Reliability Engineering and System Safety 228 (2022) 108785

Contents lists available at ScienceDirect

Reliability Engineering and System Safety

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

A novel probabilistic approach to counterfactual reasoning in system safety

Andres Ruiz-Tagle [a,*], Enrique Lopez-Droguett [b], Katrina M. Groth [a]

[a] Department of Mechanical Engineering, Center for Risk and Reliability, University of Maryland, College Park, 20742, MD, United States
[b] Department of Civil and Environmental Engineering & Garrick Institute for the Risk Sciences, University of California, Los Angeles, 90095, CA, United States

ARTICLE INFO

Keywords:
Counterfactuals
Causality
Bayesian network
System safety
Accident analysis

ABSTRACT

Safety–critical systems cannot afford to wait for data from multiple high-consequence events to become available in order to inform safety recommendations. Counterfactual reasoning has been widely used in system safety to address this issue, enabling the incorporation of evidence from single events with an analyst's current knowledge of a system to learn from past events. However, current counterfactual methods have been criticized for making analysts prone to linearizing and oversimplifying complex events. In order to overcome these limitations, this work establishes a novel probabilistic approach to counterfactual reasoning called "possible worlds" counterfactuals. This methodology enables the integration of an analyst's causal knowledge about a system (in the form of a Bayesian network-based risk assessment model) with the best available evidence about an event of interest (e.g., an accident). As a result, counterfactual hypotheses, commonly used in the practice of system safety, can now be rigorously assessed through causally-sound probabilistic methods. We demonstrate the capabilities of "possible worlds" counterfactuals with a real-world case study on the 2018 Sun Prairie gas explosion and show how this approach can provide additional lessons and insights beyond those provided by authorities at the time of the event.

1. Introduction

In Safety–critical systems such as oil & gas pipelines and infrastructure, transportation, and power plants, it is essential to learn from every event; we cannot afford to wait for data from multiple high-consequence events to become available in order to inform safety recommendations [1]. System safety researchers and practitioners have explored different ways to tackle this issue, from which a *"learning from events"* strategy is considered fundamental to the field [2–4]. The strategy's main objective is the search of safety improvement opportunities informed by evidence on single events. Investigation reports then make recommendations on areas of improvement. These recommendations are usually based on rectifying the identified causes of the undesired event, and counterfactual reasoning is widely used to identify these causes [5,6].

Counterfactual reasoning is a type of causal reasoning concerning hypotheses and queries on possible realizations of a past event (i.e., what could have happened, but did not). The use of counterfactual reasoning in system safety has been historically based, at least implicitly, on qualitative "but-for" investigations [1]; that is, finding the necessary causes of an event through statements of the form *"but-for this cause, the event could have been avoided"*. A number of examples can be found in official accident and incident reports, such as the U.S. NTSB's investigations into the Asiana Flight 214 [7] and the U.S. Air Flight 1016 accidents [8]. Both reports used counterfactuals to imply that the accidents could have been avoided, *"but-for the lack of pilot training"* for the Asiana Flight 214, and *"but-for the misjudgment of the crew regarding weather conditions"* for the U.S. Air Flight 1016. In addition, counterfactuals are a central feature of several guidelines and formal analysis techniques for incidents and accidents. For example, Branford and Hopkins recommended using "but-for" counterfactuals to identify a causal structure for Accimap analysis [9]. Likewise, Ladkin's popular Why-Because Analysis method relies on a causal network constructed through "but-for" counterfactual statements [10]. Moreover, counterfactual reasoning has been recommended to safety practitioners as a heuristic for root cause analysis in manuals and guidelines [6,11].

Some researchers, however, have criticized current counterfactual reasoning methods in system safety. These criticisms can be summarized into the following limitations: (1) *Linearity*: "but-for" counterfactuals can make investigators prone to turn complex events into linear cause-and-effect chains [12,13]; (2) *Incompleteness*: event investigators tend to use counterfactuals to transform an event's evidence into a proof of failure to perform according to a system's procedures and norms, therefore missing additional underlying reasons for the occurrence of an event [12,14]. We add a third limitation (3) "*Uncomparable*": there is no rigorous method to compare and prioritize

[*] Correspondence to: 0151.C Glenn L. Martin Hall, 4298 Campus Drive, College Park, MD 20742, USA.
E-mail address: aruiztag@umd.edu (A. Ruiz-Tagle).

https://doi.org/10.1016/j.ress.2022.108785
Received 28 April 2022; Received in revised form 6 August 2022; Accepted 21 August 2022
Available online 28 August 2022

Notation

V_i	Relevant variable of a system's model.
$V_x = v_x$	"factual world" antecedent variable and state.
$V_y = v_y$	"factual world" consequent variable and state.
$V_x = v'_x$	"possible world" antecedent variable and state.
$V_x = v'_y$	"possible world" consequent variable and state.
$do(\cdot)$	do-operator for intervention reasoning.
M	Bayesian network model.
$M_{V_i=v_i}$	Intervened Bayesian network model through $do(V_i = v_i)$.
M^c	Pre-intervention twin network model of a counterfactual.
$M^c_{\hat{V}_x=v'_x}$	Twin network model of a counterfactual.
Q	Counterfactual query.
$\{V_e = v_e\}$	Set of "factual world" evidence from a system event.
$\hat{V}_i$	"possible world" variable in a twin network model of a counterfactual.
W	Factual world nodes in a twin network model of a counterfactual.
B	Common background nodes in a twin network model of a counterfactual.
PW	Possible world nodes in a twin network model of a counterfactual.

counterfactual hypotheses on event investigations. Consequently, current counterfactual methods limit the bounds of what can be learned from a past event.

Nevertheless, "but-for" counterfactuals are still supported by researchers who pose them as a crucial method for incorporating evidence from single events into their current knowledge of a system to learn from them and to inform safety recommendations [1,15,16]. In fact, "but-for" counterfactuals are currently being used to inform causal relationships on the construction of modern systemic methods for accident investigations such as Accimaps and HFACS [9,17]. However, some authors contend that the limitations of current counterfactual methods still outweigh their benefits [12].

Although we agree that the aforementioned shortcomings of counterfactuals are well justified, we argue that these criticisms do not apply to counterfactual reasoning *per se*, but specifically for their use on qualitative "but-for"-based accident and incident investigations. The main issue with the "but-for" approach to counterfactual reasoning in system safety is that it is typically used in terms of learning through *attribution*, that is, in trying to infer what caused an event's outcome [18]. Instead, we hold that counterfactual reasoning should be used to enable investigators to learn from a single event by generating and analyzing counterfactual scenarios that are motivated by their hypotheses on why the event happened the way it did. In order to do this, counterfactual reasoning must explicitly acknowledge the stochastic nature of history; that is, a past event is only one of the many possible realizations that could have unfolded. As such, we hold that counterfactual hypotheses can only be rigorously studied through the use of quantitative risk assessment (QRA) techniques since QRA provides a comprehensive characterization of the risks of a system, its associated uncertainties, and different scenarios and events evaluations that can be studied through probabilistic inference methods [19].

This work establishes a new perspective on counterfactual reasoning for system safety by going beyond "but-for" investigations towards a probabilistic "possible worlds" approach. This approach aims to formally integrate the knowledge from a risk assessment model of a system with the best available evidence on a past event to quantitatively simulate counterfactual scenarios of the event. By doing so, "possible worlds" counterfactual reasoning enhances a *"learning from events"* system safety strategy by enabling investigators to assess their current qualitative counterfactual hypotheses on a past event through a rigorous quantitative evaluation.

To enable a probabilistic "possible worlds" approach to counterfactual reasoning, we present a first-of-its-kind system safety-oriented methodology to assess counterfactual hypotheses through Bayesian network models. Bayesian networks were selected due to their wide popularity in both the system safety [20–22] and risk assessment [23,24] literature, which stems from their ability to integrate multiple sources of information and probabilistically model a complex system's risk-influencing factors and causal dependencies. In addition, Bayesian networks are capable of performing probabilistic reasoning under uncertainty [25–27], making them an ideal framework for capturing an event's complexity and reason about its potential counterfactual realizations. We demonstrate the proposed "possible worlds" approach to counterfactual reasoning through a case study on the 2018 Sun Prairie gas explosion in the U.S. [28], showing how "possible worlds" counterfactuals can be used to inform safety recommendations that overcome the limitations of current counterfactual analyses.

The rest of this work is structured as follows: Section 2 presents a background on counterfactuals and the relevant methods used in this work. Section 3 describes the proposed methodology for enabling a "possible world" approach to counterfactual reasoning in system safety. This methodology is applied to a real-world case study on the Sun Prairie gas explosion in Section 4. Section 5 provides a discussion on the implications of "possible world" counterfactuals in system safety. This work ends with concluding remarks in Section 6.

2. Relevant methods

This section presents the relevant definitions and methods needed to develop a methodology to perform "possible worlds" counterfactual reasoning in system safety. First, we provide a review on the constitutive elements of counterfactual hypotheses and queries that will be used throughout this work. Second, causal Bayesian network models and associated intervention reasoning methods, the backbones that supports counterfactual reasoning, are presented. Last, we explain the graphical twin network approach to counterfactual modeling with Bayesian networks, which enables the probabilistic assessment of counterfactual hypothesis and queries.

2.1. Counterfactuals

Take an event V_y, in which $V_y = v_y$ occurred. A counterfactual hypothesis looks for an answer to queries of the type *"Could the consequent state of V_y be v'_y in an event, instead of the observed v_y, had the antecedent state of V_x been v'_x, instead of the observed v_x?"* Two "worlds" can be identified in this query. First, a *"factual world"* that represents what actually happened in the event, which is composed of the states of the antecedent $V_x = v_x$ and consequent $V_y = v_y$. Second, a *"possible world"* (also called *counterfactual world*) that indicates what could have happened to the "factual world" consequent, namely $V_y = v'_y$ instead of $V_y = v_y$, had the antecedent been different, namely $V_x = v'_x$ instead of $V_x = v_x$. The factual and possible worlds of a counterfactual query share a common set of background conditions; that is, the variables in an event that are assumed to be causally independent from the factual and possible worlds [29,30]. This notation will be used for the rest of this work.

Counterfactual reasoning in system safety has traditionally relied on a "but-for" approach. This approach is used to identify the causes of an event by testing the validity of a counterfactual hypothesis of the form:

but-for the antecedent $V_x = v_x$, the consequent $V_y = v_y$ would have been $V_y = v'_y$ in a "possible world". For technical causes, logic is often used to test the validity of the "but-for" statement, while human-related causes are often assessed through expert judgment [1]. An example of "but-for" counterfactuals can be found in many applications, such as in the use of the Swiss cheese model of accident causation, which holds that *but-for* a failure of a safety barrier, an event would not have happened.

The problem of using logic or expert judgment to assess the validity of a "but-for" statement is that it requires a thoroughly understood causal mechanism to support that the consequent state $V_y = v'_y$ would be true in an event had the antecedent been $V_x = v'_x$ in a "possible world". A probabilistic approach is required to this issue in counterfactual reasoning. We hold that it is valid to state that the consequent $V_y = v'_y$ is a *possible* counterfactual outcome of an event had the antecedent been $V_x = v'_x$ in a "possible world" [31]. Therefore, probabilistic inference methods must be used to assess the probability, rather than the certainty, of a counterfactual hypothesis. To ensure the feasibility of a possible counterfactual outcome, it is necessary to assess its compliance with the constraints entailed by the causal model that characterizes the system in which the event of interest happened. As such, it is crucial that counterfactuals are grounded on the current knowledge of this system and its risks. A QRA model therefore serves as an ideal framework in which to study counterfactuals. In fact, a QRA model is built towards the thorough characterization of the scenarios, consequences, and uncertainties that are involved in a system [32,33], with cause-and-effect relationships as the basis of the modeling methods that have been used for this purpose (e.g., fault trees, event trees, and currently, Bayesian networks) [26,32].

The definitions and assumptions presented above are the basis of what we propose as a probabilistic "possible world" approach to counterfactual reasoning in system safety. To enable this approach, the following methods are used.

2.2. Causal Bayesian networks

Bayesian networks are widely used in risk assessment and system safety as a causal model to express the joint probability distribution of a system's events [23,24]. To do so, a Bayesian network model M is used to represent a system's variables $V = \{V_1, \dots, V_n\}$ and their dependencies as the nodes and edges of a Directed Acyclic Graph (DAG), G. Dependencies among variables are modeled as conditional probability distributions, $Pr(V_i \mid V_j)$, which can be quantified through discrete conditional probability tables or continuous probability functions. A basic Bayesian network model is shown in Fig. 1a.

The Bayesian network model M represents the prior joint probability distribution for a system's events. Mathematically, this distribution can be computed using the *factorization formula* [27]:

$$Pr(V_1, \dots, V_n) = \prod_{i=1}^{n} Pr(V_i \mid pa(V_i)) \tag{1}$$

where $pa(V_i)$ corresponds to the "parent nodes" of V_i; that is, all nodes in $G \in M$ with an outgoing edge into the node V_i. Further, a causal Bayesian network model assumes that $pa(V_i)$ causes V_i.

