

Date of publication xxxx 00, 0000, date of current version xxxx 00, 0000.

Digital Object Identifier 10.1109/ACCESS.2022.0092316

Causal Signal Temporal Logic for The Environmental Control and Life Support System's Fault Analysis and Explanation

ZIQUAN DENG¹, (Student Member, IEEE), SAMUEL P. ESHIMA², JAMES NABITY², and ZHAODAN KONG¹ (Member, IEEE)

¹Department of Mechanical and Aerospace Engineering, University of California, Davis, CA 95616, USA

²Department of Smead Aerospace Engineering Sciences, University of Colorado, Boulder, CO 80303, USA

Corresponding author: Zhaodan Kong (e-mail: zdkong@ucdavis.edu).

This work was supported by NASA's Space Technology Research Grants Program (grant number 80NSSC19K1052).

ABSTRACT Modern cyber-physical systems would often fall victim to unanticipated anomalies. Humans are still required in many operations to troubleshoot and respond to such anomalies, such those in future deep space habitats. To maximize the effectiveness and efficiency of the anomaly response process, the information provided by anomaly response technologies to their human operators must be epistemically accessible or explainable. This paper offers a first step towards developing explainable anomaly response systems. It proposes a logic, Causal Signal Temporal Logic (CaSTL), which can formally describe cause-effect relationships pertaining to fault explanation. Moreover, it develops an algorithm to infer a CaSTL formula that explains why a fault has happened in a system, given the model of the system and an observation about the fault. The effectiveness of the proposed algorithm is demonstrated with a simulated Environmental Control and Life Support System (ECLSS).

INDEX TERMS Explanation, diagnostic reasoning, failure analysis, simulation, temporal logic, troubleshooting.

I. INTRODUCTION

MODERN cyber-physical systems (CPSs) are becoming increasingly complex and being deployed in more safety-critical missions such as transportation, health care, manufacturing, aerospace, and defense [1], [2]. Since the concept of CPS was first proposed in [3], its intent has been constantly expanded by a series of works [4]–[6]. One notable characteristic is the increasing attention paid to the role of humans in a CPS. This shift in the role of humans, from being an operator of machines to a strategic decision-maker and a flexible problem solver, presents new design requirements for CPSs, particularly in terms of enhancing their ability to interact with humans [7]. Many technologies, such as an interactive human-machine interface [8], a human digital twin [9] and a virtual-physical collaboration controller [10] have been designed for the interaction between CPSs and humans in the robotics domain. However, in comparison, despite the widespread study of the fault detection and diagnosis (FDD) problem in CPSs, most works focus solely on the development of FDD algorithms [11]–

[13]. To date, limited works have proposed a comprehensive system-level CPS design that considers both physical components, simulation systems, and an interaction design specifically applied to FDD scenarios. Small errors in CPSs may lead to disastrous consequences. Even with the most recent advancements in FDD, human experts still (and in the foreseeable future continue to) play irreplaceable roles in anomaly response, particularly in safety-critical domains. To enable anomaly response technologies in CPSs to collaborate effectively and efficiently with humans, we must guarantee that the information provided by them to the humans about the anomaly (e.g., a system fault, an off-nominal behavior, or a cascading set of system disturbances) is not only accurate in and of itself but also *epistemically accessible* or *explainable* to their human operators. In this paper, we will call anomaly response technologies with adequate epistemological underpinnings as *fault explanation (FE)*.

To give a more concrete example, it has great significance to monitor the process effectively and deal with abnormal conditions promptly to ensure the success of long-duration,

deep space missions [14]. An Environmental Control and Life Support System (ECLSS), designed to meet the environmental and metabolic needs of the crew for such missions, is especially sensitive to failures since it directly relates to the crew's life and mission success [15]. Traditionally, the FDD method has been one of the core technologies to ensure the system's normal operation. However, humans, either onboard crew members or remote operators, have an irreplaceable advantage in the face of unpredictable anomalies. Thus, generating explanations for anomalies is indispensable in making anomaly response technologies more complete.

Related Work: Formal methods can serve as a starting point for developing FE technologies [16]. Existing formal verification tools can generate either a counterexample or a certificate to justify whether a CPS violates or satisfies a specification, which describes the desired/normal behavior [17]. Such certificates and counterexamples, particularly the latter, can already provide humans with some rudimentary understanding of the fault (which violates the specification) under investigation. However, the specification needs to be provided prior to the FE process, which may not be possible in many anomaly response scenarios. An alternative way is to learn the desired/normal or undesired/abnormal behavior of a CPS from data [18]–[21]. For instance, a (temporal) logical formula can be learned from the time series data of normal (positive) and abnormal (negative) behaviors to classify these two different types of behaviors. However, formulas learned in such a data-driven fashion can only explain “*What* has happened?” not “*Why* has the fault happened?”, a more practically important question in anomaly response. Moreover, it is difficult for such methods to generate explanations for token (singular) cases if they are unseen in the data, e.g., corner cases.

