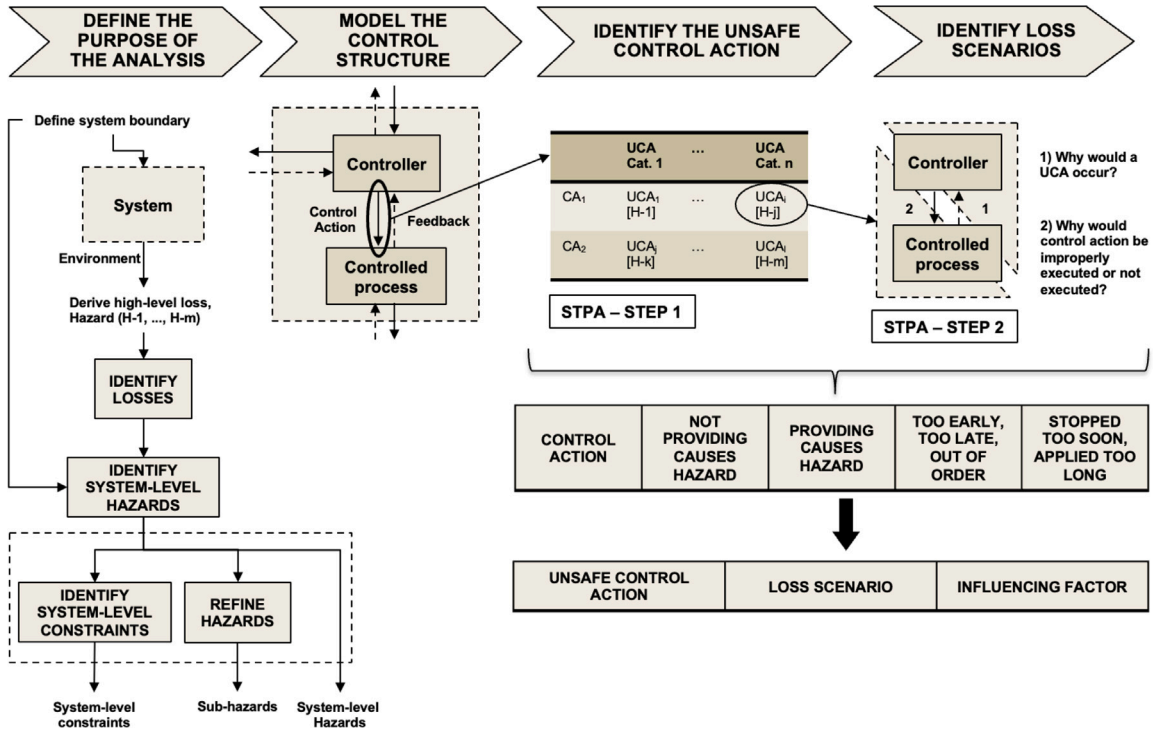


Fig. 2. STPA analysis process [34].

1. Define the purpose of the analysis, which is done through four steps:
   (a) Identify losses;
   (b) Identify system-level hazards;
   (c) Identify system-level constraints;
   (d) Refine hazards (optional);
2. Identify a hierarchical control structure, which contains at least five types of elements:
   (a) Controllers;
   (b) Control Actions;
   (c) Feedback;
   (d) Other inputs to and outputs from components;
   (e) Controlled processes;
3. Identify UCAs. There are four ways a control action can be unsafe:
   (a) Not providing the control action leads to a hazard;
   (b) Providing the control action leads to a hazard;
   (c) Providing a potentially safe control action but too early, too late, or in the wrong order;
   (d) The control action lasts too long or is stopped too soon;
4. Identify the loss scenarios by answering two questions:
   (a) Why would UCAs occur?
   (b) Why would control actions be improperly executed or not executed, leading to hazards?

3. Structural causal models

Causality focuses on the cause–effect relationship among variables or events in physical, behavioral, social and biological sciences, evaluating explanations for an observed scenario and predicting the effects of actions and policies and its development [26].

Fig. 3 shows a simple SCM, which will be used as a reference example to facilitate the exposition, whereas Table 1 shows that SCMs generalize other types of models. For a deep dive into the basics of SCMs, the reader is referred to [26,31,35].

Formally, an SCM is defined as a triplet $M = \langle U, V, \mathcal{F} \rangle$ where:

- U is the set of background variables, representing unmodeled causal influences.
- $V = \{V_1, V_2, \ldots, V_N\}$ is the set of variables of interest, influenced by $U \cup V$.
- $\mathcal{F} = \{f_1, f_2, \ldots, f_N\}$ is the set of functions such that $v_i = f_i(pa_i, u_i), i = 1, \ldots, N$, where pa_i represents the set of parent variables of V_i and u_i represents the set of background variables on V_i. The functional form of f_i can characterize any stochastic mapping.

When we define the distribution $P(u)$ over the domain of U, then $\langle M, P(u) \rangle$ is called Probabilistic Causal Model [26], whereas when the background variables are different for each node and statistically independent, then the model is referred to as a Markov model.

Any SCM can be associated with a corresponding Directed Acyclic Graph (DAG) $G(M)$, whose nodes correspond to variables U and V, whereas directed edges link U_i and pa_i to V_i, $i = 1..., N$. Every parent is a direct cause for all its children, whereby the absence of a direct link between two nodes captures our understanding that the influence on each other is mediated by other nodes (minimality assumption, [36]).

The DAG in Fig. 3 represents a simple safety system in which an emergency procedure is performed by an operator with the aim of activating a safety device. This activates two redundant systems that try to decrease the system temperature under safety threshold T_C. The model includes the following $N = 7$ variables:

1. $V_1 = X$: Emergency procedure, the completion of an emergency procedure by the operator in response to an anomalous scenario.
2. $V_2 = Y$: Safety Device Activation, indicating whether the safety device is activated upon completing the emergency procedure.

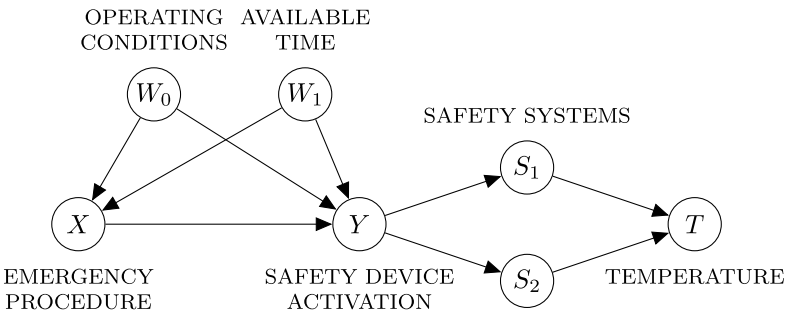


Fig. 3. Graphical causal model represented using a causal graph.

3. $V_3 = W_0$: Operating conditions, modeling the effect that poor conditions can have on both the capability of the operator to perform the procedure and the reliability of the device to switch the safety systems.
4. $V_4 = W_1$: Available time, modeling the fact that a shorter time to complete the procedure is beneficial for the activation of the safety device, which is not stressed by the anomalous conditions, whereas it is detrimental for the possibility of completing the procedure.
5. $V_5 = S_1$ and $V_6 = S_2$: Redundant safety systems activated by Y.
6. $V_7 = T$: Temperature that must not exceed a fixed threshold T_C.

As usual, for clarity the variables U, one for each node V, are not explicitly shown in Fig. 3. Notice that the direct cause is often referred to as *Treatment* (e.g., X in Fig. 3), whereas a variable that affects both treatment and outcome is called *Confounder* (e.g., W_1 in Fig. 3).

The SCMs can be used for probabilistic inference. In this case, they can generalize Bayesian Networks. For instance, with respect to the reference example, if we assume tabular functions $\mathcal{F}$, then knowing that the emergency procedure is performed correctly (X is true), one can rely on conditional probability tables to infer the probability of the safety device activation (Y is true) and, then, the probability of the activation of both redundant safety systems (S_1 and S_2 are true) and the resulting probability that temperature T does not exceed threshold T_C to guarantee safety. This model provides a powerful tool for understanding how various factors, including operator skills and actions, impact the overall safety and functionality of the system.

However, each SCM naturally defines a qualitative hierarchy of concepts, described as the "ladder of causation" [37]. The associational level described above for probabilistic inference is at the basis of the ladder, where we use statistical learning to try to infer properties of dependence among random variables from observational data.

The other two levels are interventional and counterfactual. Causal reasoning concerns these levels: it aims at drawing conclusions from the knowledge of the causal mechanisms [35]. Each level corresponds to a distinct notion within human cognition and allows formally articulating qualitatively different types of questions regarding the observed variables of the underlying system [38].