The evidence obtained after the investigation of a particular system event (e.g., an incident or accident) can be easily incorporated into the Bayesian network as a set of specific states of the system's variables, denoted as $\{V_e = v_e\}$. Then, Bayes theorem can be used to obtain a posterior distribution for a system's events, $Pr(V_1, \dots, V_n \mid \{V_e = v_e\})$. In system safety, this process of evidence updating is named *associative reasoning* [25], and is mainly used to provide associative insight such as "if we observe an excavation being performed by an untrained excavator, we expect to observe a damage to underground utilities with a probability of 0.5".

2.3. Modeling interventions

As we showed in our previous work [25], Bayesian networks can also be used to model the effect of interventions on a system, which will be required for the probabilistic assessment of counterfactual hypotheses. An intervention is equivalent to enforcing a fixed value, state, or distribution, to a variable (or set of variables) V_j. As such, an intervention estimates the causal effect of a variable on another by blocking any correlational influence from their potential common causes. Eq. (1) only guarantees *associative* insights regarding what is expected to be observed in a system when new evidence is observed from an event. On the other hand, *intervention* modeling can be used to get causal insights on a system such as "if an excavator is trained, underground utility damage is expected to decrease by 50%". To enable intervention reasoning in Bayesian networks, the joint probability distribution for a system's events conditioned by an intervention $V_j = v_j$ is computed by modifying Eq. (1) into the *truncated factorization formula* [27]:

$$Pr\left(V_1, \dots, V_n \,\middle|\, do(V_j = v_j)\right) = \prod_{i|V_i \notin V_j} Pr\left(V_i \mid pa(V_i)\right)\Big|_{V_j = v_j} \tag{2}$$

where $do(\cdot)$ is the *do*-operator, which simulates an intervention in a Bayesian network model by fixing the value of a node and removing all incoming edges to it[1] [27]. The subsequent Bayesian network with removed edges is referred as a submodel $M_{V_j = v_j}$ of the original Bayesian network model M.

As shown in the next section, interventions allow the simulation of the "possible world" in a counterfactual model by forcing the antecedent variables to have a different state in a "possible world" from the ones observed in the "factual world".

2.4. The twin network representation of counterfactuals

In order to represent and evaluate counterfactual hypotheses and queries on a system's event, Pearl's twin network graphical approach to counterfactuals [27,34] will be used in the context of Bayesian network models. Twin networks graphically represent Pearl's perspective on counterfactual reasoning to predict the possible world consequent of an event given a hypothetical possible world antecedent. This perspective can be explained in three steps [27]. First, the factual world evidence on an event's antecedent and consequent is used to update the past information on the event's background variables. Second, the course of history is bent to comply with the hypothetical possible world antecedent. Third, the possible world consequent can be predicted based on the new understanding of the past information on the event's background variables and the newly established possible world antecedent.

To illustrate the twin network approach, consider the Bayesian network model of Fig. 1a. The nodes of this network are "V_1: Maintenance", "V_2: Safety barrier", and "V_3: System failure". Consider also a system event in which a system failure and a safety barrier failure were observed (i.e., "V_3: System failure = yes" and "V_2: Safety barrier= fail" in Fig. 1a). In an event investigation, a counterfactual hypothesis could be "had the safety barrier worked, the failure could have been avoided". This hypothesis looks for an answer to the query "*Could the system's failure been avoided had the safety barrier worked?*" Fig. 1b shows a twin network representation of the query. Given that we are interested in the effect that "V_2 = work" could have had on the outcome of "V_3", a copy of these two variables, identified with a hat accent,

[1] It is important to highlight that removing the incoming edges to the intervened node is not strictly necessary to simulate an intervention. For instance, an alternative approach (with equivalent results) can be the use of a Boolean switch variable such that conditioning on the switch either sets the intervention on or off. In both cases, the result is a node whose states do not condition the states of its parents.

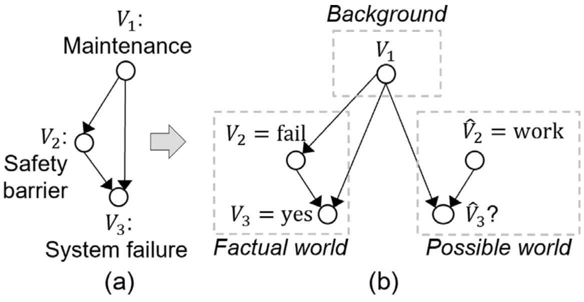


Fig. 1. Twin network model construction [26]. (a) Original Bayesian network model M. (b) Twin network model $M^c_{\hat{V}_3=v'_3}$.

are added to the model. In Fig. 1b, the portion of the network that includes the antecedent and consequent $\{V_2, V_3\}$ represents the "factual world", while $\{\hat{V}_2, \hat{V}_3\}$ represents the "possible world". In the "possible world" portion of the network, all incoming edges to the antecedent $\hat{V}_2$ are removed, thereby emulating the action of the do-operator. This is done to force the antecedent of the "possible world" to take a different value to the one observed in the "factual world" (i.e., "$\hat{V}_2$ = work" instead of the observed "V_2 = fail"). Last, the variables whose posterior probabilities remain invariant in both worlds, such as V_1 in Fig. 1a, represent the background conditions of the event. For the rest of this work, a twin network model will be denoted by $M^c_{\hat{V}_x=v'_x}$, where $\hat{V}_x = v'_x$ represents the "possible world" antecedent of interest.

It is important to note that twin network theory has been mainly developed for functional models [27]. Nevertheless, works by Fenton and Neil have successfully explored its use in Bayesian networks for health risks applications [26,35,36]. Although their research has demonstrated the value of performing counterfactual reasoning with Bayesian networks for decision-making support, their work is unrelated to system safety and does not provide the requirements or guidelines needed to construct a causally-sound twin network model from a Bayesian network model; two gaps that are tackled in this work.

2.5. Previous methods used for quantitative counterfactual analyses in risk assessment and system safety

Although scarce, there is relevant literature on system safety and risk assessment that has performed quantitative counterfactual analyses as a part of their work. For instance, Lam and Cruz [37] investigated consumer-level utility gas incidents through a probabilistic network approach to represent cause–effect chains based on historical incident investigations. Here, counterfactual scenarios were modeled by removing specific sets of causes from their incident network to analyze their effect on the likelihood of relevant consequences. A similar approach can be seen in Hughes et al. [38] on a hybrid physics-based and data-driven model for power distribution infrastructure hardening and outage simulation. To analyze the effects of pole hardening on an outage probability in storm events, they set up a counterfactual study by fitting their model to historical storms. Then, they set specific pole hardening levels to estimate if an outage probability could have been decreased. Likewise, Oughton et al. [39] used a similar methodology to set up a counterfactual study on a cyber–physical attack performed in 2015 on the Ukrainian electricity distribution network. Their model represented the Ukrainian network in 2015, and multiple disruption levels were set to estimate their impact on different socioeconomic variables.

We argue that the studies mentioned above use modeling methods that, although used in a counterfactual setting, are not suited for simulating a counterfactual outcome of an event. Rather, these studies should be interpreted as intervention analyses (similar to the methods shown in Section 2.3), in which the calculated outcome corresponds to the expected causal effect of an intervention on a system. There is a crucial difference between counterfactuals and interventions that should be considered in a counterfactual model: all evidence on a past event is included in the model, which can contain information on the variable that we want to intervene in and the outcome of interest, that is, the event's "factual world" antecedent and consequent. As it was explained in Section 2.4, the evidence on the "factual world" must be used to update the probability of the states of the variables representing the background conditions in which an event happened. However, the methods used in the works presented above do not incorporate the information provided by an event's "factual world" antecedent and consequent to evaluate counterfactual outcomes in a "possible world". The methodology proposed in Section 3 addresses this gap through the use of the twin network approach to counterfactuals presented in Section 2.4. For a further discussion on the distinction between interventions and counterfactuals, and their implications for risk and safety assessment, we refer the reader to [27] and [25,40], respectively.

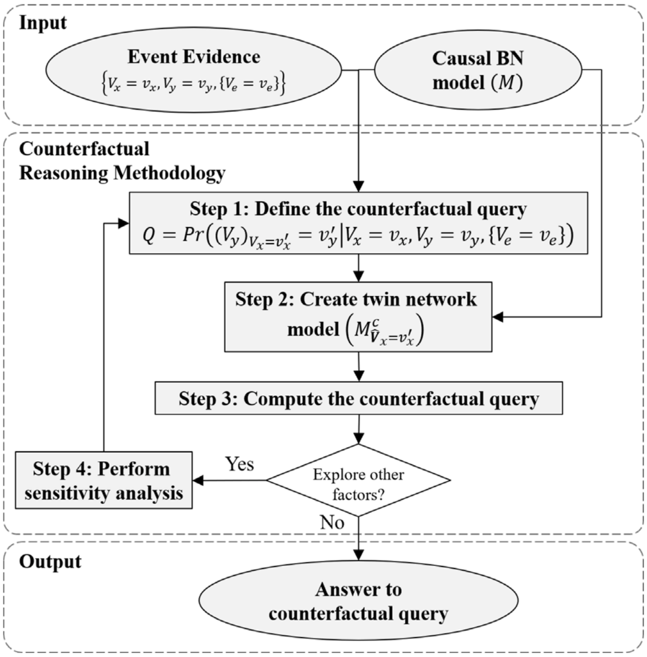

Input
Event Evidence
$\{V_x = v_x, V_y = v_y, \{V_e = v_e\}\}$
Causal BN model (M)
Counterfactual Reasoning Methodology
Step 1: Define the counterfactual query
$Q = Pr\big((V_y)_{V_x=v'_x} = v'_y \mid V_x = v_x, V_y = v_y, \{V_e = v_e\}\big)$
Step 2: Create twin network model $\big(M^c_{\hat{V}_x=v'_x}\big)$
Step 3: Compute the counterfactual query
Explore other factors?
Yes
No
Step 4: Perform sensitivity analysis
Output
Answer to counterfactual query

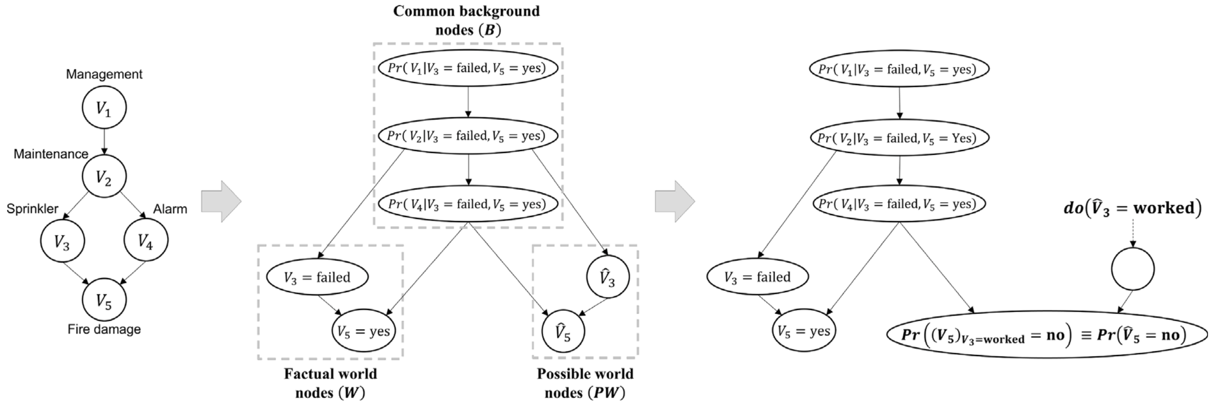
Fig. 2. Proposed methodological framework for counterfactual reasoning. The acronym BN refers to Bayesian network.

3. A methodology to enable a "possible world" approach to counterfactual reasoning in system safety

To enable a "possible world" approach to counterfactual reasoning in system safety, we propose the methodological framework presented in Fig. 2. The framework's objective is to assess the likelihood of counterfactual hypotheses on a past event in a system. To do this, the framework uses the best available evidence about an event together with a causal Bayesian network-based risk assessment model of the system in which the event occurred. The elements of the framework are described as follows.

3.1. Input: Event evidence and causal Bayesian network model

A counterfactual hypothesis studies a possible outcome that a system event could have had, but did not (i.e., an outcome in a "possible world"). Therefore, in order to assess the likelihood of a counterfactual hypothesis, an analyst only has access to the evidence gathered from


Management
V_1
Maintenance
V_2
Sprinkler
Alarm
V_3
V_4
V_5
Fire damage
Common background nodes ($\boldsymbol{B}$)
$Pr(V_1|V_3 = \text{failed}, V_5 = \text{yes})$
$Pr(V_2|V_3 = \text{failed}, V_5 = \text{yes})$
$Pr(V_4|V_3 = \text{failed}, V_5 = \text{yes})$
$V_3 = \text{failed}$
$V_5 = \text{yes}$
$\hat{V}_3$
$\hat{V}_5$
Factual world nodes ($\boldsymbol{W}$)
Possible world nodes ($\boldsymbol{PW}$)
$Pr(V_1|V_3 = \text{failed}, V_5 = \text{yes})$
$Pr(V_2|V_3 = \text{failed}, V_5 = \text{Yes})$
$Pr(V_4|V_3 = \text{failed}, V_5 = \text{yes})$
$do(\hat{V}_3 = \text{worked})$
$V_3 = \text{failed}$
$V_5 = \text{yes}$
$Pr\left((V_5)_{V_3=\text{worked}} = \text{no}\right) \equiv Pr(\hat{V}_5 = \text{no})$

(a) System BN model - $\boldsymbol{M}$

(b) Pre-intervention twin network model - $\boldsymbol{M^c}$

(c) Intervened twin network model - $\boldsymbol{M^c_{\hat{V}_3=\text{worked}}}$

Fig. 3. Simple illustrative example of a fire event (a) Bayesian network, (b) pre-intervention twin network, and (c) twin network model for the fire event example. In this example, the following counterfactual query is solved: “Could the fire damage state be “No”, instead of the observed “Yes”, had the sprinkler sate been “worked”, instead of the observed “failed”, in the event?”

the event (i.e., the “factual world”) and their knowledge of the system in which the event happened. As it was explained in Section 2, a Bayesian network-based risk assessment model of a system is a suitable model for expressing an analyst’s expert knowledge of a system and for performing reasoning under uncertainty when evidence becomes available. As such, the input to the counterfactual reasoning methodology of Fig. 2 consists of the tuple:

$$\text{Input} = \left\langle M, \left\{ V_x = v_x, V_y = v_y, \{V_e = v_e\} \right\} \right\rangle \tag{3}$$

where M corresponds to a system’s Bayesian network-based risk assessment model and $\{V_x = v_x, V_y = v_y, \{V_e = v_e\}\}$ corresponds to the set of evidence gathered from a past event in terms of the nodes in M. The specific elements of the evidence set $\{V_x = v_x, V_y = v_y, \{V_e = v_e\}\}$ are explained in Section 3.2.1.