Causality theory [22]–[24] offers a wide range of tools to address these issues. It provides mathematical formulations to define and reason about cause-effect relationships, a crucial step towards answering “*Why*” questions. Moreover, based on studies on how humans explain decisions to one another [25], it has been shown that causal explanations are preferred by humans. Therefore FE systems should be able to provide causal explanations, if possible. There have been some recent efforts on integrating formal methods and causality, even thought not in the context of FE. For instance, a new temporal logic was proposed in [23] to characterize cause-effect relationships and causal inference was then used to learn logical formulas from data. The method is data-driven therefore doesn't utilize knowledge of the system under investigation. Moreover, the proposed logic is associational, which doesn't take the full advantage of causality theory (Pearl's causal hierarchy [22] consists of three layers from the least to most powerful: associational, interventional, and counterfactual). Ideally, a comprehensive explanation for a fault should include the description of the abnormal behavior, identification of the cause, and quantification of the causal relationship between the cause and the behavior [26]. However, existing fault explanation (FE) methods, such as re-

quirement learning methods [18]–[21] discussed above only capture the characteristics of the abnormal behavior, lacking causal implications. Despite the work in [23] attempts to express causal relationships by using temporal logic, the Probabilistic Causation [27] it chooses as the metric often fails to distinguish spurious correlations.

Considering the issues mentioned above, we propose a temporal logic as the formal explanation that can not only capture the properties of abnormal behavior and the corresponding potential causes but also provide a quantitative description to measure the causal effect of the potential cause for the abnormal behavior. Further, we propose an algorithm to infer this causal explanation, which serves as the crucial technology to solve the FE problem. The proposed explanation method sheds light on the way for the interaction design in CPSs, particularly under the human-involved FDD scenario.

Contributions: The main contribution of this paper is three-fold. First, we propose a new logic, Causal Signal Temporal Logic (CaSTL), which can be used to specify and reason about cause-effect relationships pertaining to FE. We endow CaSTL with interventional and counterfactual syntax and semantics. To the best of our knowledge, this is the first instance where a temporal logic has been proposed to incorporate the interventional and counterfactual aspects of causality. Second, we develop an algorithm that, given the simulation model of a system and the time series data of a fault, can infer a CaSTL formula explaining the fault from a causal perspective. Our algorithm is both model-based and data-driven. It generates an explanation that is consistent with both the observational data and the knowledge about the system. Finally, we validate the effectiveness of the proposed method via a simulated ECLSS developed by us.

II. PRELIMINARIES

The formal definition of CPS is firstly proposed in [3] as: “CPS is an integration of computation with physical processes whose behavior is defined by both cyber and physical components of the system. Embedded computers and networks monitor and control the physical processes, with feedback loops where physical processes affect computations and vice versa.” The cyber elements include embedded systems and network controllers, which are usually modeled as discrete events. Whereas the physical components exhibit continuous dynamics and are commonly modeled using differential equations [28]. For example, in our work, an ECLSS is considered a CPS composed of cyber components, e.g., some flight software supervision control and data acquisition systems, and physical components such as an array of sensors and actuators that interact with the internal or external environment.

In the context of our FE problem, upon observing an abnormal behavior in a physical component, such as an actuator in an ECLSS, humans are naturally inclined to engage in causal reasoning [25]. Often, we believe (or want to believe) that one action caused another. It is a truth universally echoed

by scientists that correlation does not imply causation. In daily life, however, the former is frequently mistaken for the latter. Causal inference aims at estimating the causal effects of an intervention or treatment on an outcome, which increasingly plays a vitally important role in scientific investigations and real-world applications [22]–[24]. Next, we will present a few key concepts in causality that are relevant to the rest of the paper. We use capital letters to denote variables, e.g., X , and small letters to denote their values, e.g., x . Similarly, we will use bold capital letters to denote sets of variables, e.g., \mathbf{X} , and bold small letters to denote sets of values, e.g., \mathbf{x} . If a set is indexed, e.g., with \mathbf{X} being a vector, we will use subscripts to denote its elements, e.g., \mathbf{X}_i being \mathbf{X} 's i -th element/dimension (and similarly \mathbf{x}_i being the value of \mathbf{X}_i).

In order to place generating an explanation for CPSs in the concept of causality, we first recall the original definition of the Structural Causal Model.

Definition 1. Structural Causal Model (SCM) [22]: An SCM is a triple $M = \langle \mathbf{U}, \mathbf{V}, \mathbf{F} \rangle$, where \mathbf{U} is a finite set of background or exogenous variables, which cannot be observed or experimented on, \mathbf{V} is a finite set of observable or endogenous variables (these variables are assumed to be functionally dependent on some subsets of $\mathbf{U} \cup \mathbf{V}$), and $\mathbf{F} = \{F_V | V \in \mathbf{V}\}$ is a finite set of functions such that each F_V is a mapping from a subset of $\mathbf{U} \cup \mathbf{V} \setminus V$ to V .

A causal model M induces a directed (causal) graph, $G(M)$, where each vertex corresponds to a variable in $\mathbf{U} \cup \mathbf{V}$ and each directed edge points from a variable in the domain of a function F_V (i.e., $Pa(V)$) to another variable in V . The (part of) physical component in a CPS can be transferred into an SCM where \mathbf{V} could represent a series of actuators' states at different time points, e.g., $X(t)$, \mathbf{F} represent any linear or non-linear functions to model dynamic relationships between the elements in \mathbf{V} . Generally speaking, in an SCM, the dynamics can be described using piece-wise functions [29] or differential equations [30]. In this paper, we assume the dynamics of actuators in our focused CPS are modeled by piece-wise functions and only consider the corresponding causal models that induce directed acyclic graphs (DAGs). Namely, for any element ($X(t)$ for dynamic case) in \mathbf{V} , it has no effect on itself, i.e., $X(t)$ does not appear in $F_{X(t)}$. It has been shown in [24] that for an SCM that induces a DAG, the values of all its endogenous variables are determined given a context \mathbf{u} , i.e., a setting \mathbf{u} for all the exogenous variables in \mathbf{U} ¹. We call a pair (M, \mathbf{u}) consisting of a causal model M and a context \mathbf{u} a (factual) causal setting (or a possible world). To make it easier to understand, we give an example of a static SCM without introducing time t .