Specifically, the interventional level concerns actions taken to modify the values of some variables and are used to establish a causal relationship between the manipulated variables and the outcome of interest. This allows understanding the results of specific actions or changes.

Formally, interventions are denoted using the *do*-operator and define a submodel of SCM $M = \langle U, V, \mathcal{F}\rangle$. For a set of variables $H \in V$, we can consider a realization η and the submodel $M_\eta = \langle U, V, \mathcal{F}_\eta\rangle$, where $\mathcal{F}_\eta = \{f_i : V_i \notin H\} \cup \{H = \eta\}$.

With respect to the reference example, we might be interested in the effect of the intervention that activates S_1. This formally reads

$$\mathbb{P}(T < t|\text{do}(S_1 = ON)) \neq \mathbb{P}(T < t|S_1 = ON)$$

Practically, to build M_η, $H = S_1$ and $\eta = ON$, we delete the link $Y \longrightarrow S_1$, set S_1 to "ON" and replace equation $S_1 = f_{S_1}(y, u_{S_1})$ with $S_1 = ON$ (Fig. 4).

Notice that there is a substantial difference between associations and interventions. In the former case (e.g., Bayesian Networks), when we condition on $S_1 = ON$, we restrict our focus to the subset of the population of S_1 systems that experienced $S_1 = ON$ and use conditional probabilities to represent the strength of the relationships between variables. In contrast, an intervention corresponds to the situation in which we take the entire population of S_1 systems and assume every safety system S_1 is ON to estimate the causal effect of a variable on other variables in the system.

Counterfactual reasoning, the third level of the ladder, is related to the *Potential Outcome Framework* [36], which assumes that each variable can be given hypothetical or unobserved consequences that would have occurred under different treatment or intervention conditions. The basic idea, thus, is to ask what would have happened in a situation had certain things been different. This is like rewinding the world, changing a few crucial details and, then, predicting what happens in the fictional world. By tweaking the right variables, it is possible to separate true causation from correlation and coincidence [39].

Formally, given a Probabilistic causal Model (PCM) $\langle M, P(u)\rangle$ and some evidence $E = e$ (i.e., observable variables set to values e), the probability $P(v_{j_{v_i}}|E = e)$ of the counterfactual "given $E = e$, if it were $V_i = v_i$, then $V_j = v_j$" is computed through three steps:

1. Abduction: Bayesian update $P(u|e)$, as usual in Bayesian frameworks.
2. Action: build submodel M_{v_i}
3. Prediction: compute $P(v_{j_{v_i}}|e)$ through PCM $\langle M_{v_i}, P(u|e)\rangle$

With respect to the reference case, we can use counterfactuals to question whether the temperature might have been under the threshold had the safety system S_1 been activated, given that S_1 was not and S_2 was, and the temperature was above the critical value. This translates into the evaluation of the following counterfactual outcome probability

$$P(T_{M_{S_1=ON}}\langle t|T_{M_{S_1=OFF,S_2=ON}}\rangle t)$$

The practical difficulty in the procedure just described lies in the computation of $P(u|e)$, since conditioning on e undermines the independence between the background variables U [26]. The twin network method is usually adopted to overcome these difficulties [26,32]. This is briefly discussed in Section 4.2.

The causal effect of a binary treatment is usually estimated by the Average Treatment Effect (ATE). With reference to Fig. 4, the ATE of X on its child Y is defined as $ATE(X, Y) := \mathbb{E}[Y(1) - Y(0)]$, where $Y(1)$ is the potential outcome of the device activation one would observe if the emergency procedure were correctly performed ($X = 1$), whereas $Y(0)$ is the potential outcome of Y that one would observe if the procedure were to not properly implemented ($X = 0$). The estimation of the ATE relates to the fundamental problem of causal inference [40], which is fundamental because we cannot observe both $Y(1)$ and $Y(0)$ for the

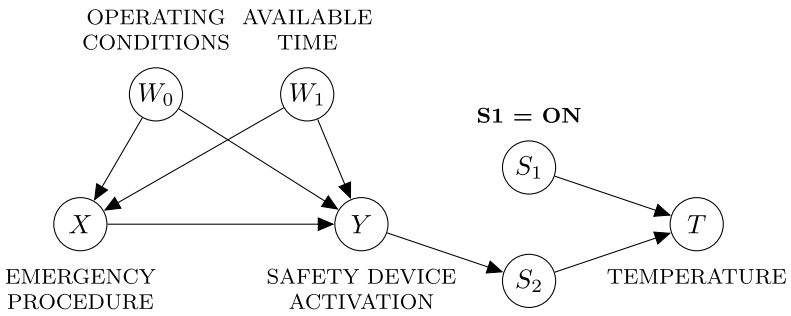


Fig. 4. Causal graph representation of the intervention "do S_1 = ON".

same unit. This problem is unique to causal inference, where we deal with causal claims defined in terms of potential outcomes [36].

When the ATE is identifiable, it can be estimated using purely observational variables. In this work, we rely on the Double Machine Learning (DML, [41]) technique to estimate the ATE, which is made up of two stages. In the first stage, we fit two models:

- model 1, to predict Y from W, obtaining the predicted $\hat{Y}$;
- model 2, to predict X from W, obtaining the predicted $\hat{X}$.

In the second stage, we "partial out" W and fit a model to predict $Y-\hat{Y}$ and $X-\hat{X}$. This gives the causal effect estimate. DML has been preferred to other methods such as Linear Regression and Causal Trees and Forests [42] because of its main characteristics: it is non-parametrically efficient (i.e., the ATE estimator has the smallest possible variance of any estimator that does not make parametric assumptions) and it is doubly robust (i.e, the estimator is consistent and the error goes to zero when at least one of the estimated model goes to zero) [36,43].

4. SCM integrated with STAMP

As mentioned in Section 2, SCMs are here introduced to complement STAMP-STPA. Specifically, through STPA we identify both the variables that represent boundaries and constraints, which reflect essential indicators for the system-level hazards and the accidental scenarios, and the design parameters or system properties and variables that influence the overall behavior, including the worst-case conditions. These results are leveraged to build SCMs that describe the causal mechanisms defining the evolution of the system and estimating the probability that hazardous events lead to the loss events.

For brevity, we focus only on two additional tasks enabled by SCM, which can enhance the comprehension of the system functioning and, thus, the robustness of risk assessment:

1. Sensitivity Analysis to Unobserved Confounding.
2. Counterfactual Analysis.

Additional benefits of the integration of STPA with SCM will be discussed in future research work.

4.1. Sensitivity analysis to unobserved confounding

In general, the simulation models developed to assess the consequences of risk scenarios (i.e., the part on the right in Fig. 1) rely on the assumption that all the relevant variables and causal mechanisms are included in the model. This assumption is fundamentally untestable, although its violation can heavily bias the risk analysis outcomes. Therefore, it is of paramount importance to question if the model actually includes all relevant factors. SCMs are particularly effective for that task, as they allow identifying missing confounders that impact most on the analysis results. This at the basis of expert brainstorming aimed at seeking for meaningful confounders that might have been excluded from the model. In turn, this is a very effective way to cope with the unknowns [9].

Different techniques are available in the literature to investigate the possible biases induced by confounders that are not actually considered in the risk model. In causal modeling, the techniques that question whether the model is missing relevant factors are referred to as sensitivity analysis techniques. In the PRA literature, however, sensitivity mainly refers to assessing how changes in input values affect the risk assessment results, as the aim is to estimate the range of possible outcomes and their associated probabilities to give a more comprehensive view of the risk landscape. Of course, these slightly different sensitivity analyses can be integrated, building on the capability of the techniques considered in this work of estimating the uncertainty in the causal effects (e.g., Table 4). This issue will be tackled in future research work. This will be done in future research work.

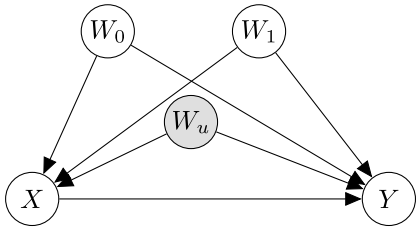
The first way we can do this is by getting an upper and lower bound on the causal effect based on credible assumptions [44]. Otherwise, we can estimate how strong the unobserved confounder's effect on the treatment and outcome need to be to make the true causal effect substantially different from our estimate. This can be done in different ways:

- Austen Plots [30], which provide a sensitivity analysis tool to aid in reasoning about potential biases induced by unobserved confounding.
- Linear partial R^2: Sensitivity Analysis for linear models using the Linear partial R^2 method [45]. A novel procedure to formally bind the strength of the confounders based on the observed covariates is introduced in [46].
- Non-Parametric partial R^2 based: Sensitivity Analysis for non-parametric models using a non-parametric partial R^2 method [47].
- Compute bounds on the Conditional and Average Treatment Effect (CATE) in the presence of unobserved confounding factors [48].
- Perform a sensitivity analysis for Unobserved Confounding in Nonexperimental research [49].