3.2. Counterfactual reasoning methodology

Using the input tuple in Eq. (3), we propose the following four-step methodology to assess the likelihood of counterfactual hypotheses on a system safety context. It is important to note that the explanations presented below are written in terms of answering a counterfactual query regarding a single antecedent variable V_x. This methodology, however, also holds for a query on a set $\{V_x\}$ of antecedent variables.

To exemplify the steps of the proposed counterfactual reasoning methodology, consider a hypothetical event in which a fire damaged a facility. In this example, an investigation found that the sprinkler system did not work at the time of the event. In addition, the event investigators have access to a Bayesian network-based risk assessment model of the fire protection system of the facility, M, as the one shown in Fig. 3a. Therefore, the tuple $\langle M, \{V_5 = \text{yes}, V_3 = \text{failed}, \{V_e = \emptyset\}\}\rangle$ will be used as an input to exemplify the proposed methodological steps for counterfactual reasoning.

3.2.1. Step 1: Define the counterfactual query

First, an analyst transforms their counterfactual hypothesis into a counterfactual query, and expresses it in probabilistic terms. The goal is to express the query in a form that can be answered by setting evidence on the causal Bayesian network model of a system. The evidence is gathered from a past event.

To clearly express a counterfactual’s antecedent and consequent in a query, this must be structured as: *“Given evidence $\{V_e = v_e\}$, could the consequent state of V_y be v'_y, instead of the observed v_y, had the antecedent state of V_X been v'_x, instead of the observed v_x, in the event?”*

In this query, the “factual world” antecedent and consequent variables and states of interest are $V_x = v_x$ and $V_y = v_y$, respectively. Likewise, the “possible world” antecedent and consequent are $V_x = v'_x$ and $V_y = v'_y$, respectively. In addition, all evidence on the “factual world” regarding other set of variables relevant to the event but different from V_x and V_y are expressed as $\{V_e = v_e\}$. Therefore, a counterfactual query can be translated into a probabilistic expression as:

$$Q = Pr\left((V_y)_{V_x=v'_x} = v'_y \,\middle|\, V_x = v_x, V_y = v_y, \{V_e = v_e\} \right) \tag{4}$$

where Q is the query of interest and $(V_y)_{V_x=v'_x} = v'_y$ is the possible outcome state of the consequent V_y had $V_x = v'_x$ in the event.

To illustrate this step through the fire damage example, consider that, after gathering evidence on the event, investigators hypothesized that the fire could have been mitigated had the sprinklers worked. Following Step 1 of the proposed counterfactual reasoning methodology, this hypothesis can be transformed to the counterfactual query: “Could the fire damage state be “No”, instead of the observed “Yes”, had the sprinkler sate been “worked”, instead of the observed “failed”, in the event?” Then, the input tuple $\langle M, \{V_5 = \text{yes}, V_3 = \text{failed}, \{V_e = \emptyset\}\}\rangle$ can be used with Eq. (4) to generate the counterfactual query $Q = Pr((V_5)_{V_3=\text{worked}} = \text{no} \mid V_3 = \text{failed}, V_5 = \text{yes})$.

3.2.2. Step 2: Create a twin network model

The term $(V_y)_{V_x=v'_x} = v'_y$ in Eq. (4) corresponds to an unobserved counterfactual outcome of a past event. Thus, computing Eq. (4) through traditional probabilistic methods is not straightforward. This computation is instead done with a twin network model as shown in Section 2.4. Therefore, the input causal Bayesian network model of the system, M, is transformed into a twin network model, $M^c_{\hat{V}_x=v'_x}$, of the counterfactual query obtained in Step 1.

The main feature of a twin network model is that it simulates the same background conditions of the “factual world” in an event and allows the transfer of this knowledge to a “possible world” scenario. As explained in the beginning of Section 2, this is a key requirement for assessing the likelihood of a counterfactual hypothesis.

In order to guide the construction of a twin network model $M^c_{\hat{V}_x=v'_x}$ from an initial Bayesian network M, the following three node types are defined:

- *Factual world nodes* (W). Corresponds to the nodes in the system’s Bayesian network model M representing the counterfactual query antecedents $\{V_x\}$, consequents $\{V_y\}$, and all of their direct descendants (i.e., all nodes in M connected by a direct path starting with an outgoing edge from a node in $\{V_x\}$ or $\{V_y\}$). In addition, the “factual world” event evidence on these nodes must be instantiated. If the evidence on the “factual world” antecedent is uncertain or unavailable from an event's investigation, no evidence is instantiated on the node representing it.

- *Common background nodes* (B). Corresponds to all other nodes in M different from those identified as factual world nodes. The "factual world" evidence on these nodes must be instantiated. These nodes represent the common background conditions between the factual and possible worlds in the counterfactual. As such, these nodes will enable the transfer of knowledge gathered on the "factual world" to a counterfactual "possible world".
- *Possible world nodes* (PW). Corresponds to a copy of the nodes in W (and their causal dependencies), but without instantiating the event's evidence. These nodes are differentiated from the ones in W by a hat accent (e.g., if $V_x \in W$, then $\hat{V}_x \in PW$).

The node types defined above are sufficient to construct a "pre-intervention twin network" of a counterfactual query, which will be denoted by M^c. This network encodes all the information needed to perform counterfactual reasoning on a past event. However, the "possible world" antecedent has yet to be forced to be different from what it was in the "factual world". In order to do this, the pre-intervention model M^c is intervened to generate the final twin network model $M^c_{\hat{V}_x = v'_x}$. This is done through the do-operator by forcing $do(\hat{V}_x = v'_x) \in PW$ in a possible world, instead of $V_x = v_x \in W$ as it was observed in the event. Consequently, estimating the probability of $(V_y)_{V_x = v'_x} = v'_y$ in the counterfactual query of Eq. (4) is equivalent to estimating the probability of $\hat{V}_y = v'_y$ in the twin network model $M^c_{\hat{V}_x = v'_x}$.

To illustrate this step, consider the input tuple of the fire event example, $\langle M, \{V_5 = \text{yes}, V_3 = \text{failed}, \{V_e = \emptyset\}\}\rangle$, and the corresponding counterfactual query $Q = Pr((V_5)_{V_3=\text{worked}} = \text{no} \mid V_3 = \text{failed}, V_5 = \text{yes})$ constructed in Step 1. Analysts can identify the factual world nodes as $\{V_3 = \text{failed}, V_5 = \text{yes}\} \in W$; the common background nodes as $\{V_1, V_2, V_4\} \in B$; and the possible world nodes as $\{\hat{V}_3, \hat{V}_5\} \in PW$. Then, the pre-intervention twin network M^c shown in Fig. 3b can be constructed. Last, the intervention $do(\hat{V}_3 = \text{worked})$ is applied to evaluate what could have happened if the sprinkler worked. The intervention eliminates all incoming edges to $\hat{V}_3 \in PW$, thus generating the finalized twin network model $M^c_{\hat{V}_3=\text{worked}}$ shown in Fig. 3c.

3.2.3. Step 3: Compute the counterfactual query

As shown in the previous step, a counterfactual's twin network model $M^c_{\hat{V}_x = v'_x}$ is a post-intervention distribution of M^c in Fig. 3. As such, its joint probability distribution can be expressed through the truncated factorization formula (see Eq. (2)) as:

$$Pr\left(V_i \in \{W \cup B\}, \hat{V}_{i \neq x} \in PW \,\middle|\, do(\hat{V}_x = v'_x)\right) = \prod_{i|\{V_i, \hat{V}_i\} \neq \hat{V}_x} Pr\left(\{V_i, \hat{V}_i\} \,\middle|\, pa\left(\{V_i, \hat{V}_i\}\right)\right)\Big|_{\hat{V}_x = v'_x} \tag{5}$$

Then, if the network variables are discrete, Eq. (5) can be marginalized to compute the counterfactual query Q of Eq. (4) as:

$$Q = \sum_{i|\{V_i, \hat{V}_i\} \neq \{\hat{V}_x, \hat{V}_y\}} \left(\prod_{i|\{V_i, \hat{V}_i\} \neq \hat{V}_x} Pr\left(\{V_i, \hat{V}_i\} \,\middle|\, pa\left(\{V_i, \hat{V}_i\}\right)\right)\Big|_E \right) \tag{6}$$

where $E = \left\{V_x = v_x, V_y = v_y, \{V_e = v_e\}, \hat{V}_x = v'_x\right\}$, that is, all the event's evidence gathered by an investigator and the "possible world" antecedent of interest. Given that Eq. (6) can be computed from a single twin network model $M^c_{\hat{V}_x = v'_x}$, Bayesian network software (e.g., GeNIe [41]) or programming libraries (e.g., Python's pgmpy [42]) can be used for this task. Additionally, Eq. (6) is extendable to continuous nodes through integration instead of summation. However, its computation becomes difficult, requiring specialized algorithms such as dynamic discretization for hybrid Bayesian networks [26].

This step can be illustrated in the fire event example, in which estimating the counterfactual query $Q = Pr((V_5)_{V_3=\text{worked}} = \text{no} \mid V_3 = \text{failed}, V_5 = \text{yes})$ is equivalent to estimating the probability of $\}\}\hat{V}_5 = \text{no}\varepsilon$ in the twin network model $M^c_{\hat{V}_3=\text{worked}}$ of Fig. 3c.

3.2.4. Decision: Explore other factors?

In this step, an analyst decides if the estimated answer of the counterfactual query is enough to inform their analysis and safety recommendations. This decision is not trivial, and should align with the objectives of the analysis. If no further information is needed from counterfactual queries, the analyst proceeds to provide the generated knowledge to risk managers as decision support. An example of this can be illustrated using the fire event example. If the objective of the analysis was only to determine if a working sprinkler could have helped to avoid fire damage, the outcome of Step 3 of the proposed methodology provides enough information. However, a simplistic objective like this is rarely the goal of an event investigation.

To learn the most and inform the best decisions from past events, an analyst is encouraged to include in their analysis' objectives the exploration of potential safety improvement opportunities and to expand their knowledge on a past event through the analysis of multiple counterfactual scenarios. This additional objective has the potential of revealing a number of factors that could have had a relevant influence on the outcome of a past event which might have been overlooked at first glance. In fact, the importance of this additional objective has been stressed by authors such as Woo et al. [15,43] and Oughton et al. [39], which have proposed a downward counterfactual searching approach (that is, an analysis on how an event could have had a worse consequence, such as higher failure probabilities or more severe losses) to elucidate the potential of black swans and extreme events based on past hazardous experiences (we refer to [43] for further information on the counterfactual search approach). Leveraging this objective into quantitative system safety, Step 4 provides a sensitivity analysis-based method for exploring additional factors that could have a relevant effect on the outcome of a past event.

3.2.5. Step 4: Perform a sensitivity analysis

In order to explore other factors that could have had a relevant effect into the outcome of a past event in a counterfactual setting, we propose a "downward" approach to counterfactual reasoning. In particular, we focus on which variables of the system could have increased the probability of an undesired consequence state in a "possible world" scenario. This is done through a Bayesian network-based local sensitivity analysis over the "possible world" consequent $\hat{V}_y = v'_y \in PW$. Then, the most influential "possible world" nodes on the consequent node's sensitivity are selected to be further analyzed through counterfactual reasoning; that is, going back to Step 1 and analyze how they could have affected the outcome of the event of interest.

This step is illustrated on the fire damage example as follows: after computing the result of the counterfactual query "Could the fire damage state be "No", instead of the observed "Yes", had the sprinkler state been "worked", instead of the observed "failed", in the event?" the analysts found that a working sprinkler would have not significantly reduce the probability of fire damage. Consequently, the analysts decide to explore if other safety barriers, such as the correct functioning of the alarm system and the maintenance schedule, could have further reduced the likelihood of the event. Using the twin network $M^c_{\hat{V}_3=\text{worked}}$ built in step 2, in which $\{V_3, V_5\} \in W$, $\{\hat{V}_3, \hat{V}_5\} \in PW$, and $\{V_1, V_2, V_3\} \in B$, a sensitivity analysis is performed over $\hat{V}_5 = \text{yes}$. Imagine that the sensitivity analysis results show that "V_4 : Alarm" is the most influential factor on the counterfactual outcome. As such, analysts can decide on going back to Step 1 of this methodology to answer the counterfactual $Q = Pr((V_5)_{\{V_3=\text{worked}, V_4=\text{worked}\}} = \text{No} \mid V_3 = \text{failed}, V_5 = \text{Yes})$.