Example 1. Fig. 1. (a) shows the causal graph $G(M)$ induced by an SCM M , where $\mathbf{U} = \{U_X, U_Y, U_Z\}$, $\mathbf{V} = \{X, Y, Z\}$, and $\mathbf{F} = \{F_X, F_Y, F_Z\}$, e.g., $X = F_X(Z, U_X) = 5Z + 3U_X$, $Y = F_Y(Z, U_Y) = 4Z + 7U_Y$.

¹The reverse is not true though. There might be multiple \mathbf{u} 's corresponding to the same set of endogenous variable values.

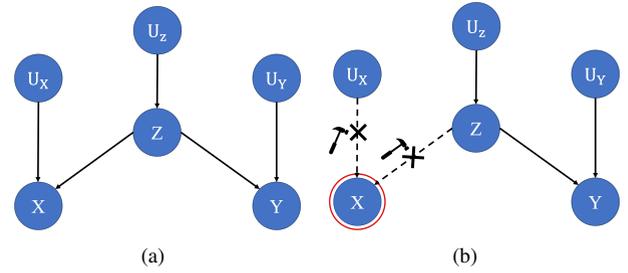


FIGURE 1. The causal graph $G(M)$ induced by the SCM of Example 1.

In the above example, we can observe X and Y are correlated since they're both caused by Z . However, there is no causal relationship between X and Y . Such structures widely exist in physical systems, e.g., two or more lights controlled by a single switch. However, we cannot determine the causal relationship between X and Y by only observing X and Y . Instead, we will need to conduct an intervention, which can be understood as a mathematical formulation of randomized experiments.

Definition 2. Intervention [22]: Forcing the value of some variable $V \in \mathbf{V}$ to v in an SCM M (called an “intervention” and denoted by $\text{do}(V = v)$) results in a new SCM denoted by $M_{\text{do}(V=v)}$. In the new SCM (called “intervention SCM” or “submodel”), the function for V , F_V , is set to v while the remaining functions in \mathbf{F} are unchanged. A pair $(M_{\text{do}(V=v)}, \mathbf{u})$ consisting of a submodel $M_{\text{do}(V=v)}$ and a context \mathbf{u} represents a counterfactual causal setting (or a parallel world). An intervention can be carried out on multiple endogenous variables simultaneously, denoted by $\text{do}(\mathbf{X} = \mathbf{x})$, where $\mathbf{X} \subseteq \mathbf{V}$. The resulting submodel is denoted by $M_{\text{do}(\mathbf{X}=\mathbf{x})}$.

Example 1. (Continued). An intervention $X = x$ means that we should replace the original function F_X which maps from U_X and Z to X with a new function $X = x$. In Fig. 1. (b), this corresponds to removing the edges from U_X and Z to X and setting the value of node X to x .

It can be easily seen that if all functions in \mathbf{F} of an SCM M or a submodel $M_{\text{do}(V=v)}$ are deterministic (an assumption we make in this paper), for a given context \mathbf{u} , there is only one possible value for each $v \in V$. We call the resulting set of endogenous variable values $\mathbf{v}(M, \mathbf{u})$ or $\mathbf{v}(M_{\text{do}(V=v)}, \mathbf{u})$ as potential outcome. With such, we can use any appropriate logic, e.g., the propositional or predicate logic, to reason about (factual and counterfactual) causal settings. For instance, given a formula $\varphi := y > 5$, a causal model M satisfies φ under context \mathbf{u} , written as $(M, \mathbf{u}) \models \varphi$, if $y(M, \mathbf{u}) > 5$. Moreover, Halpern and Peal have proposed a definition of actual cause with the aforementioned formalism.

Definition 3. HP Definition of Actual Cause [24]: $\mathbf{X} = \mathbf{x}$, $\mathbf{X} \subseteq \mathbf{V}$ is an actual cause of φ under the causal setting (M, \mathbf{u}) if all the following conditions hold: (1) **Sufficiency:** $(M, \mathbf{u}) \models \varphi$ and $(M, \mathbf{u}) \models \mathbf{X} = \mathbf{x}$; (2) **Necessity:** There

exists an intervention $\text{do}(\mathbf{X} = \mathbf{x}')$ with $\mathbf{x}' \neq \mathbf{x}$ such that $(M_{\text{do}(\mathbf{X}=\mathbf{x}')}(\mathbf{u}), \mathbf{u}) \not\models \varphi$.

III. CASTL: CAUSAL SIGNAL TEMPORAL LOGIC

Even though the HP definition of actual cause can be used to reason about causality in many real life scenarios [24], it is not powerful enough for the purpose of explaining and reasoning about faults in CPSs, e.g., ECLSS, where temporal information is crucial for both the description of faults and the analysis of their causes. Thus, in this section, we extend **Dfn. (3)** and propose the syntax as well as the qualitative and quantitative semantics of Causal Signal Temporal Logic (CaSTL).