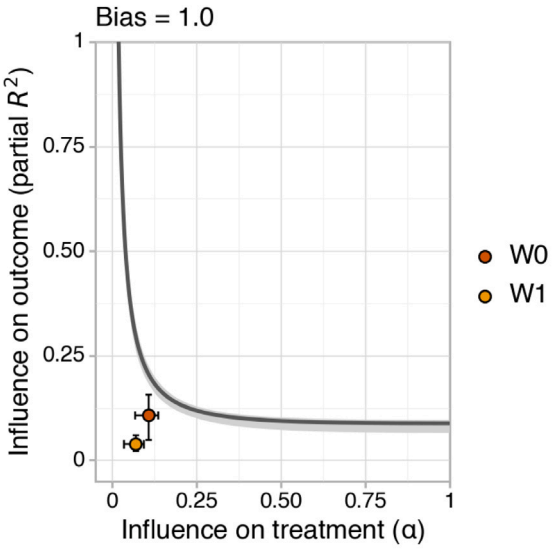
To the Authors' best knowledge, none of these techniques has ever been employed in reliability and risk engineering to discover unknowns. For brevity, we show how to proceed with the first approach, only. The main advantages of this approach are:

1. Plausible judgments are made on directly interpretable quantities, i.e., the confounding influence on treatment and effect.
2. The strength of the unobserved confounder can be directly compared with the strength of observed covariates.
3. The analyst does not have to consider any aspect of the sensitivity analysis when modeling the observed data.

We only provide an intuitive explanation of the plots considering the reference example, the theoretical details being available in [43]. In abscissa of Fig. 13, we find parameter α representing the strength of the influence of the unknown confounder on treatment, whereas in ordinate we find parameter R^2_Y, representing the outcome-confounder



(a) Model with an unobserved confounder



(b) Model with an unobserved confounder

Fig. 5. Austen plots.

strength. The observed variables are X, Y, W_0, W_1, whereas W_u represents a possible unobserved confounder. To question if any other unknown confounder is missing, we use the following general procedure, whereby the domain experts are asked to translate judgments about the strength of the unobserved confounding into judgments about the bias induced in the estimate of the impact [43]:

1. Produce an estimate of the ATE by using any modeling tools (DML in our case).
2. Pick a level of bias that would suffice to change the qualitative interpretation of the estimate, such as the lower bound of the 90% confidence interval of the ATE.
3. Plot the values of α and R^2_Y that would suffice to induce that much bias. This is the black curve in Fig. 5(b).
4. Compute the influence level for the observed covariates. These are the circles in Fig. 5(b), with relevant 90% confidence bars.

Examining the plot, we can see that an unobserved confounder as strong as W_0 could induce an amount of confounding not sufficient to invalidate the outcome of the safety device. Then, if the experts believe that such a confounder as strong as W_0 is unlikely, it may be concluded that the model is effective and conclusions for risk analysis are robust to missing confounding. Otherwise, brainstorming is start to identify possible confounders for which that bias is plausible.

4.2. Counterfactual inference

Counterfactual reasoning might be very useful to complement STAMP-STPA for improving the "Learning from Events" task. In this respect, Causal Analysis based on STAMP (CAST) has been introduced to identify the questions that need to be answered to fully understand why an accident has occurred, to maximize the learning from events. To do this, CAST analyzes the highest levels of the safety control structure derived from STPA to understand how and why each successively higher level has allowed or contributed to an inadequate control at the level under analysis, the failures in the process models of those who made the decisions and why those failures existed, etc. To do this, graphical representations that illustrate interactions and causality are advocated as useful tools to assist in the analysis.

SCMs answer this request, as they allow for the integration of an analyst's causal knowledge of a system with the evidence on an event of interest. As a result, counterfactual hypotheses, which are of common

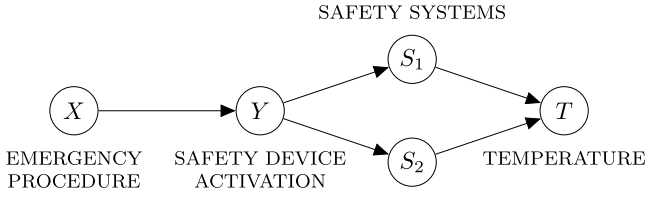


Fig. 6. Graphical causal model.

use in the practice of system safety, can be rigorously assessed through causally-sound probabilistic methods. Moreover, SCMs can leverage quantitative data from multiple sources to provide qualitative insights into human performance for building a predictive HRA model [50]. This is of particular interest to CAST.

We consider the reference example of the redundant safety system (Fig. 6) to present counterfactual reasoning in support of STPA-driven risk analysis. We rely on the twin network approach [26], which operationalizes the three steps presented in Section 3. For simplicity, we consider Conditional Probability Tables (CPTs, Table 5) to represent the causal mechanisms.

A twin network consists of two interlinked networks, one representing the real world and the other the counterfactual world of interest. Specifically, to construct a twin network given a SCM and use it to compute a counterfactual query is as follows. First, we duplicate the given SCM, denoting the nodes in the duplicated model by superscript $*$. If $V = \{V_1, \ldots, V_n\}$ is the set of observable nodes, then $V^* = \{V_1^*, \ldots, V_n^*\}$ is the duplicated set. However, for every node V_i^* in the counterfactual world, the background parent U_i^* is replaced with the background parent U_i of the original, "factual" model, such that each original background variable U_i is now a parent of two nodes, V_i and V_i^*. The two graphs are linked only by the background variables, while sharing the same node structure and generating mechanisms. To compute a general counterfactual query $P(T|E = e, \text{do}(V_i = v_i))$, we modify the structure of the counterfactual network by dropping arrows from parents of V_i^* and set $V_i^* = v_i$. Then, in the twin network with this modified structure, we compute $P(T^*|E = e, V_i^* = v_i)$ via standard inference techniques, where $E = e$ are evidences in factual nodes [39].

With reference to the SCM in Fig. 6, for example, we query the counterfactual: "Given that temperature T exceeded the critical threshold T_C

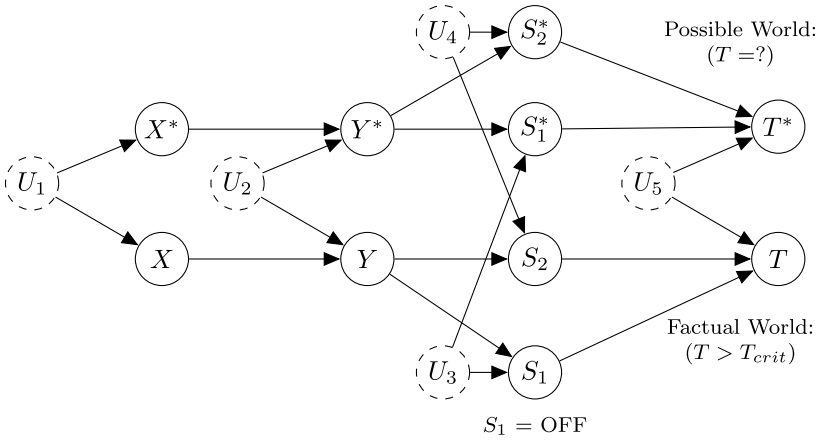


Fig. 7. Twin network.

in the scenario in which S_1 was failed, what is the probability that the temperature would have not exceeded T_C even if S_1 had worked?"

The twin network is shown in Fig. 7. We set the evidence $S_1 = OFF$ and $T > T_C$ and update the CPTs (Table 6). Then, we set the intervention "S_1* = ON" to estimate the counterfactual $P(T^*_{S_1^*=ON} < T_C | S_1 = OFF, T > T_C)$ Fig. 8. The computation results in the following estimation:

$$P(T^*_{S_1^*=W} < T_C | S_1 = F, T > T_C) = 0.7345$$

This result shows that in the conditions in which the temperature reached the critical value T_C, guaranteeing that the safety system S_1 works leads to a probability of successfully reducing the temperature smaller than that corresponding to the scenario in which both S_1 and S_2 certainly work (i.e., 0.9, Table 5), but larger than that of the scenario in which S_1 works and S_2 is failed. This is due to the fact that the evidence in the factual world (i.e., S_2 was not effective in avoiding the increase in the temperature), undermines the trust in this device also in the counterfactual world (through the background variables).