3.3. Output: Answer to counterfactual query

The output of the counterfactual reasoning methodology presented above corresponds to the probability of a possible outcome of a past event on a system. Depending on how the Bayesian network model of the system is specified, this result can be in the form of an expected value or a probability distribution. However, we discourage investigators to use these specific values as the sole basis of their safety improvement recommendations. Rather, the two following points must be taken into consideration:

- The power of counterfactual reasoning in system safety lies on assessing the contrasts between an event and other possible realizations of it [44]. Consequently, the output value of a counterfactual query should not be studied on its own, but rather in comparison to other possible outcomes of a past event.
- The output value of a counterfactual query should be studied as a complement to the results obtained through traditional *"learning from events"* safety strategies. For instance, "possible worlds" counterfactual reasoning can be used to assess the distance between models and practice. This can be done by evaluating the differences between the answer of counterfactual queries estimated through the proposed methodology of Section 3.2 and the narrative and recommendations provided in an accident investigation report [45].

These points will be illustrated in the case study presented in Section 4 and further discussed in Section 5.

4. A case study on excavation damage of natural gas pipelines

A leading cause of failure across natural gas pipelines in the U.S. is third-party damage caused by an excavator that has no relationship to a utility company [46,47]. According to the Pipeline and Hazardous Materials Safety Administration (PHMSA), third-party damage was, on average, the cause of 21.4% of excavation incidents on distribution and transmission lines between 2016 and 2021 [46]. These incidents resulted in 11 fatalities, 34 injuries, and $140M USD in property damage.

To prevent excavation damages, the following dig-in best practices are encouraged (and regulated in some states) [48]:

1. The excavator provides notice of intent (also referred as locate request) to an 811 One Call center or submits an online request;
2. The utility operator locates and marks their underground facilities at the excavation site, and;
3. The excavator proceeds to dig carefully as understood in [49].

Although damage prevention practices are well established for underground utilities, third-party damage rates have been roughly constant in time [46]. Due to the significant risks posed by third-party damage, Ruiz-Tagle et al. developed BaNTERA, a comprehensive Bayesian network model for third-party excavation risk assessment [50]. Preliminary results demonstrated how BaNTERA can be used beyond a probability estimation of third-party damage, offering valuable insights into cause–effect relationships, "what-if" scenarios, and identifying and prioritizing ways to prevent and mitigate third-party damage risk.

Currently, safety recommendations for natural gas utilities are generated from either qualitative or quantitative studies. Examples of these are PHMSA's event investigations [51] and the yearly CGA DIRT report [47]. As such, probabilistic "possible world" counterfactuals represent an opportunity to integrate these data with the risk assessment model BaNTERA to learn from past events caused by third-party damage and inform safety recommendations. This section presents a case study that uses BaNTERA and a real-world accident report narrative to demonstrate the capabilities of "possible world" counterfactuals for learning from a past event.

4.1. The 2018 Sun Prairie gas explosion

Around 6 p.m. on July 10, 2018 in Sun Prairie, Wisconsin, U.S., a third-party excavator from VC Tech struck a natural gas main while performing drilling activities to install fiber–optic lines for Bear Communications, LLC. Gas was released, migrating below ground into nearby buildings. One hour later the gas was ignited by an unknown source, provoking an explosion that resulted in 1 fatality, 8 injures, and approximately $20M USD in property damage [51]. A picture of the event is shown in Fig. 4.

Fig. 4. The 2018 Sun Prairie gas explosion.
Source: The Daily Reporter [28].

An investigation by the U.S. Department of Labor's Occupational Safety and Health Administration (OSHA) [52] determined, through an administrative trial, that both Bear Communications, LLC and its contractor, VC Tech, were liable for this event. Each company failed to notify an intent to excavate, which is a mandatory practice by Wisconsin's state law. However, the event was more complicated than the judgment entails. The following points summarize the event's storyline presented by OSHA during the administrative trial, which has been widely covered by specialized media [28]:

- Spring of 2018 — Bear Communications, LLC contracted Jet Underground Drilling Company to perform an excavation for fiber–optic cable installations.
- June 2018 — Jet Underground provided a locate request through a notification to Diggers Hotline, Wisconsin's 811 One Call center. USIC Locating Service performed the location and marking of underground utilities in the site.
- July 2018 — Due to timing issues, Bear Communications LLC decided that a competing drilling company, VC Tech, would proceed with the excavation activities. VC Tech had not notified their intent to excavate to an 811 One Call center, relying on previous markings made for Jet Underground in June. While performing drilling operations, VC Tech strucked a natural gas main, causing an explosion.
- Public hearings and event's aftermath — In public hearings, VC Tech acknowledged that they were knowingly violating state law by excavating without a locate request notification. However, they blamed USIC Locating Service for the event, arguing that the marks performed for Jet Underground were incorrect. This information could not be verified nor refuted. USIC Locating Service answered that the purpose of the hearing was not to blame them, but rather to determine whether VC Tech notified their intent of excavation or not. Moreover, USIC Locating Service argued that, in addition to a lack of notification, VC Tech did not follow dig-in best practices which could have avoided the event even though the site was mismarked.

The investigation performed by OSHA indicates clear liability; the excavators did not comply with state law by failing to notify their intent to excavate. Additionally, a separate investigation by PHMSA determined that the root cause of the event was a lack of notification (see event #20180073 in [51]); the lesson seems clear, *"but-for the lack*

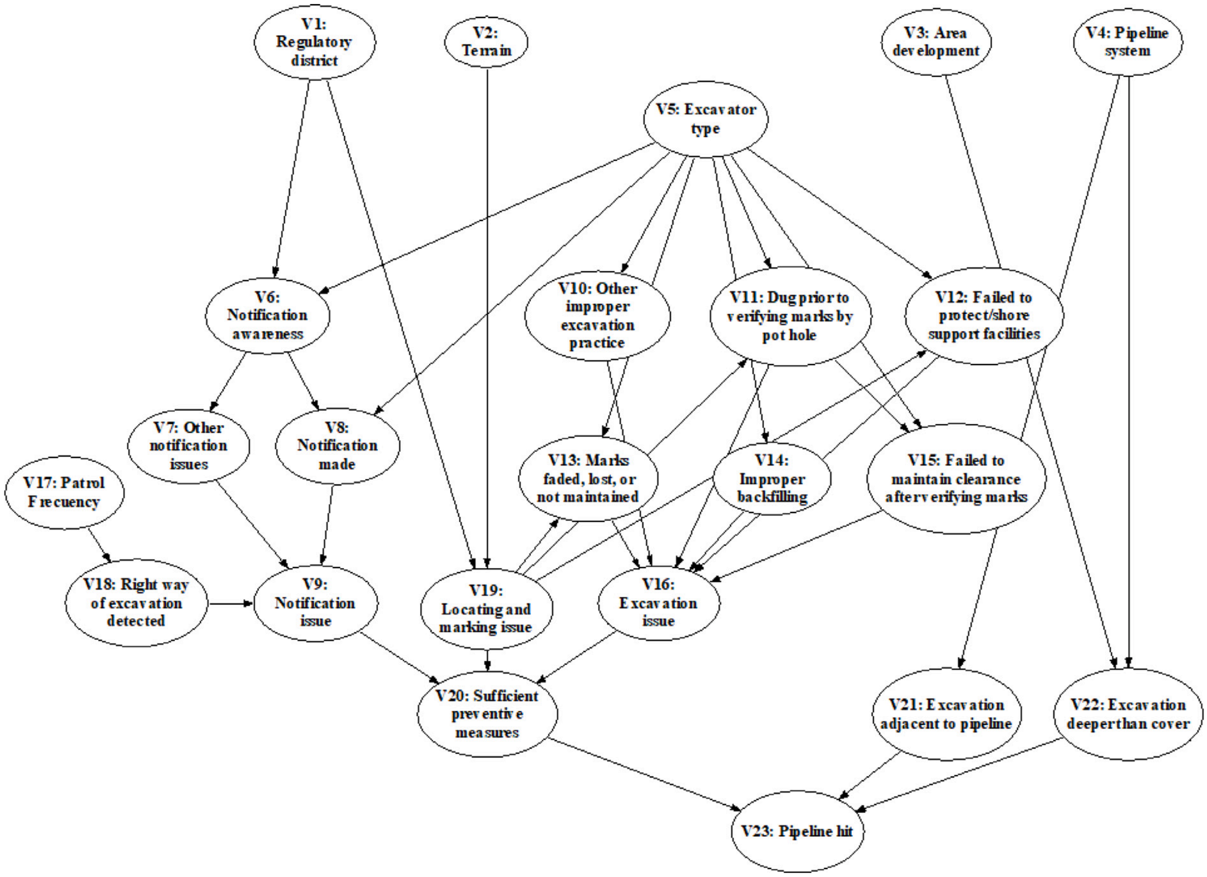

V1: Regulatory district
V2: Terrain
V3: Area development
V4: Pipeline system
V5: Excavator type
V6: Notification awareness
V7: Other notification issues
V8: Notification made
V9: Notification issue
V10: Other improper excavation practice
V11: Dug prior to verifying marks by pot hole
V12: Failed to protect/shore support facilities
V13: Marks faded, lost, or not maintained
V14: Improper backfilling
V15: Failed to maintain clearance after verifying marks
V16: Excavation issue
V17: Patrol Frecuency
V18: Right way of excavation detected
V19: Locating and marking issue
V20: Sufficient preventive measures
V21: Excavation adjacent to pipeline
V22: Excavation deeper than cover
V23: Pipeline hit

Fig. 5. Marginalized version of BaNTERA, a Bayesian network model for third-party excavation risk assessment. This model will be identified as M.
Source: Adapted from [50].

of notification, the event could have been avoided". However, the story-line presented above shows that there were multiple event precursors that could have influenced the outcome of the accident. As system safety practitioners, we need to learn from this event as thoroughly as possible so we can prevent future accidents. To expand OSHA's and PHMSA's investigations on the Sun Prairie gas explosion, we analyzed it through the "possible world" counterfactual reasoning methodological framework presented in Section 3.

4.1.1. Input: BaNTERA and accident narrative

Following the guidelines presented in Section 3.1, the following input tuple (see Eq. (3)) is selected for the "possible world" counterfactual reasoning case study on the Sun Prairie gas explosion:

- *Causal Bayesian network model (M)*: Since BaNTERA [50] represents a thorough risk assessment model for the third-party damage problem, a marginalized version of it is used for this case study. This model is presented in Fig. 5 and will be identified as M. A marginalized version of BaNTERA is selected to make the case study clear in this work. Originally, BaNTERA has 64 nodes and 101 edges, representing a level of complexity that could hinder the presentation of the case study results. On the other hand, the model in Fig. 5 was marginalized to only 23 nodes and 36 edges. Additionally, it must be highlighted that BaNTERA was marginalized in order to only model a pipe hit probability, not pipe damage. This was done to focus this study on the excavation process. We refer the reader to Appendix A for a list of the nodes in the model of Fig. 5, their states, and their prior marginal probabilities. We also refer the reader to BaNTERA's original paper [50] for more information on the model's construction, parameterization, and validation.
- *Event evidence*: The evidence on the Sun Prairie gas explosion is obtained from PHMSA's event report #20180073 in [51]. This evidence was mapped to the nodes of the Bayesian network model M, resulting in the node states presented in the "Prior" column of Table 1. An additional piece of evidence that is not explicitly shown in this column is that a pipe was hit in the excavation. In terms of the model M, this evidence is equivalent to "V_{23} = Yes".

A value of interest that can be obtained using the information presented above is the prior expected probability of a pipe hit by VC Tech given the conditions in which the excavation took place. This value can be calculated through Eq. (1) using the input model M of Fig. 5 and the event's evidence presented in Table 1. As a result, a prior pipe hit probability of $Pr(V_{23} = \text{Yes}) = 25.75\%$ is obtained; that is, it is expected to observe 25.75 pipe hits in 100 excavations in which the same event evidence is observed. This value will be used for comparison with the outcomes of the counterfactual scenarios analyzed in the next section.

4.1.2. Scenario analysis through counterfactual reasoning

The probabilistic "possible world" counterfactual reasoning methodology presented in Section 3.2 is applied to four distinct scenarios regarding the Sun Prairie gas explosion. These scenarios were formulated to thoroughly study the different event precursors presented in the trial story-line shown in Section 4.1. Scenarios A and C are used to study the effectiveness that a notification could have had on impacting the probability of a pipe hit by VC Tech. Additionally, the scenarios B and D are used to evaluate if there were any opportunities to decrease the likelihood of a pipe hit at the time of the event had the previous marks made by the USIC Locating Service, and used by VC Tech, were incorrect.