Definition 4. Syntax of CaSTL: The syntax of CaSTL is defined as follows: $\Phi ::= \text{do}(\varphi_c) \rightsquigarrow \varphi_e$, where \rightsquigarrow is the “lead-to” operator and φ_c and φ_e are the cause and effect event formulas, respectively, which have the same syntax as the Signal Temporal Logic (STL) [31]: $\varphi_\rho ::= X_\rho(t_\rho) \sim d_\rho | \varphi_1 \wedge \varphi_2 | \varphi_1 \vee \varphi_2 | \diamond_{[\tau_1, \tau_2]} \varphi | \square_{[\tau_1, \tau_2]} \varphi$, where subscript $\rho = \{c, e\}$, $X_\rho(t_\rho)$ denotes the value of variable X_ρ at time t_ρ , $t_\rho \in \mathbb{N}$, $d_\rho \in \mathbb{R}$, and $\sim \in \{\leq, <, >, \geq\}$. The Boolean operators \vee and \wedge are disjunction (“or”) and conjunction (“and”), respectively. The temporal operators \diamond and \square stand for “eventually” and “always,” respectively. CaSTL is equipped with qualitative and quantitative semantics defined as follows.

Definition 5. Qualitative Semantics of CaSTL: A causal setting (M, \mathbf{u}) (or the resulting trace $\pi(M, \mathbf{u})$) satisfies Φ , i.e., $(M, \mathbf{u}) \models \Phi$ (or $\pi(M, \mathbf{u}) \models \Phi$) if and only if the following two conditions are satisfied: (1) **Sufficiency**: $\forall \text{do}([X_c(\tau_1) = x_c(\tau_1), \dots, X_c(\tau_n) = x_c(\tau_n)])$ such that $\pi_c = [x_c(\tau_1), \dots, x_c(\tau_n)] \models \varphi_c$, i.e., $\pi_c \models \diamond_{[\tau_1, \tau_n]}(X_c \geq d_c)$, $\pi(M_{\text{do}(\pi_c)}, \mathbf{u}) \models \varphi_e$, i.e., $\pi(M_{\text{do}(\pi_c)}, \mathbf{u}) \models \square_{[\tau_3, \tau_4]}(X_e \geq d_e)$. Here X_c is the variable related to the cause and X_e is related to the effect. (2) **Necessity**: $\exists \text{do}([X_c(\tau_1) = x'_c(\tau_1), \dots, X_c(\tau_n) = x'_c(\tau_n)])$ such that $\pi'_c = [x'_c(\tau_1), \dots, x'_c(\tau_n)] \not\models \varphi_c$, i.e., $\pi'_c \not\models \diamond_{[\tau_1, \tau_n]}(X_c \geq d_c)$, $\pi(M_{\text{do}(\pi'_c)}, \mathbf{u}) \not\models \varphi_e$, i.e., $\pi(M_{\text{do}(\pi'_c)}, \mathbf{u}) \not\models \square_{[\tau_3, \tau_4]}(X_e \geq d_e)$.

Remark 1. Intuitively, according to our semantics, we say that φ_c is the cause of φ_e in the context of an SCM M (e.g., one converted from a simulation model) and a context \mathbf{u} iff (1) any intervention satisfying φ_c will always result in a trace satisfying φ_e and (2) there exists an intervention violating φ_c which will result in a trace violating φ_e . We are extending the HP definition of token cause (**Dfn. (3)**) here, i.e., “if the cause φ_c had not occurred, then the effect φ_e would not have happened” (also called the counterfactual principle). Different from the original HP definition, which doesn't naturally afford reasoning about temporal relationships, our CaSTL definition of token cause (by combining the powers of causality and temporal logic) enables us to reason about and explain both causal and temporal relationships. Moreover, φ_c of CaSTL is more expressive than the HP causal expres-

sion, either a single variable or a set of variables.

The qualitative semantics can be used to check whether a causal setting (M, \mathbf{u}) satisfies or violates a proposed causal relation expressed in CaSTL. However, it does not provide any information about how strongly the **Sufficiency** and **Necessity** are satisfied or violated. Quantitative semantics for STLs, called robustness degrees, has been proposed to provide a measure of satisfiability of a trace with respect to (w.r.t.) a STL formula [32], [33]. We define the qualitative semantics of CaSTL by modifying these existing definitions as follows:

Definition 6. Quantitative Semantics of CaSTL: The degrees of sufficiency and necessity of a causal setting (M, \mathbf{u}) (or the resulting trace $\pi(M, \mathbf{u})$) w.r.t. Φ are defined as:

(1) The **degree of sufficiency** of a CaSTL formula Φ w.r.t. a model M and a trace π is

$$S(\Phi) = \mathbb{E}[\rho_e(\varphi_e, \pi(M_{\text{do}(\pi_c)}, \mathbf{u})) | \forall \rho_c(\varphi_c, \pi_c) > 0], \quad (1)$$

where $\rho_c(\cdot)$ are the robustness degrees of the interventional traces π_c w.r.t. the cause formula φ_c (see **Dfn. 5**), which is calculated based on [33]. $\rho_e(\cdot)$ are the robustness degrees of the traces $\pi(M_{\text{do}(\pi_c)}, \mathbf{u})$ w.r.t. the effect formula φ_e . Both of these two robustness degrees are bounded to the interval $[-1, 1]$. Therefore, $\rho_c(\cdot) > 0$ are corresponding to traces generated by applying interventions satisfying φ_c .