Notice that these results are different from those one would get when applying standard PRA techniques like FTA: in that case, the analysis would have selected one out of the four combinations of S_1 and S_2 working or failed, and mapped that combination in the $T > T_C$ event probability (i.e., the probability values in Table 5). The difference lies in that we are no longer referring to the population of S_2 systems, but to that specific system which did not work in the specific situation.

Notice also that the counterfactual value highlights the relevance of the specific safety barrier S_1 in avoiding the accident. Then, it can be intuitively interpreted as an importance measure [51] for that specific safety barrier in the considered system. The relationship between counterfactuals and importance measures will be investigated in future research works.

5. Case study

In this Section, some details are provided on the application of the proposed framework to the risk analysis for the introduction of Hydrogen as a fuel for the trains on the Brescia–Iseo–Edolo railway line. The entire study is quite copious, complex and cannot be completely disclosed, whereby it is not possible to report the complete process undertaken and show all the results. Then, the aim of this Section is to give insights into how the proposed framework can help the risk analysis of a system equipped with new technology. To do this, we extract the main features of the STAMP application and show one application of the confounder analysis and one of the counterfactual reasoning to illustrate how they can be useful for the investigation of the accidental scenarios emerging from complex aspects associated with the integration between new fuel and well-known background systems.

5.1. STAMP/STPA

As mentioned before, reporting the entire STPA analysis is out of the scope of the present work, also considering that the application of STAMP per se is not a novel contribution. Then, the focus of the Section is on the information that STPA can provide for building SCMs to investigate the consequences of the hazards.

Notice that this analysis is carried out for a new technology in support to the design activities for its adoption, i.e., at a project stage when the system is "on paper" and neither field data nor a digital model are available. Once they will, different approaches will be investigated for the integration of SCM and STAMP/STPA. The main research paths will focus on the development of algorithms for causal discovery, which try to derive the SCM directly from the data [52,53], and frameworks to validate the safety barriers effectiveness (e.g., [54]).

5.1.1. Define the purpose of the analysis

The experts highlighted various areas of concern toward the overarching goal of guaranteeing the train service, including passenger safety, in-time transportation and environmental protection. Accordingly, different losses have been identified such as loss of human lives, loss of duties and customer satisfaction, economic loss, vehicle or environmental damages, etc.

We focus on the losses related to the introduction of the emerging fuel, only. Specifically, hydrogen is a highly combustible gas and can be easily ignited by small sources such as sparks, electrical discharges, etc., within a very wide flammability range: at atmospheric pressure, H_2 is combustible when its concentration lies within the range $[4 - 74, 2]\%$ (the results of this study concerns mainly this threat), known as Lower and Upper Flammable Limits. Therefore, one of the most relevant losses is "*Fire/Explosion*". Here we report some findings of STPA concerning this loss.

According to STPA, we link this loss to a set of relevant hazardous events undergoing further investigation. According to the STPA procedure, System Level Hazards, System Level Constrains and Responsibilities have been listed in this phase, although they are not reported. To wit, from STAMP it emerges that Fire/Explosion requires the presence of ignition sources and a H_2 concentration inside the Flammable Limits, which are two relevant hazardous events. For simplicity, this is summarized in Fig. 9(b), through the FTA perspective, and ().

5.1.2. Hierarchical control structure

Fig. 9(a) shows a simplified control structure for the Compressed Hydrogen Storage System (CHSS), which has been extracted from the complex one used for the entire study. The influencing variables for the H_2 atmospheric concentration have been identified and will be investigated in the SCM.

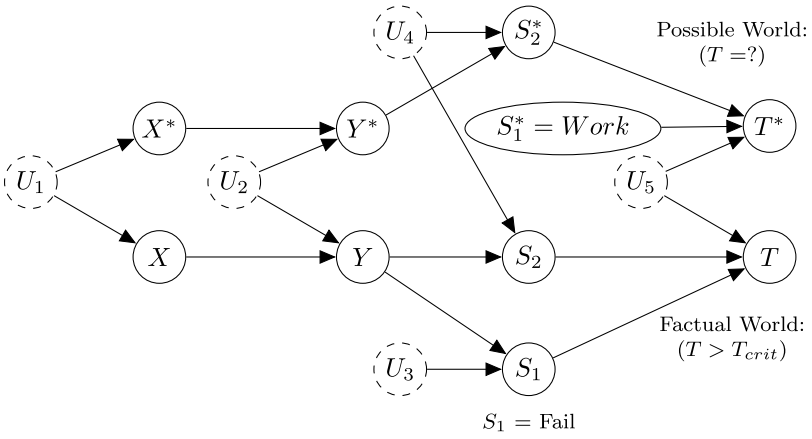


Fig. 8. Action step: Do($S_1^* = Work$).

A list of Control Actions has been derived, which include the reference to Source, System, Action, Context, Process model, Interface models, Actuators and Feedback. This list is here omitted.

5.1.3. UCA analysis

To identify UCAs, it is necessary to specify the context in which the control action might lead to a hazard. Table 3 gives a short list of UCAs for the CHSS, extracted from the complete list. Specifically, only H_2 *Release/Leakage* events are considered in this work, which represent the major source of concerns for the train manufacturer in the management of the Fire and Explosion Loss. The UCAs in Table 3 highlight the fundamental role of the context: the same Control Action becomes unsafe for different causes and leads to different hazardous consequences.

5.1.4. Loss scenarios

The link between UCAs and potential accidents results in the identification of the loss scenarios. This last step of STAMP/STPA is at the basis of the causal modeling task, which complements the analysis with quantitative estimations and the capability of providing a better understanding of the risk scenarios. For example, **UCA-4** entails damages to the fuel cells, whereas **UCA-3** and **UCA-5** result in an explosive or flammable gas atmosphere outside the train. In the same perspective, **UCA-6** leads to a tank burst.

Hydrogen release from the Gas Treatment Unit (GTU) or the Thermal Pressure Relief Device (TPRD) represents safety barriers to avoid damage to CHSS components (**UCAs 4 and 7**), provided they are correctly managed. Otherwise, they yield to explosive or flammable gas atmospheres (e.g., in case of limited available dilution volume). This risk is considered by **UCAs 3 and 6**. It is mandatory to ensure that whenever these two releases occur, the flammable gas has sufficient space to dilute in the atmosphere and reach safe concentration levels. For this, as emerged from STPA, a fundamental variable to consider to quantitatively assess the risk is the safety distance, which is the distance in any direction from the release point up to where the gas/air mixture dilutes to a concentration below the Lower Flammable Limit. In [55], this is referred to as the extent of a zone.

However, to depict the worst scenario according to STPA, we need also to consider the variables that can influence the hazard consequences. For example, another fundamental variable to consider is the "Available Space", which represents the distance along the release direction up to which the gas flow does not encounter any obstacle that might slow down the dilution process. The other variables identified by STPA as contributing to determining the worst scenario are reported in the SCM in Fig. 10.

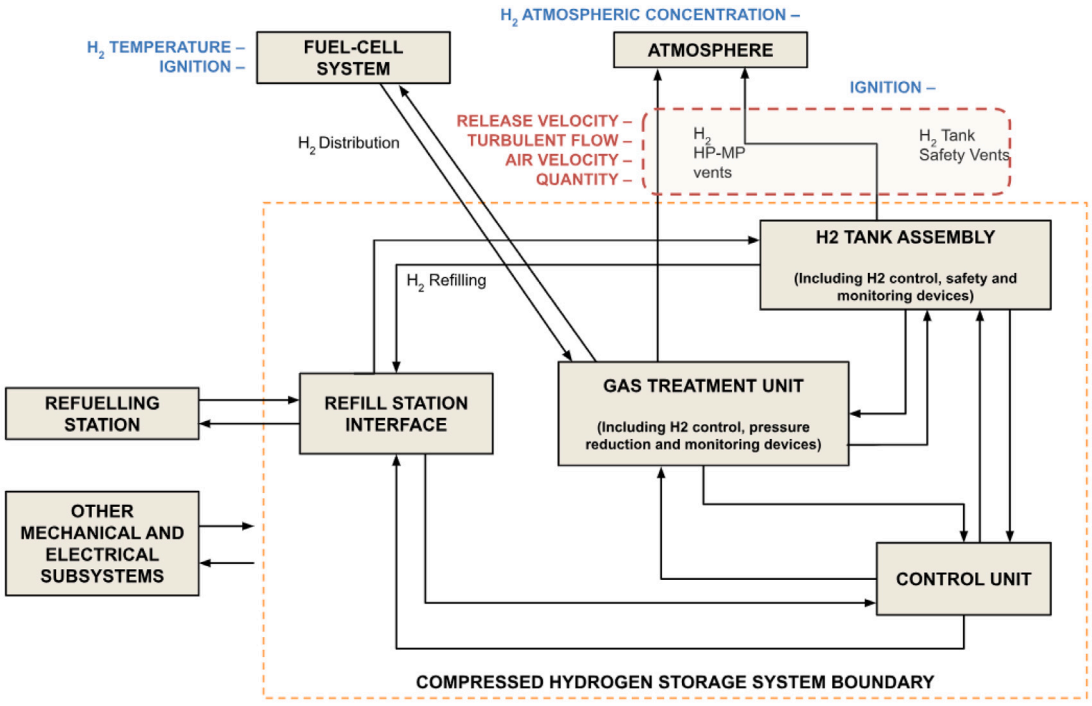
Table 2
H_2 Release/Leakage Hazards, IFs and variables.