Table 1
Summary of the inputs and results of Sun Prairie gas explosion case study. The "-" symbol indicates that no information was instantiated in the Bayesian network model used in the study. For counterfactual scenarios, W, PW, and B indicates information that was instantiated in the "factual world," "possible world" and background nodes, respectively, of the scenario's twin network. $do(\cdot)$ indicates the use of the *do*-operator in the scenario's twin network "possible world" nodes.

Node (Fig. 5)	Prior	Scenario A	Scenario B	Scenario C	Scenario D
V_1: Regulatory district	East north central	B : East north central	B : East north central	B : East north central	B : East north central
V_2: Terrain	Under pavement	B : Under pavement	B : Under pavement	B : Under pavement	B : Under pavement
V_3: Area development	Class 3	B : Class 3	B : Class 3	B : Class 3	B : Class 3
V_4: Pipeline system	Distribution	B : Distribution	B : Distribution	B : Distribution	B : Distribution
V_5: Excavator type	Professional	B : Professional	B : Professional	B : Professional	B : Professional
V_6: Notification awareness	Aware	B : Aware	B : Aware	B : Aware	B : Aware
V_7: Other notification issues	–	–	–	–	–
V_8: Notification made	No	W: **No**; PW: do(**Yes**)	W : No	W: **No**; PW: do(**Yes**)	W : No
V_9: Notification issue	–	–	–	–	–
V_{10}: Other improper excavation practice	–	–	–	–	–
V_{11}: Dug prior to verifying marks by pot hole	–	–	–	–	W: -; PW: do(**No**)
V_{12}: Failed to protect/shore/support facilities	–	–	–	–	–
V_{13}: Marks faded, lost, or not maintained	–	–	–	–	–
V_{14}: Improper backfilling	–	–	–	–	–
V_{15}: Failed to maintain clearance after verifying marks	–	–	–	–	W: -; PW: do(**No**)
V_{16}: Excavation issue	–	–	–	–	–
V_{17}: Patrol frequency	–	–	–	–	–
V_{18}: Right way of excavation detected	No	B : No	B : No	B : No	B : No
V_{19}: Locating & marking issue	–	–	W: -; PW: do(**Yes**)	W: -; PW: do(**Yes**)	W: -; PW: do(**Yes**)
V_{20}: Sufficient preventive measures	–	–	–	–	–
V_{21}: Excavation adjacent to pipeline	Yes	B : Yes	B : Yes	B : Yes	B : Yes
V_{22}: Excavation deeper than cover	Yes	B : Yes	B : Yes	B : Yes	B : Yes
V_{23}: **Pipeline hit**	Pr(**No**) = 0.7425 Pr(**Yes**) = 0.2575	W: Pr(**Yes**) = 1; PW: Pr(**No**) = 0.8961 PW:Pr(**Yes**) = 0.1039	W: Pr(**Yes**) = 1; PW: Pr(**No**) = 0.3197 PW:Pr(**Yes**) = 0.6803	W: Pr(**Yes**) = 1; PW: Pr(**No**) = 0.8815 PW:Pr(**Yes**) = 0.1185	W: Pr(**Yes**) = 1; PW: Pr(**No**) = 0.3925 PW:Pr(**Yes**) = 0.6075

- *Scenario A.* First, we will analyze the lack of notification present in the excavation process, which was hypothesized by authorities as the root cause of the Sun Prairie gas explosion. In order to do this, the following query is studied through probabilistic "possible world" counterfactuals: *Given evidence* $\{V_i = v_i\}$ *gathered from the event, could the consequent state of "*V_{23} *: Pipeline hit" be "No", instead of the observed "Yes", had the antecedent state of "*V_8 *: Notification made" been "Yes", instead of the observed "No", in the event?* Here, and in the subsequent scenario analyses, $\{V_i = v_i\}$ corresponds to all evidence gathered from the event that is different from the antecedent and consequent of interest (V_8 and V_{23} respectively for this scenario). This evidence is presented in Table 1 under the "Scenario A" column. Following Step 1 of the methodology, this query can be expressed mathematically as:

$$Q_A = Pr\left(\left(V_{23}\right)_{V_8=\text{Yes}} = \text{No} \,\middle|\, V_8 = \text{No}, V_{23} = \text{Yes}, \{V_i = v_i\}\right) \quad (7)$$

  Then, following Step 2 of the methodology, the twin network presented in Fig. B.8 (in Appendix B) is created to compute the answer to the query in Eq. (7). Last, following Step 3 of the methodology, the answer to the counterfactual query is calculated, obtaining $Q_A = 89.61\%$. As such, a notification by VC Tech could have reduced the expected probability of a pipeline hit from 25.75% in the "factual world" to 10.39% in a "possible world" (see Table 1).
- *Scenario B.* The result obtained in Scenario A aligns well with the conclusions provided by the OSHA and PHMSA investigations; that is, if a notification had been provided by VC Tech, the incident could have been less likely to occur. VC Tech, however, argued that the actual root cause of the event was that the prior locating and marking works performed by USIC Locating Service were wrong. In order to assess this hypothesis, the following counterfactual query is studied: *Given evidence* $\{V_i = v_i\}$ *gathered from the event, how likely is that the consequent state of "*V_{23} *: Pipeline hit" been "No", instead of the observed "Yes", had the antecedent state of "*V_{19} *: Locating & marking issue" been "Yes" in the event?* Here, the "factual world" antecedent state on V_{19} is unknown; that is, there was no clear evidence supporting a locating and marking issue at the time of the event. Following Step 1 of the methodology, this query can be expressed mathematically as:

$$Q_B = Pr\left(\left(V_{23}\right)_{V_{19}=\text{Yes}} = \text{No} \,\middle|\, V_{23} = \text{Yes}, \{V_i = v_i\}\right) \quad (8)$$

  The twin network shown in Fig. B.9 is used to compute Eq. (8), obtaining $Q_B = 31.97\%$. Therefore, the "possible world" counterfactual suggests that if the marks used by VC Tech to perform their excavation activities had been incorrect, there was 68.03% probability of a pipeline hit at the time of the event (see Table 1).
- *Scenario C.* Even though the potential negative effect of locating and marking issues were considered in OSHA's investigation, the focus remained on a lack of locate request notification. To assess the effect of a notification on the probability of a pipe hit in the scenario in which there were incorrect locating and marking works at the excavation site, the following counterfactual query is studied: *Given evidence* $\{V_i = v_i\}$ *gathered from the event, how likely is that the consequent state of "*V_{23} *: Pipeline hit" been "No", instead of the observed "Yes", had the antecedent state of "*V_{19} *: Locating & marking issue" been "Yes" and the antecedent state of "*V_8 *: Notification made" been "Yes", instead of the observed*

"No", in the event? Mathematically, this query is equivalent to answering:

$$Q_C = Pr\left(\left(V_{23}\right)_{V_8=\text{Yes},V_{19}=\text{Yes}} = \text{No} \,\middle|\, V_8 = \text{No}, V_{23} = \text{Yes}, \{V_i = v_i\}\right) \tag{9}$$

The twin network shown in Fig. B.10 is used to compute Eq. (9), obtaining $Q_C = 11.85\%$. Therefore, a notification by VC Tech could have reduced the expected probability of a pipeline hit from a 25.75% in the "factual world" to a 11.85% in a "possible world" in which there was a locating and marking issue in USIC Locating Service works at the site (see Table 1).

- *Scenario D.* The result obtained for scenario B shows that a pipe hit could have been highly probable (68.03%) if the marks used by VC Tech would have been incorrect. Nevertheless, Scenario C showed that this probability could have still been significantly reduced (11.85%) had VC Tech provided a locate request notification. In public hearings, however, USIC Locating Service claimed that VC Tech did not follow dig-in best practices that could have avoided the event even in the scenario in which a locating and marking issue was present. This claim motivated us to study if there were any dig-in best practices that VC tech could have followed to effectively avoid a pipe hit in this scenario.

In order to perform the abovementioned study, Scenario B will be expanded to include the fourth step of the proposed counterfactual reasoning methodology; that is, explore other risk influencing factors through a sensitivity analysis. The same evidence used in Scenario B is instantiated on the twin network presented in Fig. B.9, and a sensitivity analysis is performed over the "possible world" outcomes of the node "V_{23} : Pipeline hit". In particular, this is done by calculating Kjærulff and Van Der Gaag [53] sensitivity derivative parameter, S, for all "possible world" dig-in best practices errors that depend on a marked excavation site (that is, V_{11}, V_{12}, V_{13}, and V_{15} in Table 1)[2] The results of the sensitivity analysis are shown in Fig. 6.

Fig. 6 shows that, if there had been a marking error at the time of the event, a pipe hit (i.e., "V_{23} = Yes") would have been most sensitive to increasing its probability due to the two following dig-in issues: "V_{11}: Dug prior to verifying marks by pot hole" and "V_{15}: Failed to maintain clearance after verifying marks". Potholing and clearance maintenance are expected to be performed consecutively in any excavation. As such, the following counterfactual query is studied: *Given evidence* $\{V_i = v_i\}$ *gathered from the event, how likely is that the consequent state of "V_{23} : Pipeline hit" been "No", instead of the observed "Yes", had the antecedent state of "V_{11}: Dug prior to verifying marks by pot hole" been "No", the antecedent state of "V_{15}: Failed to maintain clearance after verifying marks" been "No", and the antecedent state of "V_{19} : Locating & marking issue" been "Yes", in the event?* Mathematically, this query can be expressed as:

$$Q_D = Pr\left(\left(V_{23}\right)_{V_1 1=\text{No},V_{15}=\text{No},V_{15}=\text{Yes}} = \text{No} \,\middle|\, V_{23} = \text{Yes}, \{V_i = v_i\}\right) \tag{10}$$

The twin network shown in Fig. B.11 is used to compute Eq. (10), obtaining $Q_C = 39.25\%$. Therefore, if VC Tech had excavated using wrong marks, a pipe hit probability could have only been reduced from 68.03% (as shown in Scenario B) to 60.75% had they performed a pothole and a subsequent maintenance of clearance at the excavation site (see Table 1).

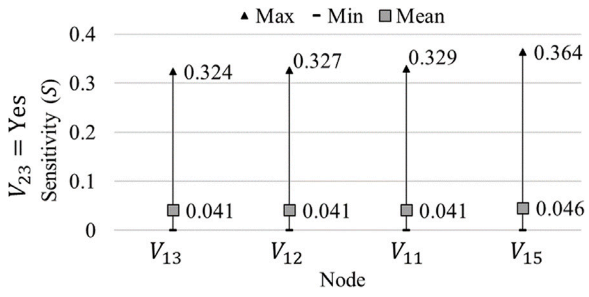


Fig. 6. Min, max, and mean sensitivity derivative parameter (S) of the nodes V_{11}, V_{12}, V_{13}, and V_{15} on the target node V_{23}.

4.1.3. Output: Lessons learned

Both OSHA's and PHMSA's investigations identified a lack of a locate request notification as the root cause of the event. This claim is supported by the results obtained for the scenarios A and C of this study. In fact, the "possible world" counterfactual analysis showed that a pipe hit probability could have been reduced from a 25.75% to a 10.39% had a notification been provided to authorities (Table 1, Scenario A). In addition, this study showed that a pipe hit probability could have increased only to 11.85% had locating and marking works by USIC Locating service on the site had been incorrect (Table 1, Scenario C).

The results obtained for the scenarios B and D, however, showed that if it was true that VC Tech used incorrect marks to perform the excavation activity, the probability of a pipe hit could have been of 68.03% (Table 1, Scenario B); 2.25 times more likely that the expected prior of 25.75%. Furthermore, this value could have been reduced only to 60.75% had VC Tech performed the dig-in best practices of potholing and subsequent site clearance, which were found to be the most influential on a pipe hit probability in this scenario (Table 1, Scenario D).

Therefore, the "possible world" counterfactual analysis performed on the Sun Prairie gas explosion showed that, in the particular case in which an excavator uses incorrect marks from a previous locating and marking work, only a locate request notification has the potential to significantly reduce a pipe hit probability. However, considering that a lack of notification is the major root cause of third-party damage in the U.S. [46], we conclude that it is necessary to find additional safety barriers beyond current dig-in best practices such as potholing and clearance maintenance to avoid similar events in the future.

5. Discussion

Counterfactual reasoning is prevalent in system safety in the form of "but-for" analyses due to the ability to incorporate the evidence from single events into an analyst's knowledge of a system in order to identify an event's causes and inform recommendations [1,5,54]. However, researchers have raised three important limitations on "but-for" counterfactuals use for learning from events: *linearity*, *incompleteness*, and *uncomparable*. In this work, we created the probabilistic "possible worlds" approach to counterfactuals to tackle these limitations while keeping the benefits of counterfactual reasoning in system safety.

The first limitation on the current use of counterfactuals in system safety is *linearity*, meaning that event investigators are prone to turn complex events into linear cause-and-effect chains. Probabilistic "possible worlds" counterfactuals address this limitation through the use of a Bayesian network-based QRA model to assess the likelihood of counterfactual hypotheses. Bayesian networks are built representing the current state of knowledge on the uncertainties about the phenomena, processes, and activities involved in a system [33]. As such, the use of a Bayesian network model enabled us, for instance, to incorporate into the Sun Prairie gas explosion case study different

[2] The sensitivity derivative parameter S of a node state "$V_i = v$" given a target node state "$Y_i = y_i$" means that a change c on the value of $Pr(V_i = v_i)$ will cause a change of $S \cdot c$ on the value of $Pr(Y_i = y_i)$. For more information on the method, we refer the reader to [53].