(2) The **degree of necessity** of a CaSTL formula Φ w.r.t. a model M and a trace π is defined as

$$N(\Phi) = -\mathbb{E}[\rho_e(\varphi_e, \pi(M_{\text{do}(\pi'_c)}, \mathbf{u})) | \forall \rho'_c(\varphi_c, \pi'_c) < 0], \quad (2)$$

where $\rho'_c(\cdot)$ are the robustness degrees of the interventional traces π'_c that violate the cause formula φ_c (see **Dfn. 5**). $\rho_e(\cdot)$ are the robustness degrees of the traces $\pi(M_{\text{do}(\pi_c)}, \mathbf{u})$ w.r.t. the effect formula φ_e . Similarly, $\rho'_c(\cdot) < 0$ are corresponding to traces generated by applying interventions violating φ_c .

IV. PROBLEM FORMULATION: FAULT EXPLANATION FROM A CAUSAL PERSPECTIVE

Based on the fact that causal inference is not possible without manipulating the variable(s) to which a cause will be attributed to, we will utilize a simulation model M of the system of interest, such as an ECLSS, to provide causal explanations for anomalies.

Problem 1. Fault Explanation (FE) Problem. Given a fully observable simulation model M of the system of interest, e.g., an ECLSS, the time series data (trace) of a fault $\pi = [\mathbf{x}(0), \dots, \mathbf{x}(T)]$, the time series data (trace) of a normal behavior $\pi^* = [\mathbf{x}^*(0), \dots, \mathbf{x}^*(T)]$, and a formula φ_e describing the effect of the fault, find another formula φ_c (i.e., find the variable X_c , the time bond $[t_{c_1}, t_{c_2}]$, and the threshold d_c) such that the CaSTL formula $\Phi = \text{do}(\varphi_c) \rightsquigarrow \varphi_e$ explains fault φ_e in the context of model M and trace π in the sense that: (1) **Existence**: there should be at least one context \mathbf{u} such that $\pi = \pi(M, \mathbf{u})$, i.e., the trace generated by M and \mathbf{u} satisfies $\pi \models \varphi_c$ and $\pi \models \varphi_e$. Moreover, $\pi^* \not\models \varphi_c$ and

$\pi^* \not\models \varphi_e$. (2) **Sufficiency and Necessity:** For all contexts \mathbf{u} 's satisfying the **Existence** condition, $(M, \mathbf{u}) \models \Phi$ according to the token level semantics of CaSTL, i.e., **Dfn.** (5).

Remark 2. The assumption that φ_e is given is not a very strong one, given all the existing temporal logic inference algorithms [18]–[21] that can be used to infer φ_e from data. We merely make the assumption to keep the paper compact. Besides, in practice, φ_e can be set according to the alarm threshold of the abnormal behavior.

Remark 3. We believe explanations written as CaSTL formulas are epistemically accessible to human operators. Such explanations satisfy a range of criteria of what humans see as “good” explanations [25]. For instance, CaSTL explanations are contrastive. Humans tend to think in counterfactual cases, e.g., “What would have happened if variable X had been different?” CaSTL naturally affords this aspect in its semantics. Furthermore, CaSTL explanations are compact due to the expressiveness of CaSTL as mentioned in **Remark 1**. At last but not least, CaSTL explanations can be selected. Human usually do not expect explanations that cover the complete list of causes of an event. We are used to selecting one or two causes from a list of possible causes as “the” explanation [34]. As such, in our framework, we first compute the degrees of sufficiency and necessity of each possible cause and then only provide those with highest degrees to humans.

V. METHODOLOGY

In this section, we will first convert the FE problem into an optimization problem (**Sec. V. V-A**). Then we introduce a way of formulating simulation models as SCMs (**Sec. V. V-B**) and propose a method to generate traces for quantifying the degrees of sufficiency and necessity (**Sec. V. V-C**). Finally, we present an overall algorithm to solve the FE problem (**Sec. V. V-D**).

A. FE AS AN OPTIMIZATION PROBLEM

In practice, it is quite possible that there doesn't exist an explanation Φ that satisfies all three conditions mentioned in **Problem 1**. To accommodate this, we will utilize the quantitative semantics of **Sufficiency** and **Necessity** defined in **Sec. III** and convert the FE problem to an optimization one. To identify the (possible) cause event $\varphi_{c,i}$ associated with an element/dimension \mathbf{X}_i of a component \mathbf{X} (with the corresponding explanation being $\Phi_i := \text{do}(\varphi_{c,i}) \rightsquigarrow \varphi_e$), we solve the following optimization problem:

Problem 2. FE Optimization Problem. Given a fully observable simulation model M of the system of interest, e.g., an ECLSS, the trace of a fault $\pi = [\mathbf{x}(0), \dots, \mathbf{x}(T)]$ the trace of a normal behavior $\pi^* = [\mathbf{x}^*(0), \dots, \mathbf{x}^*(T)]$, a formula φ_e describing the effect of the fault, the dimension of interest \mathbf{X}_i , find the optimal value of parameter vector $\theta_i^* := [t_{c_1}^*, t_{c_2}^*, d_{c_1}^*, \dots]$ (with the resulting CaSTL formula being $\Phi_i^* = \text{do}(\varphi_{c,i}^*) \rightsquigarrow \varphi_e$) that solves the following

optimization problem:

$$\sup J_i(\theta_i) = -E(\Phi_i) + \lambda_S S(\Phi_i) + \lambda_N N(\Phi_i) \quad (3)$$

where $E(\Phi_i) = e^{-(\rho_{c\pi} - \rho_{c\pi^*})}$ is the degree of existence, $\rho_{c\pi}$ and $\rho_{c\pi^*}$ are robustness degrees of π and π^* w.r.t. $\varphi_{c,i}$, respectively. $S(\Phi_i)$ and $N(\Phi_i)$ are **degrees of sufficiency** and **necessity** of the CaSTL formula Φ_i (as defined in **Eqns. 1** and **2**), λ_S and λ_N (both positive) balance the degrees of existence, sufficiency and necessity.

Finally, with all the Φ_i^* 's (θ_i^* 's) and their corresponding costs J_i^* , we can select the ones with the lowest costs (highest combined degrees of existence, sufficiency, and necessity) and provide them to human operators.

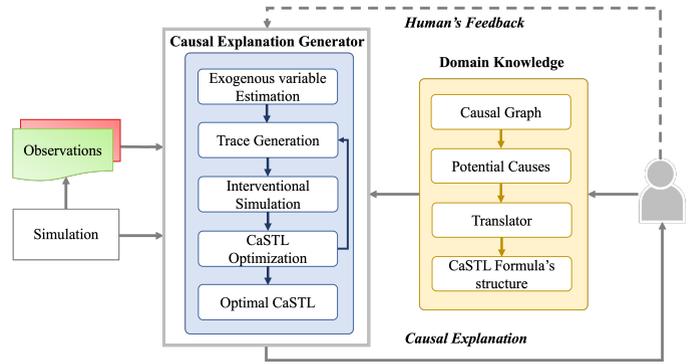


FIGURE 2. The framework of our proposed method. The dashed line is one of our future work. The proposed framework can not only provide causal explanations to human operators but also can be used to transmit human's feedback back Causal Explanation Generator to achieve a cooperative anomaly response.

B. SIMULATION MODELS AS SCMS

From **Example 1**, we are informed that the framework of the causal inference needs to manipulate the variable(s) to which a cause will be attributed. Such interventions are clearly unfeasible in many real-life's cases; however, we can manipulate variables with the help of simulation models to take the intervention. Because in simulation models, it is often possible to manipulate a variable individually, this naturally conforms to the concept of intervention and thus facilitates causal inference.

In order to take advantage of the flexibility of the simulation model to solve the FE optimization problem from the perspective of the causal inference, such a model can be converted to an SCM as follows. First, set \mathbf{U} as the sensor noises. Second, set \mathbf{V} as the collection of all components of the state variable at all time instants, i.e., $\mathbf{V} = \{\mathbf{X}(0), \dots, \mathbf{X}(T)\}$. Finally, set \mathbf{F} as $\{F_{\mathbf{X}(t)} | t = 0, \dots, T\}$ with each $F_{\mathbf{X}(t)}$ is the underlying relations among components.

C. DEGREES OF SUFFICIENCY AND NECESSITY COMPUTATION

Based on **Def. 6**, we need to generate traces satisfying the cause formula $\varphi_{c,i}$ to calculate the **degree of sufficiency**. As for the **degree of necessity**, traces violating $\varphi_{c,i}$ need to be

generated. We construct an optimization problem for trace generation. For preparing the calculation of the **degree of sufficiency**, we first solve the following problem:

$$\inf\left(-\sum_{n=1}^I \rho_{c_n} - \frac{1}{2} \sum_{n=1}^I \sum_{m=1}^I \|\rho_{c_n} - \rho_{c_m}\|^2\right), \quad (4)$$

$$s.t. \quad \pi_{c,i} \models \varphi_{c,i},$$

$$\delta\rho_{c_i} \geq \delta\epsilon_{\min},$$

where I is the number of traces to be generated, ρ_{c_i} is the robustness degree of each trace in $\pi_{c,i}$ w.r.t. $\varphi_{c,i}$, and the scalar $\epsilon_{\min} \geq 0$ is a desired minimum robustness. According to the cost function in Eqn. (4), $\forall \rho_{c_i} \in (0, 1]$, the generated traces satisfy $\varphi_{c,i}$ and the corresponding robustness degrees have a uniform distribution in the interval $(0, 1]$.