Hazard ID	Description
H-1	Unintentional release of hydrogen in a confined space (inside CHSS compartments, during refueling operations, etc.)
H-2.1	Hydrogen venting in well-ventilated space as part of a safety protocol results in unexpected accumulation
H-2.2	Hydrogen venting in a confined space as part of a safety protocol results in unexpected accumulation
H-3	Presence of heat sources from overheated equipment or components
H-4	The system (train/railway infrastructure) generates uncontrolled ignition sources (sparks from electrical components or mechanical operations, open flames, etc.)
Influencing factors	
Atmospheric conditions Surrounding environment System design (e.g. ventilation) Operational practices and Human behavior Dilution volume	
Variables	
H_2 Atmospheric Concentration (hazardous if lies within flammability range)	
H_2 Temperature (hazardous when its control is not maintained)	
Distance from flammable substances and other ignition sources (hazardous if below specific safety thresholds)	

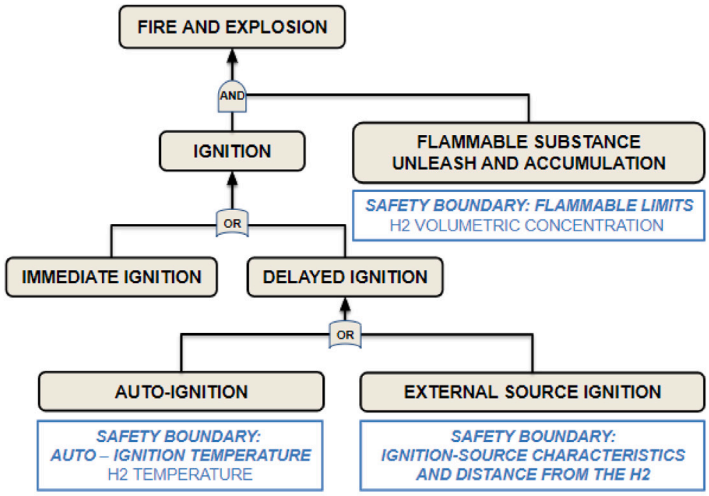
5.2. Causal modeling

The SCM model builds on the International Standard for Explosive Atmospheres [55], considering the variables identified through STPA. Given the broad scope of the model, the hazardous event investigated is the general H_2 *release/leakage"*, which may be the result of safety measures or unintentional leakage due to malfunction.

Notice that Y (the outcome of interest) refers to a distance instead of an atmospheric concentration. This is because a single-dimension space-related variable simplifies the identification of the areas where the mixture of air with flammable gases or vapors can ignite and permit self-sustaining flame propagation [55]. Extending the analysis to the three-dimensional volume surrounding the leakage is the next step of the study, according to the view that SCM is a rough M&S approach that allows exploring the potential accidents in preliminary studies, to pave the way for further investigations to be performed by more refined simulation approaches. Given the complexity of the H2 dispersion



(a) Simplified control structure for the CHSS. Due to the complexity of railway infrastructure systems, a comprehensive description of the hierarchical control structure is not reported this work.



(b) Fire and Explosion Prevention: synoptic view.

Fig. 9. STPA outcomes.

behavior near leaks, this more refined simulation model will rely on computationally intensive CFD analysis. Notice that this double-step M&S approach is fully compliant with the IEC International Standard, which proposes a rough model based on critical distances (rather than volumes) to provide preliminary estimations on gas dilution behavior.

The cause–effect relations between the variable in the DAG of Fig. 10 grounds on both physical and expert-driven functions. For example, the release speed X_1 is estimated by

$$X_1 = \frac{W_g}{\rho S} + U_1 \tag{1}$$

where $U_1 \sim N(0,1)$ and

$$W_g = C_d S p \sqrt{\gamma P_0 \left(\frac{2}{\gamma+1}\right)^{(\gamma+1)/(\gamma-1)}} \tag{2}$$

is the release rate with choked gas velocity (sonic release).

The airspeed results from the sum of three independent contributions: the wind effect, the train moving and a ventilation system. These have been combined through CPTs based on expert qualitative statements.

Table 3
UCA analysis for H_2 Release/Leakage events.

Control action	Source/Target	UCA types		
		(a) Provided but not needed and unsafe	(b) Provided, but with the incorrect intensity & timing	(c) Not provided, when needed to maintain safety
CHSS Isolation Valve Closing	CHSS Leak Detection System/CHSS compartment	N/A	**UCA-1.1**: CHSS Leak Detection System provide partial Isolation Valve Closure during H_2 leakage from the Pipeline. [H-1]	**UCA-1.2**: CHSS Leak Detection System does not provide Isolation Valve Closure during H_2 leakage from Pipeline. [H-1]
GTU Venting Valve Opening	GTU relief valve vent/Atmosphere & Fuel-cell system	**UCA-2**: GTU valve provide H_2 venting into the atmosphere during normal operating conditions. [H-2.1]	**UCA-3**: GTU valve provide H_2 venting with improper environmental conditions (e.g. inside a tunnel) and/or insufficient flow. [H-2.2]	**UCA-4**: GTU valve does not provide H_2 venting during H_2 tank overpressure event. [H-2.1, H-2.2, H-3, H-4]
TPRD open vent	TPRD & H_2 tank/Atmosphere	**UCA-5**: TPRD valve provide H_2 venting into the atmosphere during normal operating conditions. [H-2.1]	**UCA-6**: TPRD provide H_2 venting with improper environmental conditions (e.g. inside a tunnel) and/or insufficient flow. [H-2.2]	**UCA-7**: TPRD does not provide H_2 venting opening during H_2 Tank overpressure and overheating event. [H-2.1, H-2.2, H-3, H-4]

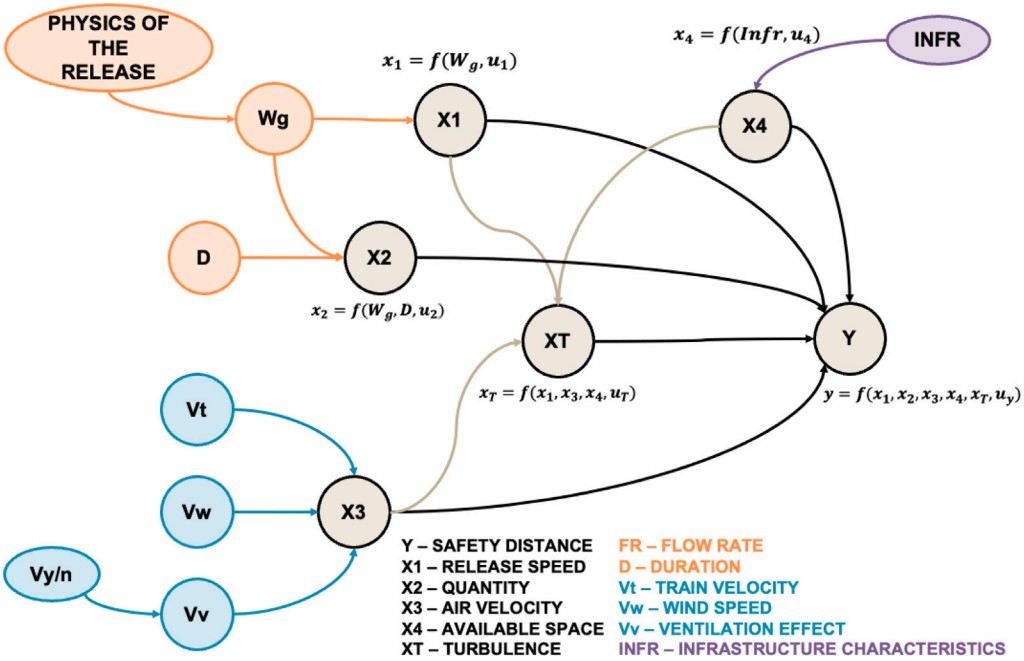


Fig. 10. Structural causal model.