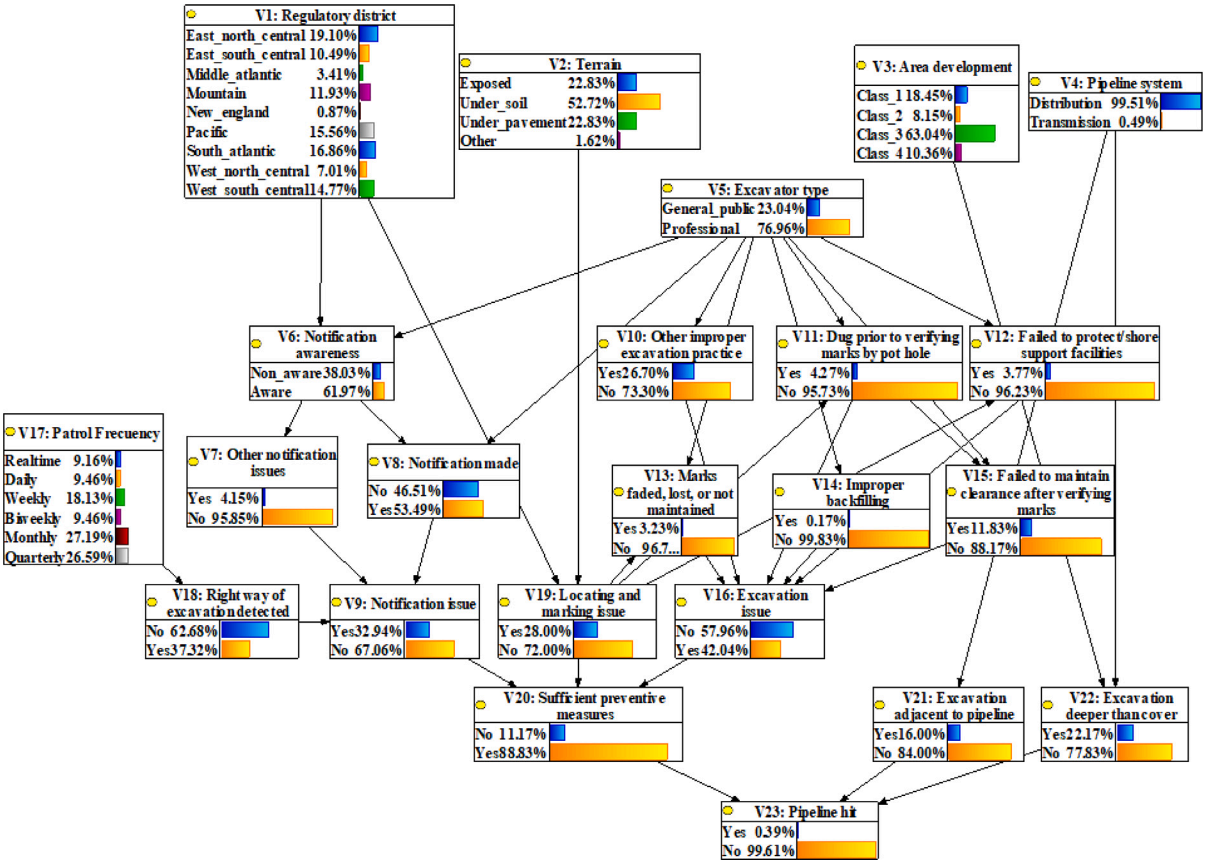


Fig. A.7. Nodes, states, and marginal prior probabilities of the Bayesian network model used in the Sun Prairie gas explosion case study.

relevant factors that are not currently considered in dig-in best practices procedures [49] such as notification awareness, excavator type, and a utility company's patrol frequency (see Appendix A). Furthermore, the use of the BaNTERA QRA model enabled us to better express the causal complexity of an excavation activity by incorporating commonly overlooked causal dependencies such as the effect that a type of terrain has on the likelihood of locating and marking issues ($V_2 \rightarrow V_{19}$ in Fig. 5) and the effect that a regulatory district has on an excavator's awareness on notification procedures and the quality of locating and marking works ($V_6 \leftarrow V_1 \rightarrow V_{19}$ in Fig. 5).

The second limitation on the current use of counterfactuals is *incompleteness*, namely, event investigators tend to use counterfactuals to transform an event's evidence into a proof of failure to perform according to a system's procedures and norms, therefore missing underlying reasons of an event. *Incompleteness* was clear in the official investigations following the Sun Prairie gas explosion. Both OSHA [52] and PHMSA [51] concluded, without further analysis, that the root cause of the event was a lack of locate request notification. Although there was additional evidence into potential location and excavation errors, the event was reduced into additional evidence on the potential high consequences that not following notification procedures can have on pipeline safety. Drawing a parallel to a "but-for" approach to counterfactual reasoning, the statement *"but-for the lack of notification, the Sun Prairie gas explosion could have been avoided"* was found by authorities as a logical, valid, and sufficient explanation to the event. As such, the investigation was halt, and the only lesson learnt was to keep promoting and enforcing the compliance to notification procedures (see Section 4). However, as we showed in this work, much more could have been learned from this event.

As discussed above, the use of Bayesian network models in probabilistic "possible worlds" counterfactuals allows the inclusion of relevant factors beyond a system's procedures and standards in the assessment of counterfactual hypotheses. This capability, however, does not ensure that these factors will be taken into consideration in the analysis. To overcome this issue, the proposed methodological framework for counterfactual reasoning (Fig. 2) includes a sensitivity analysis step that identifies which variables of a system (different from the previously queried "possible world" antecedents) could have increased the probability of an undesired consequence in a counterfactual scenario. This step is suggested to be performed after assessing an analyst's initial counterfactual hypotheses, thereby expanding the bounds of what can be learned from a past event. This capability was demonstrated in the Sun Prairie gas explosion case study by showing that, even though dig-in best practices had been performed, a pipe hit still could have been highly probable. Considering that there has been a call for action beyond current damage prevention practices in the U.S. [49], a "possible world" counterfactual analysis of the Sun Prairie gas explosion provided relevant evidence into exploring further actions beyond dig-in best practices to prevent damages when all previous safety barriers have failed.

The last limitation on the current use of counterfactuals in system safety is that they are *uncomparable*; that is, there is no rigorous method to compare and prioritize counterfactual hypotheses on event investigations. In order to tackle this issue, probabilistic "possible worlds" counterfactuals uses the QRA principles of uncertainty quantification and scenario analysis.

To incorporate uncertainty quantification in a counterfactual analysis of an event, we acknowledge a past event as one of many possible scenarios that could have unfolded. As such, we propose to assess the likelihood of a counterfactual hypothesis by transforming it into a probabilistic query to be answered through Bayesian network-based methods (see Fig. 2). Quantifying the uncertainty on the realization of a past event through probabilities allows different counterfactual hypotheses to be directly compared. This capability of "possible worlds" counterfactuals was demonstrated, for instance, in the case study when

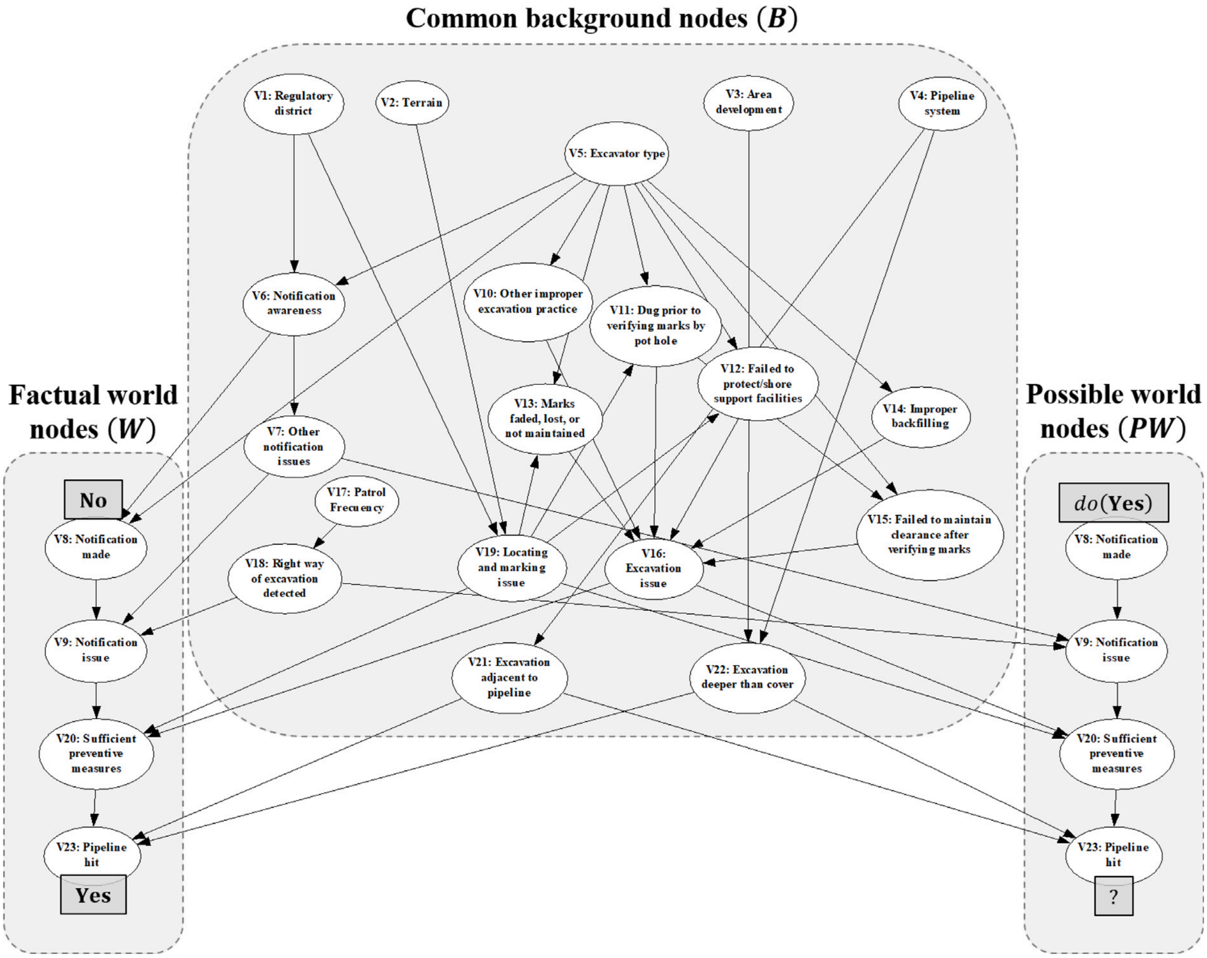

Common background nodes (B)
Factual world nodes (W)
Possible world nodes (PW)
V1: Regulatory district
V2: Terrain
V3: Area development
V4: Pipeline system
V5: Excavator type
V6: Notification awareness
V10: Other improper excavation practice
V11: Dug prior to verifying marks by pot hole
V12: Failed to protect/shore support facilities
V13: Marks faded, lost, or not maintained
V14: Improper backfilling
V7: Other notification issues
V17: Patrol Frecuency
V18: Right way of excavation detected
V19: Locating and marking issue
V16: Excavation issue
V15: Failed to maintain clearance after verifying marks
V21: Excavation adjacent to pipeline
V22: Excavation deeper than cover
No
V8: Notification made
V9: Notification issue
V20: Sufficient preventive measures
V23: Pipeline hit
Yes
do(Yes)
V8: Notification made
V9: Notification issue
V20: Sufficient preventive measures
V23: Pipeline hit
?

Fig. B.8. Twin network model used in the counterfactual scenario A of the Sun Prairie gas explosion case study. Although it is not shown in the twin network figure for clarity, some common background nodes have instantiated evidence as shown in Table 1.

assessing the likelihood of the counterfactual hypothesis “A locate request notification could have avoided a pipe hit”. In Table 1, we showed that the effect that a locate request notification could have had in a pipe hit probability could have been similar independently if a locate and marking error had been present at the excavation site (10.39% compared to 11.85%).

Although counterfactual hypotheses can be compared through probabilities, their calculation is not straightforward. Traditionally, a risk assessment quantifies the probability of a specific scenario; that is, assessing how a cause can potentially lead to a specific consequence. Counterfactuals, however, are based on the evidence gathered from a past event, namely, a scenario that have already unfolded. As such, it was necessary to establish the mathematical background and methods needed to incorporate an event’s evidence into the analysis of a counterfactual scenario in a system safety context. As demonstrated in this work’s case study, this was done through a Bayesian network-based twin network method, which allows a transparent representation of a counterfactual’s elements (namely, its factual and possible worlds antecedent, consequent, and background conditions), and the integration of evidence through probabilistic inference methods (see Section 2.3).

Despite the abovementioned benefits of “possible worlds” counterfactuals in system safety, it is important to take into account its limitations when informing “lessons learned” and recommendations. Insights that can be learned from a “possible worlds” counterfactual analysis are bounded by the Bayesian network-based QRA model used to describe the system and the event of interest. In particular, an accurate causal model of the event is necessary to provide useful information when learning from an event [55]. If a causal model is not accurate, it is possible to overlook potential causal mechanisms or background variables that can affect the correct assessment of the likelihood of a counterfactual hypothesis of interest [27]. Additionally, accurate causal modeling is also important due to the fact that counterfactual hypotheses are not possible to verify; an event already happened the way it did. Although natural experiments have been recommended to verify counterfactuals [1], we hold that the use of “possible world” counterfactuals should not be limited to the existence of these experiments. Rather, we recommend that the analyzed counterfactual hypotheses do not diverge from what is plausible to answer with the used QRA model [27,35].