Similarly, for calculating the **degree of necessity**, we solve the following problem at first:

$$\inf\left(\sum_{i=1}^I \rho'_{c_i} - \frac{1}{2} \sum_{n=1}^I \sum_{m=1}^I \|\rho'_{c_n} - \rho'_{c_m}\|^2\right), \quad (5)$$

$$s.t. \quad \pi'_{c,i} \not\models \varphi_{c,i},$$

$$\delta\rho'_{c_i} \leq -\delta\epsilon_{\min},$$

where ρ'_{c_i} is the robustness degree of each trace in $\pi_{c,i}$ w.r.t. $\varphi_{c,i}$, which satisfy $\forall \rho'_{c_i} \in [-1, 0)$. Considering the smoothness of AGM robustness degree, we take a similar gradient-descent method proposed in [35] to solve above problems. After that, we take interventions on the simulation model M using $\pi_{c,i}$ and $\pi'_{c,i}$:

$$\pi(M_{\text{do}(\cdot)}, \mathbf{u}) = \pi(M_{\text{do}(\mathbf{X}_{c_i=\cdot})}, \mathbf{u}), \quad (6)$$

where \cdot is $\pi_{c,i}$ and $\pi'_{c,i}$, respectively, and \mathbf{u} is the specific sensor noises. Finally, the degrees of **sufficiency** and **necessity** can be computed using Eqns. (1) and (2). The algorithm to calculate the degrees of **sufficiency** and **necessity** is shown in **Alg. 1**.

D. SOLUTION

The conventional approaches for the optimization of cost functions J_i are often expensive, because the number of evaluations that may be performed is limited, typically to a few hundred or even less [36]. In our problem, we evaluate J_i using a simulator, e.g., a Simulink model. Given the complexity of many CPS models, obtaining a trace from the simulator can be time-inefficient and it is hard to observe the first- or second-order derivatives when evaluating J_i . Therefore, to learn a CaSTL formula in an efficient manner, we need to decrease the number of simulations and choose a more time-efficient optimization approach such as the Bayesian optimization (BayesOpt). Combined with our focus on finding a global rather than local optimum, we decide to use BayesOpt since it provides an elegant framework for avoiding getting caught in a local minimum.

BayesOpt consists of two main components: a Bayesian statistical model, typically Gaussian process (GP) regression, for modeling the cost function, and an acquisition

Algorithm 1 Degrees of **sufficiency** and **necessity** Computation for $\varphi_{c,i}$

Input: SCM M (converted from a simulation model), a causal STL formula $\varphi_{c,i}$ with parameter set θ_i , learning rates γ_1 and γ_2 , number of traces I , two thresholds ϵ_{d_1} and ϵ_{d_2}

Output: The **degree of sufficiency** $S(\Phi_i)$ and the **degree of necessity** $N(\Phi_i)$

- 1: Obtain \mathbf{u} from the simulation model
- 2: **while** $\min \|\rho_{c_n} - \rho_{c_m}\| \leq \epsilon_{d_1}$ **do**
- 3: Solve Eqn. (4) using the gradient-descent method with γ_1 as the learning rate
- 4: **end while**
- 5: Generate interventions using Eqn. (6)
- 6: $S(\Phi_i) \leftarrow$ Eqn. (1)
- 7: **while** $\min \|\rho'_{c_n} - \rho'_{c_m}\| \leq \epsilon_{d_2}$ **do**
- 8: Solve Eqn. (5) using the gradient-descent method with γ_2 as the learning rate
- 9: **end while**
- 10: Generate interventions using Eqn. (6)
- 11: $N(\Phi_i) \leftarrow$ Eqn. (2)
- 12: **return** $S(\Phi_i)$ and $N(\Phi_i)$

function for deciding where to sample next. Let $J(\cdot) \sim GP(m(\theta), k(\theta, \theta'))$ in Eqn. (3) be an unknown function we aim to optimize over a candidate set $\theta = [t_{c_1}, t_{c_2}, d_c]$. This GP is completely specified by its mean function $m(\theta)$ and its covariance function or kernel $k(\theta, \theta')$:

$$m(\theta) = \mathbb{E}[J(\theta)], \quad (7)$$

$$k(\theta, \theta') = \mathbb{E}[(J(\theta) - m(\theta))(J(\theta') - m(\theta'))]. \quad (8)$$

In this paper, the squared exponential kernel [37] is used in GP modelling, which can be computed as:

$$k(\theta_1, \theta_2) = \exp\left(-\frac{|\theta_1 - \theta_2|^2}{(2l)^2}\right), \quad (9)$$

where l is the length scale and $|\cdot|$ is the Euclidean length.

The statistical model as shown in Eqn. (7) is then used to create an acquisition function $\alpha_t(\theta)$. As such, it can be used to suggest P_t , the next input with which to sample the system. Gaussian Process Upper Confidence Bound (GP-UCB) [38] is one intuitive acquisition function. It balances exploration and exploitation through a single hyperparameter, β_t :

$$\alpha_t(\theta) = m_{t-1}(\theta) + \sqrt{\beta_t \sigma_{t-1}(\theta)}, \quad (10)$$

A higher β_t increases the variance of the acquisition function favor points, causing more exploration. A lower β_t increases the mean of the acquisition function favor points, causing more exploitation.

The algorithm to solve **Problem 2** using the BayesOpt is shown in **Alg. 2**. The steps are self-explanatory. We adopt the classical approach of counterfactual reasoning: abduction (Line 1) and intervention, consisting of action and prediction (Lines 7-9). The code will be available [here](#). **Fig. 2** shows the framework of our proposed methods to solve the **FE** problem.