The model also accounts for the turbulent flow onset's degree by means of expert-based functions due to the phenomenon's complexity. For the distance, we consider a Gaussian linear additive model:

$$Y = \alpha X_1 + \beta X_4 + \gamma X_3 + \delta X_T + U_Y$$

where $U_Y \sim N(0, \sigma^2)$ and the model coefficients have been estimated through a linear regression on Computational Fluid Dynamics (CFD) analysis outcomes. The larger the value of Y, the more risky the scenario.

5.2.1. Confounder analysis

In this Section, we give an example of how the confounder analysis illustrated in Section 4.1 can help improve the risk model.

Specifically, to build the Austen plot, we consider a reference level of bias (e.g., 1.0), close to the ATE estimated through DML for the model with no unobserved confounders. As mentioned before, the colored dots in Fig. 11 show the influence strength of the confounders (Table 4). The domain experts' analysis of Fig. 11 highlights that the model is not very robust to unobserved confounders with an impact similar to that of the release speed and the available space. This information is very useful in the first phases of design, as it enables brainstorming sessions with domain experts to discuss the plausibility of the existence of any additional influencing factor, not emerged in STPA, with equal or larger influence with respect to the other confounders. In the considered scenario, these sessions resulted in the identification of one possible additional confounder that although not emerged in STPA cannot be preliminary excluded: Atmospheric Humidity. This variable has been considered because experts believed that its influence might be not much smaller than that of average space.

Notice that as in any setting in which interviews to experts are performed to retrieve information useful for modeling, particular care should be paid in preparing the questions, to avoid possible biases. However, to the Authors best knowledge, this issue has not been addressed for the specific case of questions based on Austen plots. Then, it is an important open issue for further research work.

Humidity is expected to affect the H_2 atmospheric concentration in two different ways:

- Given that the dilution process becomes slower in a thinner atmosphere, due to the enhanced effect of the recirculation processes and vortices, humidity can decrease the release speed and increase the turbulence impact. Nevertheless, it seemed unrealistic to experts to imagine a cause–effect relationship between atmospheric humidity and the gas pressure P_0 at the release point, considering that this parameter represents the sole free variable in the physical Eq. (2) governing the release speed X_1 (Fig. 12).

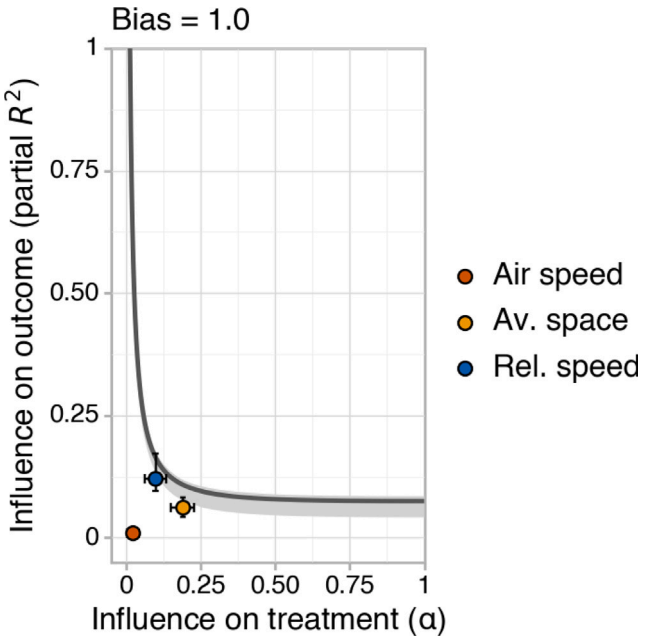


Fig. 11. Austen plot.

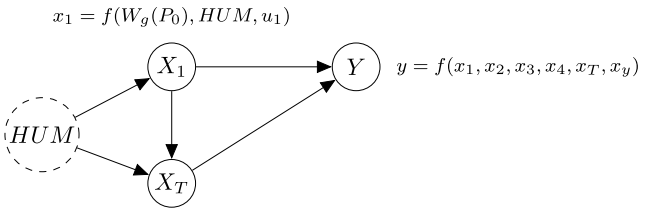


Fig. 12. Humidity as a confounder between turbulence and release speed.

- Humidity increases the H_2 Lower Flammable Level (Fig. 13(b)). Therefore, it can reduce the distance at which a flammable gas atmosphere may arise. At the same time, it can affect the turbulent flow onset's degree (Fig. 13(a)). This confounder is deemed to be potentially relevant.

To investigate the actual effect of humidity in Fig. 13(b), Table 4 compares the ATE with bootstrapped 90% confidence intervals for the causal link between X_T and outcome Y before and after including humidity into the causal model. The causal links encode the Shapiro diagram relationships (Fig. 13(b)). It can be seen that once explicitly modeled in the SCM, humidity actually increases the causal effect of Turbulence on the outcome. Then, even if the statistical evidence is not very strong (the confidence intervals for the two ATE estimations overlap with each other), experts agreed that the variable should be taken into account in more in-depth CFD simulations.

The procedure was iterated to check the opportunity to include other confounder variables to build a more robust causal model. However, no additional variable was added.

This example shows that the approach is very useful to get preliminary feedback about the estimation of the risks, especially in terms of sensitivity of the estimations to parameters. This yields a better understanding of risk, based on quantitative analysis.

Finally, notice that there exists a trade-off between complexity and completeness deriving from the necessity of including all the possible confounders that might affect the cause–effect relationships and the ability to conceive a causal model that remains feasible to interpret and treat.

Table 4
Causal effect estimates computed in the original model and after the integration of the Humidity.

	ATE	Confidence interval
Turbulence (Original model)	1.133	(1.058, 1.205)
Turbulence (Model with humidity)	1.205	(1.108, 1.361)

5.2.2. Counterfactual outcomes

This Section shows a practical example of how counterfactual reasoning can help understand risk. Specifically, we use the twin network approach discussed in Section 4.2 to statistically assess the role of turbulent flow onset's degree with respect to the considered loss event Fire & Explosion (Fig. 9(b)). A simplified version of the SCM of Fig. 10 is considered in Fig. 14, where I is a Bernoulli random variable for the presence of ignition sources at a given distance $D = 0.5$ m, whereas the occurrence of the Fire & Explosion event is modeled as:

$$F\&E = \begin{cases} 1 & \text{if } I = 1 \ \& \ Y_{obs} > D \quad \text{Fire/Explosion occurs} \\ 0 & \text{Otherwise.} \quad \text{Fire/Explosion does not occur} \end{cases}$$

This models the fact that if the estimated safety distance is larger than D, then at D we have a concentration larger than 4%.

The following analysis answers the following counterfactual query, emerged after the CFD simulations considering open-air conditions: "*Given that "Fire & Explosion" has not occurred during a release with train staying in the open air, what is the likelihood that the same event would have occurred had the same release occurred inside a tunnel with ceiling wall at a distance no longer than* 0.5 *m from the leakage source point?*". The initial probability estimate is:

$$P(F\&E = 1 | I = 1) = P(Y_{obs} > D) = 0.05$$

Formally, we are interested in estimating:

$$P(F\&E^*_{X_4^*=0.5, I=True} = 1 | F\&E_{X_4=\infty, I=True} = 0)$$

Following the three-step approach (i.e., Abduction, Action & Prediction) introduced in Section 3 and detailed in Section 4.2, we first estimate $P(u|F\&E_{X_4=\infty} = 0)$ (i.e., Abduction). This results in the update of the mean for all noise variables (i.e., $U_i \sim N(0, \sigma_1^2) \longrightarrow U_i|e \sim (\mu_i^*, \sigma_i^2)$, etc.).

In the second step, Action, we set $Do(X_4^* = 0.5)$, which directly increases the Safety Distance mean by $0.12m$ (i.e., almost 25%). We finally estimate the probability of interest:

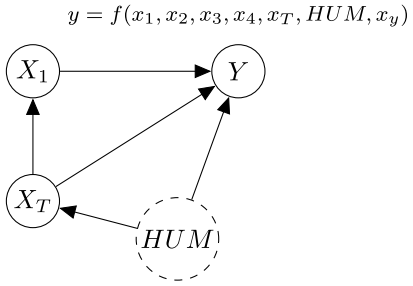
$$P(F\&E^*_{X_4^*=0.5, I=True} = 1 | F\&E_{X_4=\infty, I=True} = 0) = 0.09$$

This is larger (almost double) than the initial probability in the factual world (i.e., 0.05).