Finally, the proposed methodology in Section 3 assumes that an event’s twin network counterfactual model is solely based on the variables included in the analyzed system’s original causal Bayesian network model. Although this assumption holds for well-understood systems, for new technologies, highly-complex systems, and events with sparse/conflicting evidence, it may be necessary to add new variables to model a “possible world” event. An example concerning sparse and conflicting evidence can be seen in legal argumentation, which can be a direct ramification of a high-consequence accident such as the Sun Prairie explosion. As described by Neil et al. [56], a model representing two competing narratives of an event (one as a “factual” and the other as a “possible” world) will have commonalities. However, variables such as factual evidence and source credibility can differ in both narratives.

Similar challenges can be seen when performing counterfactual analyses on new technologies. An example of this can be seen when analyzing the transportation of hydrogen blends in natural gas pipelines.

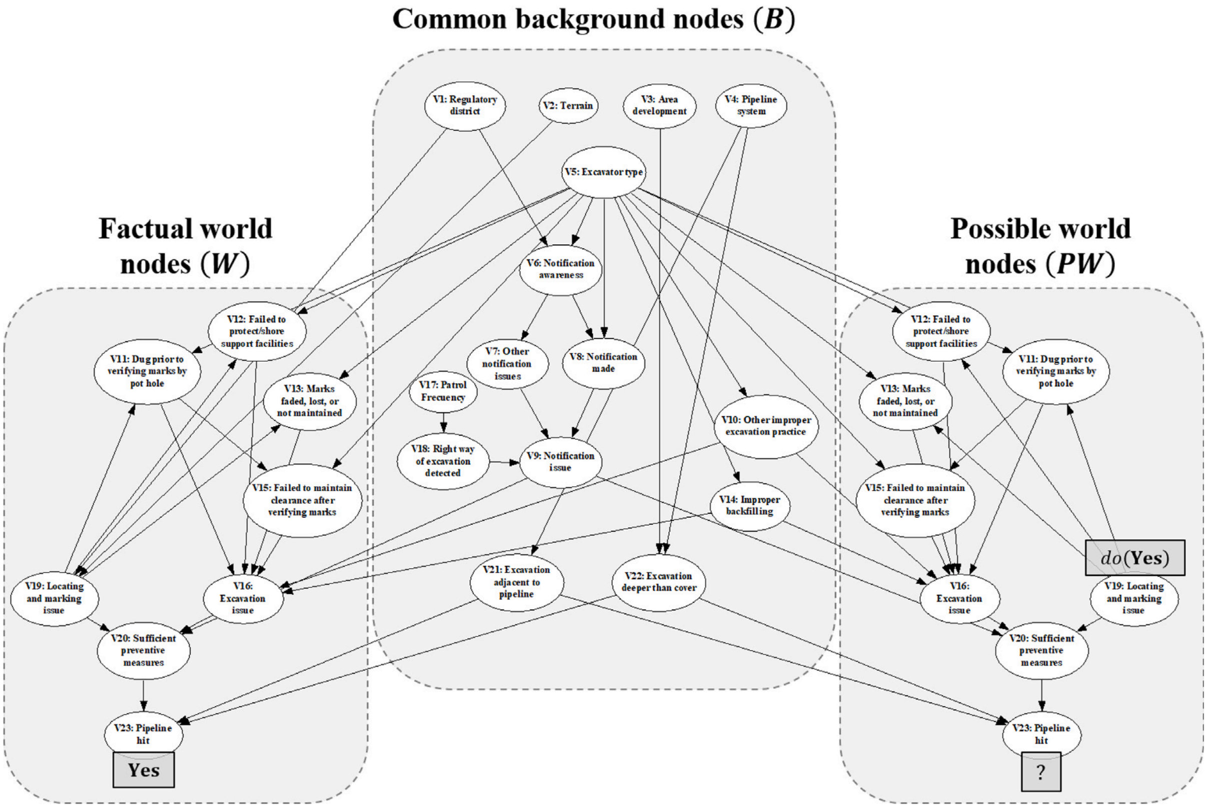

Common background nodes (B)
Factual world nodes (W)
Possible world nodes (PW)
V1: Regulatory district
V2: Terrain
V3: Area development
V4: Pipeline system
V5: Excavator type
V6: Notification awareness
V7: Other notification issues
V8: Notification made
V9: Notification issue
V10: Other improper excavation practice
V11: Dug prior to verifying marks by pot hole
V12: Failed to protect/shore support facilities
V13: Marks faded, lost, or not maintained
V14: Improper backfilling
V15: Failed to maintain clearance after verifying marks
V16: Excavation issue
V17: Patrol Frecuency
V18: Right way of excavation detected
V19: Locating and marking issue
V20: Sufficient preventive measures
V21: Excavation adjacent to pipeline
V22: Excavation deeper than cover
V23: Pipeline hit
Yes
$do(\textbf{Yes})$
?

Fig. B.9. Twin network model used in the counterfactual scenario B of the Sun Prairie gas explosion case study. Although it is not shown in the twin network figure for clarity, some common background nodes have instantiated evidence as shown in Table 1.

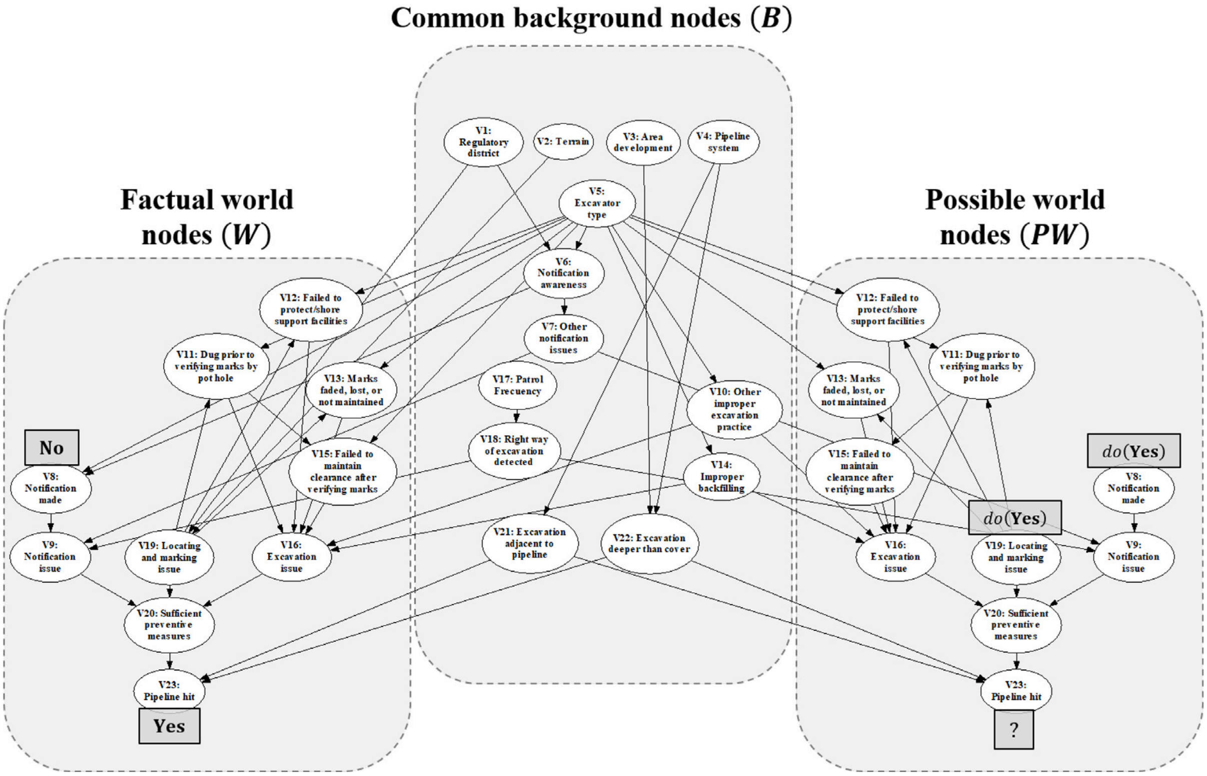

Common background nodes (B)
Factual world nodes (W)
Possible world nodes (PW)
V1: Regulatory district
V2: Terrain
V3: Area development
V4: Pipeline system
V5: Excavator type
V6: Notification awareness
V7: Other notification issues
V8: Notification made
V9: Notification issue
V10: Other improper excavation practice
V11: Dug prior to verifying marks by pot hole
V12: Failed to protect/shore support facilities
V13: Marks faded, lost, or not maintained
V14: Improper backfilling
V15: Failed to maintain clearance after verifying marks
V16: Excavation issue
V17: Patrol Frecuency
V18: Right way of excavation detected
V19: Locating and marking issue
V20: Sufficient preventive measures
V21: Excavation adjacent to pipeline
V22: Excavation deeper than cover
V23: Pipeline hit
No
Yes
$do(\textbf{Yes})$
$do(\textbf{Yes})$
?

Fig. B.10. Twin network model used in the counterfactual scenario C of the Sun Prairie gas explosion case study. Although it is not shown in the twin network figure for clarity, some common background nodes have instantiated evidence as shown in Table 1.

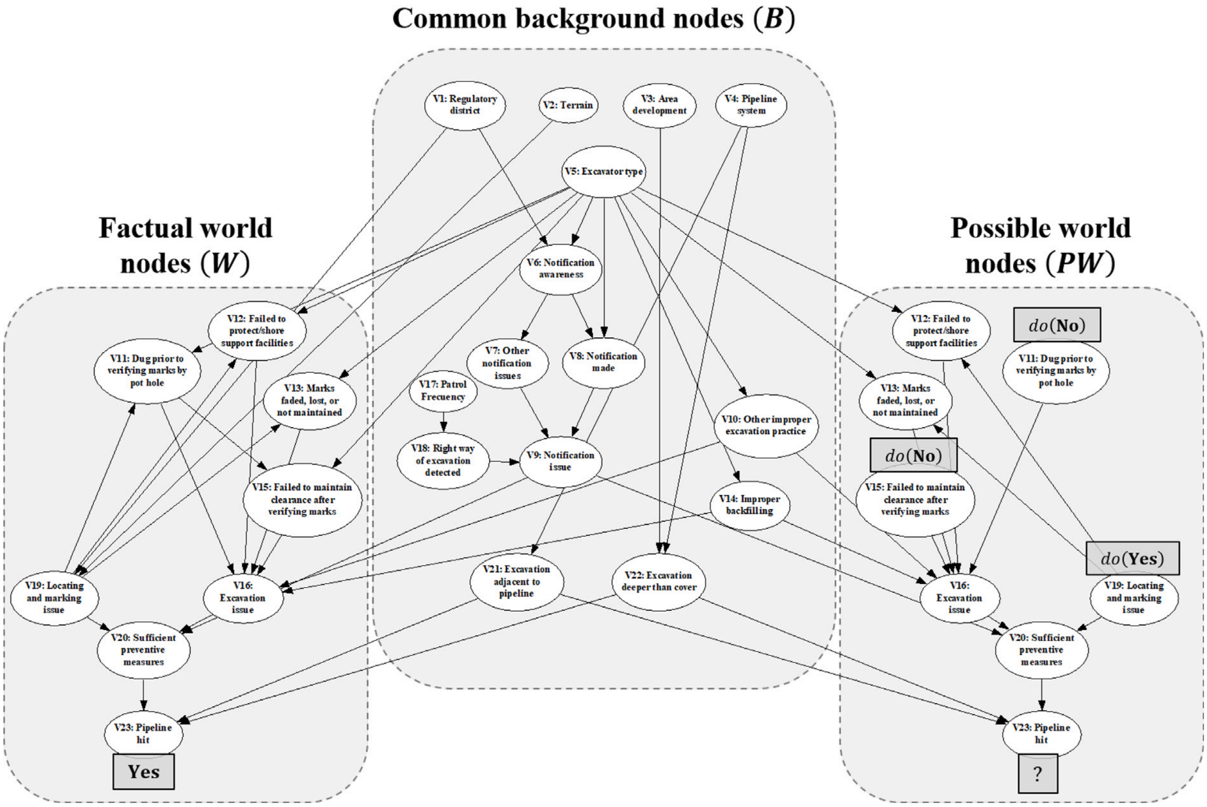

Common background nodes (B)
V1: Regulatory district
V2: Terrain
V3: Area development
V4: Pipeline system
V5: Excavator type
Factual world nodes (W)
Possible world nodes (PW)
V6: Notification awareness
V12: Failed to protect/shore support facilities
V11: Dug prior to verifying marks by pot hole
V13: Marks faded, lost, or not maintained
V7: Other notification issues
V8: Notification made
V17: Patrol Frecuency
V18: Right way of excavation detected
V9: Notification issue
V10: Other improper excavation practice
V15: Failed to maintain clearance after verifying marks
V14: Improper backfilling
V19: Locating and marking issue
V16: Excavation issue
V21: Excavation adjacent to pipeline
V22: Excavation deeper than cover
V20: Sufficient preventive measures
V23: Pipeline hit
Yes
do(No)
do(No)
do(Yes)
?

Fig. B.11. Twin network model used in the counterfactual scenario D of the Sun Prairie gas explosion case study. Although it is not shown in the twin network figure for clarity, some common background nodes have instantiated evidence as shown in Table 1.

Consider the following counterfactual query: *"Given the evidence on the Sun Prairie explosion, could the consequences have been different if the transported gas was a hydrogen blend, instead of natural gas?"* A twin network model of this query will require additional nodes in the "possible world" to incorporate the effect of hydrogen on natural gas pipelines, such as hydrogen embrittlement on steel pipes and potentially higher operational pressures. Addressing these challenges presents an excellent opportunity to expand the insights and lessons learned from an event's investigation in the face of complex systems, events, and new technologies. The proposed "possible worlds" approach can serve as the backbone of these analyses in the future.

6. Concluding remarks

This work established a "possible worlds" approach to counterfactual reasoning for system safety that improves a "*learning from events*" safety strategy. The proposed probabilistic "possible worlds" approach to counterfactuals enables the integration of an analyst's causal knowledge of a system (in the form of a Bayesian network-based risk assessment model) with the best available evidence on an event of interest. As a result, counterfactual hypotheses, which are of common use in the practice of system safety, can now be rigorously assessed through causally-sound probabilistic methods. This approach overcomes current "but-for" counterfactuals' limitations, as it was demonstrated in the case study on the 2018 Sun Prairie gas explosion.

Learning from single past events is crucial in the practice of system safety, and the new "possible worlds" approach provides a new foundation for enabling this. In summary, this work's major contributions are:

- The establishment of the "possible world" approach to counterfactual reasoning that enables, for the first time, a probabilistic assessment of counterfactuals in system safety.
- Demonstrating that "possible world" counterfactuals built on a QRA model and the best available evidence on an event provides an objective, defensible, and transparent basis for addressing counterfactual reasoning in system safety.
- Posit the mathematics of "possible worlds" counterfactuals to expand current capabilities of counterfactual methods in QRA and system safety for learning from past events (e.g., incident and accident investigations).
- The demonstration of the probabilistic "possible worlds" approach to counterfactual reasoning with the Sun Prairie explosion case study.

This work presents a first approach to the use of probabilistic counterfactuals in system safety, and future work is needed to keep developing the method and exploit its capabilities. A first direction for future research will be the development of a criteria that bounds which counterfactuals hypotheses can be accurately tested by the used risk assessment model. Future work will also explore generative modeling to account for both aleatory and epistemic uncertainties beyond the currently used conditional probability tables in the Bayesian network models shown in this work. Finally, future work will also focus on models that include not only the probabilities but also the consequences of an event. This direction will further expand the lessons that can be learned from events based on possible world counterfactuals. For instance, if the model presented in this work is expanded to include expected consequences in terms of loss of life and cost, "possible worlds" counterfactuals can provide insight into the probability that an accident such as the Sun Prairie explosion could have turned into a major catastrophe, and identify which variables are driving that scenario.

CRediT authorship contribution statement

Andres Ruiz-Tagle: Writing – original draft, Visualization, Validation, Software, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Enrique Lopez-Droguett:** Writing – review & editing, Methodology, Conceptualization. **Katrina M. Groth:** Writing – review & editing, Supervision, Resources, Project administration, Methodology, Funding acquisition, 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.

Data availability

Data will be made available on request.

Acknowledgments

This material is based upon work supported in part by the National Science Foundation, United States under Grant No. 2045519. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation. This work also received funding from the A. James Clark School of Engineering at the University of Maryland. The authors would also like to acknowledge the support from the Marvin Roush Fellowship. The technical support from Ernest Lever at GTI Energy is also gratefully acknowledged.

Appendix A. Nodes, states, and marginal prior probabilities of the Bayesian network model used in the Sun Prairie gas explosion case study

See Fig. A.7.

Appendix B. Twin network model used in the counterfactual scenarios

See Figs. B.8–B.11.

References

[1] Hopkins A. Issues in safety science. Saf Sci 2014;67:6–14. http://dx.doi.org/10.1016/j.ssci.2013.01.007.

[2] Stemn E, Joe-Asare T. The influence of accident manuals on the effectiveness of accident investigations–an analysis of accident management documents of Ghanaian mines. Saf Sci 2021;135:105129.

[3] Drupsteen L, Hasle P. Why do organizations not learn from incidents? Bottlenecks, causes and conditions for a failure to effectively learn. Accid Anal Prev 2014;72:351–8.

[4] Le Coze JC. What have we learned about learning from accidents? Post-disasters reflections. Saf Sci 2013;51(1):441–53.

[5] Rowland M, Hopkins A. Learning from high reliability organisations. In: The importance of counterfactual thinking: the case of the 2003 Canberra Bushfires. CCH Australia Ltd.; 2009, p. 237.

[6] Johnson C, Holloway C. A survey of logic formalisms to support mishap analysis. Reliab Eng Syst Saf 2003;80(3):271–91. http://dx.doi.org/10.1016/s0951-8320(03)00053-x.

[7] National Transportation Safety Board. Aircraft accident report: descent below visual glidepath and impact with seawall Asiana airlines flight 214. Tech. rep., U.S. National Transportation Safety Board; 2014.

[8] National Transportation Safety Board. Aircraft accident report: flight into terrain during missed approach U.S. air flight 1016. Tech. rep., U.S. National Transportation Safety Board; 1995.

[9] Branford K, Hopkins A, Naikar N. Guidelines for AcciMap analysis. In: Learning from high reliability organisations. CCH Australia Ltd; 2009, p. 1–12.

[10] Sanders J, Ladkin P. Introduction to why-because analysis. Tech. rep., Computer Networks and Distributed Systems Group, Bielefeld University; 2012.

[11] Johnson C. A handbook for the reporting of incidents and accidents. UK, London: Springer; 2003.

[12] Karanikas N, Chionis D, Plioutsias A. "Old" and "new" safety thinking: Perspectives of aviation safety investigators. Saf Sci 2020;125:104632.

[13] Dekker SW. Reconstructing human contributions to accidents: the new view on error and performance. J Saf Res 2002;33(3):371–85.

[14] Erik H. FRAM: the functional resonance analysis method: modelling complex socio-technical systems. CRC Press; 2017.

[15] Woo G, Maynard T, Seria J. Reimagining history: counterfactual risk analysis. Tech. rep., Lloyd's; 2017.

[16] Etienne J. Knowledge transfer in organisational reliability analysis: From post-accident studies to normal operations studies. Saf Sci 2008;46(10):1420–34.

[17] Hulme A, Stanton NA, Walker GH, Waterson P, Salmon PM. What do applications of systems thinking accident analysis methods tell us about accident causation? A systematic review of applications between 1990 and 2018. Saf Sci 2019;117:164–83.

[18] Morris MW, Moore PC. The lessons we (don't) learn: Counterfactual thinking and organizational accountability after a close call. Adm Sci Q 2000;45(4):737–65.

[19] Aven T. A risk science perspective on the discussion concerning safety I, safety II and safety III. Reliab Eng Syst Saf 2022;217:108077.

[20] Fu S, Yu Y, Chen J, Xi Y, Zhang M. A framework for quantitative analysis of the causation of grounding accidents in arctic shipping. Reliab Eng Syst Saf 2022;226:108706.

[21] Li X, Liu T, Liu Y. Cause analysis of unsafe behaviors in hazardous chemical accidents: Combined with HFACs and Bayesian network. Int J Environ Res Public Health 2020;17(1):11.

[22] Xia N, Zou PX, Liu X, Wang X, Zhu R. A hybrid BN-HFACS model for predicting safety performance in construction projects. Saf Sci 2018;101:332–43.

[23] Kabir S, Papadopoulos Y. Applications of Bayesian networks and Petri nets in safety, reliability, and risk assessments: A review. Saf Sci 2019;115:154–75.

[24] Langseth H, Portinale L. Bayesian networks in reliability. Reliab Eng Syst Saf 2007;92(1):92–108.

[25] Ruiz-Tagle A, Lopez Droguett E, Groth KM. Exploiting the capabilities of Bayesian networks for engineering risk assessment: Causal reasoning through interventions. Risk Anal 2021.

[26] Fenton N, Neil M. Risk assessment and decision analysis with bayesian networks. CRC Press; 2018.

[27] Pearl J. Causality. Cambridge University Press; 2009.

[28] The Daily Reporter Construction News. Tag archives: Sun prairie. Pages 1-2. Dly Rep 2019/2020. URL https://dailyreporter.com/tag/sun-prairie/.

[29] Lewis D. Counterfactuals and comparative possibility. In: Ifs. Springer; 1973, p. 57–85.

[30] Pearl J. Structural counterfactuals: A brief introduction. Cogn Sci 2013;37(6):977–85.

[31] Menzies P, Beebee H. Counterfactual theories of causation. In: Zalta EN, editor. The stanford encyclopedia of philosophy. Winter 2020, Metaphysics Research Lab, Stanford University; 2020.

[32] Modarres M, Kaminskiy MP, Krivtsov V. Reliability engineering and risk analysis: a practical guide. CRC Press; 2009.

[33] Aven T. Quantitative risk assessment: the scientific platform. Cambridge University Press; 2011.

[34] Balke A, Pearl J. Counterfactual probabilities: Computational methods, bounds and applications. In: Uncertainty in artificial intelligence conference proceedings. Elsevier; 1994, p. 46–54.

[35] Kyrimi E. Bayesian networks for clinical decision making: support, assurance, trust (Ph.D. thesis), Queen Mary University of London; 2019.

[36] Constantinou AC, Yet B, Fenton N, Neil M, Marsh W. Value of information analysis for interventional and counterfactual Bayesian networks in forensic medical sciences. Artif Intell Med 2016;66:41–52.

[37] Lam C, Cruz A. Risk analysis for consumer-level utility gas and liquefied petroleum gas incidents using probabilistic network modeling: a case study of gas incidents in Japan. Reliab Eng Syst Saf 2019;185:198–212.

[38] Hughes W, Zhang W, Cerrai D, Bagtzoglou A, Wanik D, Anagnostou E. A hybrid physics-based and data-driven model for power distribution system infrastructure hardening and outage simulation. Reliab Eng Syst Saf 2022;225:108628.

[39] Oughton EJ, Ralph D, Pant R, Leverett E, Copic J, Thacker S, et al. Stochastic counterfactual risk analysis for the vulnerability assessment of cyber-physical attacks on electricity distribution infrastructure networks. Risk Anal 2019;39(9):2012–31.

[40] Hund L, Schroeder B. A causal perspective on reliability assessment. Reliab Eng Syst Saf 2020;195:106678.

[41] GeNIe v3.0. 2019, URL https://bayesfusion.com/genie/.

[42] Ankan A, Panda A. Pgmpy: Probabilistic graphical models using python. In: Proceedings of the 14th python in science conference. 2015, p. 6–11.

[43] Woo G. Downward counterfactual search for extreme events. Front Earth Sci 2019;7:340.

[44] Wenzlhuemer R. Counterfactual thinking as a scientific method. Hist Soc Res/Histo Sozialforschung 2009;27–54.

[45] Salmon PM, Cornelissen M, Trotter MJ. Systems-based accident analysis methods: A comparison of accimap, HFACS, and STAMP. Saf Sci 2012;50(4):1158–70.

[46] PHMSA. Pipeline incident 20 year trends. U.S. Department of Transportation; 2020, https://www.phmsa.dot.gov/data-and-statistics/pipeline/pipeline-incident-20-year-trends [Accessed December 2020].

[47] CGA. DIRT annual report for 2019. Tech. rep., Common Ground Alliance; 2020.

[48] Santarelli JS. Risk analysis of natural gas distribution pipelines with respect to third party damage. Western University; 2019.

[49] CGA. The definitive guide for underground safety & damage prevention. Tech. rep., Alexandria, VA: common ground alliance; 2020, URL https://bestpractices.commongroundalliance.com/.

[50] Ruiz-Tagle A, Lewis AD, Schell C, Lever E, Groth KM. BaNTERA: A Bayesian network for third-party excavation risk assessment. Reliab Eng Syst Saf 2022;108507.

[51] PHMSA. Gas transmission & gathering incident data - january 2010 to present. In: Distribution, transmission & gathering, lng, and liquid accident and incident data. Tech. rep., U.S. Department of Transportation; 2021, URL https://www.phmsa.dot.gov/data-and-statistics/pipeline/distribution-transmission-gathering-lng-and-liquid-accident-and-incident-data.

[52] OSHA. U.S. department of labor cites utility contractors following fatal explosion in wisconsin. In: OSHA news release - region 5. 2019, URL https://www.osha.gov/news/newsreleases/region5/01102019.

[53] Kjærulff U, Van Der Gaag LC. Making sensitivity analysis computationally efficient. 2013, ArXiv:1301.3868.

[54] Le Coze J-C. Storytelling or theory building? Hopkins' sociology of safety. Saf Sci 2019;120:735–44.

[55] Smallman R, Summerville A. Counterfactual thought in reasoning and performance. Soc Personal Psychol Compass 2018;12(4):e12376.

[56] Neil M, Fenton N, Lagnado D, Gill RD. Modelling competing legal arguments using Bayesian model comparison and averaging. Artif Intell Law 2019;27(4):403–30.