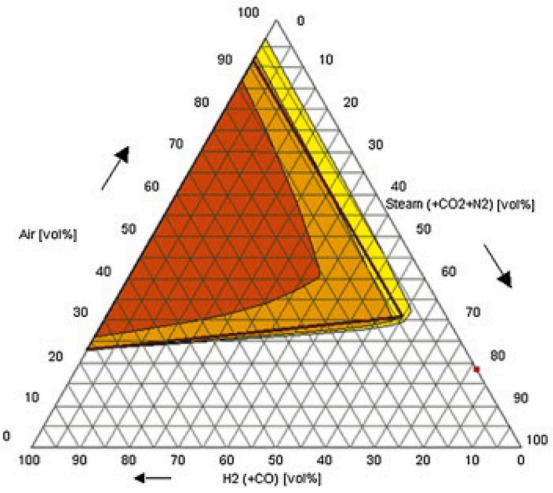
The effect of action and abduction on Y is summarized in Fig. 15: they change the mean value on the distribution of Y, which entails a different probability of exceeding the risk threshold. Based on these estimations, we can conclude that the scenario is very sensitive to the available space, whereby the event of Fire&Explosion deserves specific investigations focused on the hydrogen release in tunnels.

6. Conclusions

In this work, we have proposed to integrate STPA with SCM for building a framework that can improve both the identification of risks and their understanding, while providing a sound base for quantification and insights to build specific simulation models. This is particularly relevant for the risks deriving from the use of new technologies such as Hydrogen. Specifically, STPA is used for the qualitative identification of the system hazards, threats and loss scenarios through a thorough and systematic investigation of cause–effect relationships between objects (components, parameters and features) and functional features (properties, constraints and performance indexes), whereas SCMs are used to

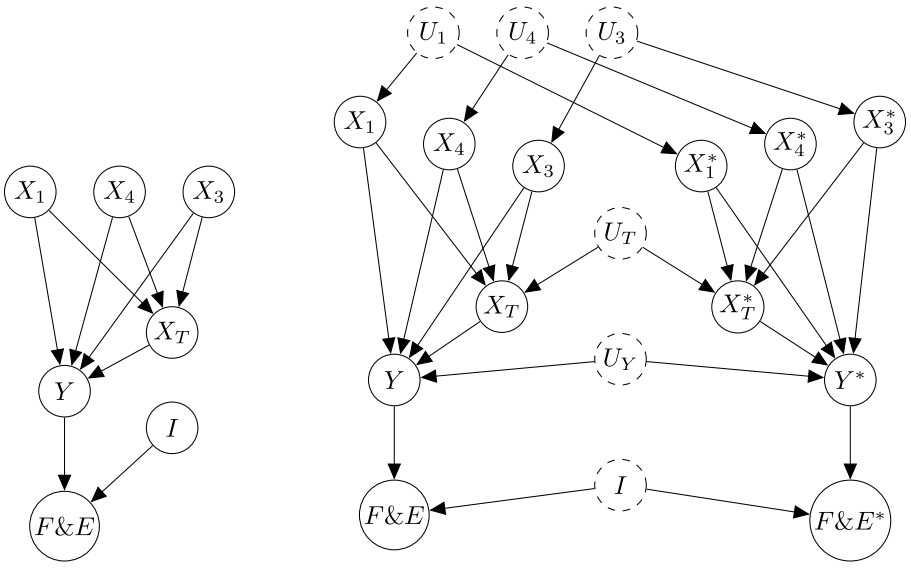


(a) Humidity as a confounder between X_T and outcome Y.



(b) Shapiro diagram for Hydrogen-air-steam mixture. [56]

Fig. 13. Humidity as a confounder: causal and Shapiro models (see [56]).



(a) SCM. (b) Twin Network.

Fig. 14. Fire&Explosion SCM, extracted from Fig. 10.

structure the cause–effect relations between system objects, exogenous conditions and relevant variables to identify and represent hazardous scenarios. Specifically, they are used to:

- Combine past experience and background knowledge to represent the analyst's understanding of the system logic to replace the complex and time-consuming modeling and simulation approaches with figurative models able to capture the scenario dynamics.
- Investigate the existence of confounding factors affecting the causal mechanisms between the components to identify and characterize undiscovered system vulnerabilities, i.e., a priori unconsidered scenario behaviors. This improves the risk understanding.
- Estimate counterfactual outcomes probabilities. The proposed methodology applies counterfactual reasoning to system safety and provides holistic approaches to "Learning from Events" strategic techniques (e.g. Root Cause Analysis) and safety barriers critical analyses.

The proposed framework has been applied to a case study concerning the risks related to the circulation of H2-powered trains. The approach was found very useful by experts.

This work paves the way for future research work to investigate the additional benefits of fostering SCM in risk analysis. These benefits include the possibility of both introducing novel importance measures that are based on the causal role played by each component in the causal scenario and give a more sound basis to the concept of cause of an accident, based on the formal definition of actual causes [57].

Table 5
CPTs.

X(work)				0.8
X(fail)				0.2
		X(work)		X(fail)
Y(work)		0.7		0.4
Y(fail)		0.3		0.6
		Y(work)		Y(fail)
S_1(work)		0.9		0.6
S_1(fail)		0.1		0.4
		Y(work)		Y(fail)
S_2(work)		0.7		0.3
S_2(fail)		0.3		0.7
	S_1(work) S_2(work)	S_1(work) S_2(fail)	S_1(fail) S_2(work)	S_1(fail) S_2(fail)
$T < T_C$	0.9	0.7	0.8	0.4
$T > T_C$	0.1	0.3	0.2	0.6

Table 6
Abduction. Update the graph nodes probabilities on the basis of the evidence $e : T > T_{MAX}$.

X(work)	0.7143	Y(work)	0.2286	S_2(work)	0.1964	$T^* < T_C$	0.6496
X(fail)	0.2857	Y(fail)	0.7714	S_2(fail)	0.8036	$T^* > T_C$	0.3504

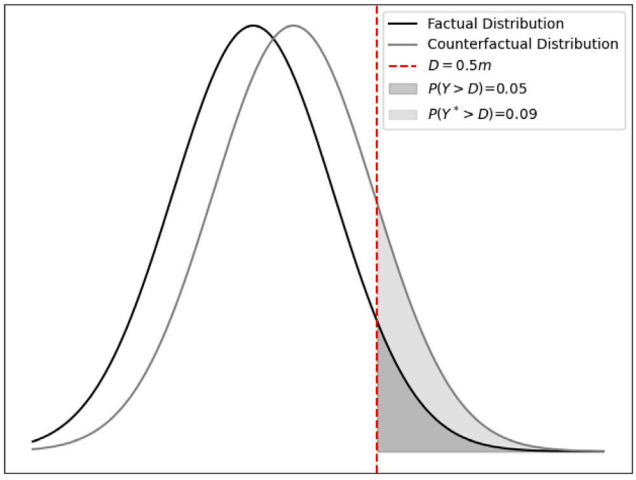


Fig. 15. Effect on Y of the counterfactual query.

CRediT authorship contribution statement

L. Riccardi: Writing – review & editing, Writing – original draft, Supervision, Methodology, Investigation, Conceptualization. **M. Compare:** Writing – review & editing, Writing – original draft, Supervision, Methodology, Investigation, Conceptualization. **R. Mascherona:** Writing – review & editing, Writing – original draft, Supervision, Methodology, Investigation, Conceptualization. **E. Zio:** Writing – review & editing, Writing – original draft, Supervision, Methodology, Investigation, Conceptualization.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Appendix

See Tables 5 and 6.

Data availability

No data was used for the research described in the article.

References

[1] Energy policy act of 1992: To provide for improved energy efficiency. U.S. Government Publishing Office; 1992.
[2] USDepartment of Energy. Energy efficiency and renewable energy - alternative fuels data center.
[3] Barbosa F, Sayed TE, Heid B, Majiti R. What is hydrogen energy? McKinsey Explain 2023.
[4] Zio E. The future of risk assessment. Reliab Eng Syst Saf 2018;177:176–90.
[5] Xie Q, Zhou T, Wang C, Zhu X, Ma C, Zhang A. An integrated uncertainty analysis method for the risk assessment of hydrogen refueling stations. Reliab Eng Syst Saf 2024;248:110139.
[6] Zio E. An introduction to the basics of reliability and risk analysis, vol. 13, World scientific; 2007.
[7] Modarres M. Risk analysis in engineering: Techniques, tools, and trends. CRC Press; 2016.
[8] Rausand M. Reliability of safety-critical systems: Theory and applications. Wiley; 2014.
[9] Aven T. A risk science perspective on the discussion concerning Safety I, Safety II and Safety III. Reliab Eng Syst Saf 2022;217:108077.
[10] Bjerga T, Aven T, Zio E. Uncertainty treatment in risk analysis of complex systems: The cases of STAMP and FRAM. Reliab Eng Syst Saf 2016;156:203–9.
[11] Leveson NG. Engineering a safer world: Systems thinking applied to safety. The MIT Press; 2012.
[12] Hollnagel E. Safety-I and safety-II: the past and future of safety management. CRC Press; 2018.
[13] Steen R. On the application of the Safety-II concept in a security context. Eur J Secur Res 2019;4.
[14] Heinrich H. In: Blanchard R, editor. Industrial accident prevention: a scientific approach 1931. Half-title: mcGraw-hill insurance series, McGraw-Hill; 1931.
[15] Reason J. Managing the risks of organizational accidents. Routledge; 2016.
[16] Leveson NG, Thomas JP. STPA handbook. Cambridge, MA, USA; 2018.
[17] Antonello F, Buongiorno J, Zio E. A methodology to perform dynamic risk assessment using system theory and modeling and simulation: Application to nuclear batteries. Reliab Eng Syst Saf 2022;228:108769.
[18] Wróbel K, Montewka J, Kujala P. Towards the development of a system-theoretic model for safety assessment of autonomous merchant vessels. Reliab Eng Syst Saf 2018;178:209–24.
[19] Cheng T, Utne IB, Wu B, Wu Q. A novel system-theoretic approach for human-system collaboration safety: Case studies on two degrees of autonomy for autonomous ships. Reliab Eng Syst Saf 2023;237:109388.
[20] Lu Y, Zhang S-G, Tang P, Gong L. STAMP-based safety control approach for flight testing of a low-cost unmanned subscale blended-wing-body demonstrator. Saf Sci 2015;74:102–13.

[21] Read G, Naweed A, Salmon P. Complexity on the rails: A systems-based approach to understanding safety management in rail transport. Reliab Eng Syst Saf 2019;188:352–65.
[22] Bensaci C, Zennir Y, Pomorski D, Innal F, Lundteigen MA. Collision hazard modeling and analysis in a multi-mobile robots system transportation task with STPA and SPN. Reliab Eng Syst Saf 2023;234:109138.
[23] https://psas.scripts.mit.edu/home/materials/. [Accessed July 2024].
[24] Sun L, Li Y-F, Zio E. Comparison of the HAZOP, FMEA, FRAM, and STPA Methods for the Hazard Analysis of Automatic Emergency Brake Systems. ASCE-ASME J Risk Uncertain Eng Syst B: Mech Eng 2022;8(3).
[25] Heikkilä E, Malm T, Sarsama J, Tiusanen R, Ahonen T. Hazard analysis of an autonomous container handling system – a comparison of STPA and HAZOP methods. Sci J Gdynia Marit Univ 2023;(125/23):25–39.
[26] Pearl J. Causality: Models, Reasoning and Inference. 2nd ed.. USA: Cambridge University Press; 2009.
[27] Occupational health and safety management systems — Requirements with guidance for use. ISO 45001:2018, Geneva, CH: International Organization for Standardization; 2018.
[28] Khakzad N, Khan F, Amyotte P. Dynamic safety analysis of process systems by mapping bow-tie into Bayesian network. Process Saf Environ Prot 2013;91(1):46–53.
[29] Mancuso A, Compare M, Salo A, Zio E. Portfolio optimization of safety measures for reducing risks in nuclear systems. Reliab Eng Syst Saf 2017;167:20–9, URL https://www.sciencedirect.com/science/article/pii/S0951832016305051, Special Section: Applications of Probabilistic Graphical Models in Dependability, Diagnosis and Prognosis.
[30] Veitch V, Zaveri A. Sense and sensitivity analysis: Simple post-hoc analysis of bias due to unobserved confounding. 2020, arXiv:2003.01747 [stat.ME].
[31] Pearl J. Causal and counterfactual inference. Handb Ration 2021;427.
[32] Ruiz-Tagle A, Lopez-Droguett E, Groth KM. A novel probabilistic approach to counterfactual reasoning in system safety. Reliab Eng Syst Saf 2022;228:108785.
[33] Leveson N. A new accident model for engineering safer systems. Saf Sci 2004;42(4):237–70.
[34] Lee SH, Shin S-M, Hwang JS, Park J. Operational vulnerability identification procedure for nuclear facilities using STAMP/STPA. IEEE Access 2020;8:166034–46.
[35] Peters J, Janzing D, Schlkopf B. Elements of causal inference: Foundations and learning algorithms. The MIT Press; 2017.
[36] Neal B. Introduction to causal inference from a machine learning perspective. 2023, https://www.bradyneal.com/Introduction_to_Causal_Inference-Dec17_2020-Neal.pdf.
[37] Pearl J, Mackenzie D. The book of why. New York: Basic Books; 2018.
[38] Bareinboim E, Correa JD, Ibeling D, Icard T. On Pearl's hierarchy and the foundations of causal inference. In: Probabilistic and causal inference: the works of judea pearl. 1st ed.. New York, NY, USA: Association for Computing Machinery; 2022, p. 507–56.
[39] Vlontzos A, Kainz B, Gilligan-Lee CM. Estimating categorical counterfactuals via deep twin networks. 2023, arXiv:2109.01904 [cs.LG].
[40] Holland PW. Statistics and causal inference. J Amer Statist Assoc 1986;81(396):945–60.
[41] Chernozhukov V, Chetverikov D, Demirer M, Duflo E, Hansen C, Newey W, Robins J. Double/debiased machine learning for treatment and structural parameters. Econom J 2018;21(1):C1–68.
[42] Wager S, Athey S. Estimation and inference of heterogeneous treatment effects using random forests. J Amer Statist Assoc 2018;113(523):1228–42.
[43] Murphy KP. Probabilistic machine learning: Advanced topics. MIT Press; 2023.
[44] Manski CF. Partial identification of probability distributions. Springer-Verlag; 2003.
[45] Cinelli C, Kumor D, Chen B, Pearl J, Bareinboim E. Sensitivity analysis of linear structural causal models. In: Chaudhuri K, Salakhutdinov R, editors. Proceedings of the 36th international conference on machine learning. Proceedings of machine learning research, Vol. 97, PMLR; 2019, p. 1252–61.
[46] Cinelli C, Hazlett C. Making Sense of Sensitivity: Extending Omitted Variable Bias. J R Stat Soc Ser B Stat Methodol 2019;82(1):39–67.
[47] Chernozhukov V, Cinelli C, Newey W, Sharma A, Syrgkanis V. Long story short: Omitted variable bias in causal machine learning. Working paper 30302, National Bureau of Economic Research; 2022.
[48] Yadlowsky S, Namkoong H, Basu S, Duchi J, Tian L. Bounds on the conditional and average treatment effect with unobserved confounding factors. 2022, arXiv:1808.09521v5 [stat.ME].
[49] Liu W, Kuramoto S, Stuart E. An introduction to sensitivity analysis for unobserved confounding in non-experimental prevention research. Prev Sci: Off J Soc Prev Res 2013;14.
[50] Groth KM, Mosleh A. Deriving causal Bayesian networks from human reliability analysis data: A methodology and example model. Proc Inst Mech Eng O: J Risk Reliab 2012;226(4):361–79.
[51] van der Borst M, Schoonakker H. An overview of PSA importance measures. Reliab Eng Syst Saf 2001;72(3):241–5.
[52] Glymour C, Zhang K, Spirtes P. Review of causal discovery methods based on graphical models. Front Genet 2019;10:524.
[53] Nogueira AR, Gama J, Ferreira CA. Causal discovery in machine learning: Theories and applications. J Dyn Games 2021;8(3).
[54] Tamascelli N, Dal Pozzo A, Scarponi GE, Paltrinieri N, Cozzani V. Assessment of safety barrier performance in environmentally critical facilities: Bridging conventional risk assessment techniques with data-driven modelling. Process Saf Environ Prot 2024;181:294–311.
[55] IEC 60079-10-1. 3rd ed.. IEC; 2020, International Standard. Explosive atmospheres - Part 10-1: Classification of areas - Explosive gas atmospheres.
[56] Bentaib A, Meynet N, Bleyer A. Overview on hydrogen risk research and development activities: Methodology and open issues. Nucl Eng Technol 2015;221.
[57] Halpern JY. Actual causality. The MIT Press; 2016.