Algorithm 2 FE Optimization for $\varphi_c(\theta)$

Input: Search space D , SCM M (converted from a simulation model), fault trace π , normal trace π^* , effect event φ_e

Output: A CaSTL formula that explains fault trace π in the context of model M : $\Phi_i^* = \text{do}(\mathbf{X}_i(t_c^*) \sim^* d_c^*) \rightsquigarrow \varphi_e$ and its cost J_i^*

- 1: Obtain \mathbf{u} from the simulation model
- 2: Select $\theta \in D$ for every different φ_c
- 3: Calculate J using **Eqn. (3)** and M
- 4: $P_0 \leftarrow [\theta, J]$
- 5: **for** t in $[1, T]$ **do**
- 6: Bayesian update $m_{t-1}(\theta)$ and $\sigma_{t-1}(\theta)$
- 7: $\theta_t \leftarrow \text{argmax}_{\theta \in D} m_{t-1}(\theta) + \sqrt{\beta_t} \sigma_{t-1}(\theta)$
- 8: Calculate S and N using **Alg. 1**
- 9: Calculate J_t using **Eqn. (3)**
- 10: $P_t \leftarrow P_{t-1} \cup (\theta_t, J_t)$
- 11: **end for**

VI. CASE STUDY

A. ECLSS SIMULATION MODEL

The Simulation Testbed for Exploration Vehicle ECLSS (STEVE) at the University of Colorado Boulder is a simplified single-bed CO_2 removal system of the Carbon Dioxide Removal Assembly (CDRA) onboard the International Space Station (ISS). Its purpose is to simulate common anomalies of CDRA and generate data under nominal and/or faulty conditions. STEVE, shown in **Fig. 3**, comprises a single sorbent bed packed with a 13X zeolite used in the 4-Bed CO_2 , a successor of the Carbon Dioxide Removal Assembly (CDRA). As indicated in the Piping and Instrumentation Diagram (**Fig. 4**), the bed either removes CO_2 from the provided air flow (adsorption) or releases CO_2 under thermal vacuum (desorption). During adsorption, the apparatus supplies the specified flow of CO_2 -laden air to the sorbent bed. Nominally, STEVE provides a gas mixture with 78.86% nitrogen, 20.84% oxygen, 0.3% carbon dioxide, and dew point of less than $-60^\circ C$; the latter achieved with a desiccant bed packed with Drierite® beads. At this concentration, the CO_2 partial pressure is approximately 2.1 mm Hg. A rope heater raises the insulated bed temperature to $200^\circ C$ and a throttled vacuum pump reduces pressure to below 20 mm Hg for CO_2 desorption and regeneration of the pellets via thermal-pressure swing. The adsorption/desorption cycle can be repeated for a set number of cycles.

In parallel, a Simulink model of the STEVE testbed was developed. The model can rapidly generate data and simulate conditions that the STEVE testbed cannot experimentally test. The model also utilizes Simscape, a physical pre-developed component model used in the Simulink environment. **Fig. 5** shows the overall model that comprises principal components and subsystems of the STEVE testbed, which are the inlet stream, the sorbent bed, valves, sensor suites, vacuum pump, flow controller, filter, and pipes.

The Simscape Moist Air components assume that gas

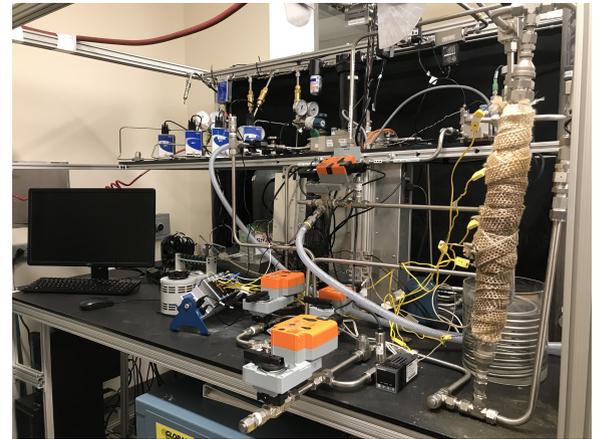


FIGURE 3. Image of the STEVE testbed.

species in the mixture are thermally perfect. Relationships between pressure, temperature, and density obey the ideal gas law. Other properties – specific enthalpy, specific heat, dynamic viscosity, and thermal conductivity – are functions of temperature only. The Simscape models conserve mass, momentum, and energy based on this assumption. In addition, Gaussian noise is added to the simulation results at the Data Acquisition (DAQ) subsystem to simulate sensor noise and saturation. Although these Simscape components enable fast and effortless simulation of basic thermal and fluid properties, there are no pre-developed models for valves and the sorbent bed which performs CO_2 adsorption and desorption. Such models are hard-coded using the “MATLAB Function block” available in the Simulink library.

The Simulink model is designed to simulate multiple failure modes similar to the STEVE testbed. For example, a leak at the outlet of the sorbent bed, which may be caused by wear and tear of connectors or the human error during maintenance, can be simulated. Air will leak out of the system during adsorption and into the system during desorption due to the pressure difference between the bed and the lab (habitat) environment. For this study, we performed a simulation with four cycles that contain 80 minutes CO_2 adsorption and 80 minutes desorption per cycle. (**Fig. 6** (b) and (c)).

B. SIMULATION RESULTS AND DISCUSSION

In order to validate the performance of proposed method for the **FE** problem, we consider the following situations: (1) there is a single fault happened to the component that we select as a potential cause and (2) there are multiple faults happened to different components that we select as potential causes.

1) Single Fault

In this case, a Leak failure is introduced to “Valve 1” in the third cycle as shown in **Fig. 6**. (a). The corresponding anomalies can be observed in a sequences sensors, e.g. *Bed*

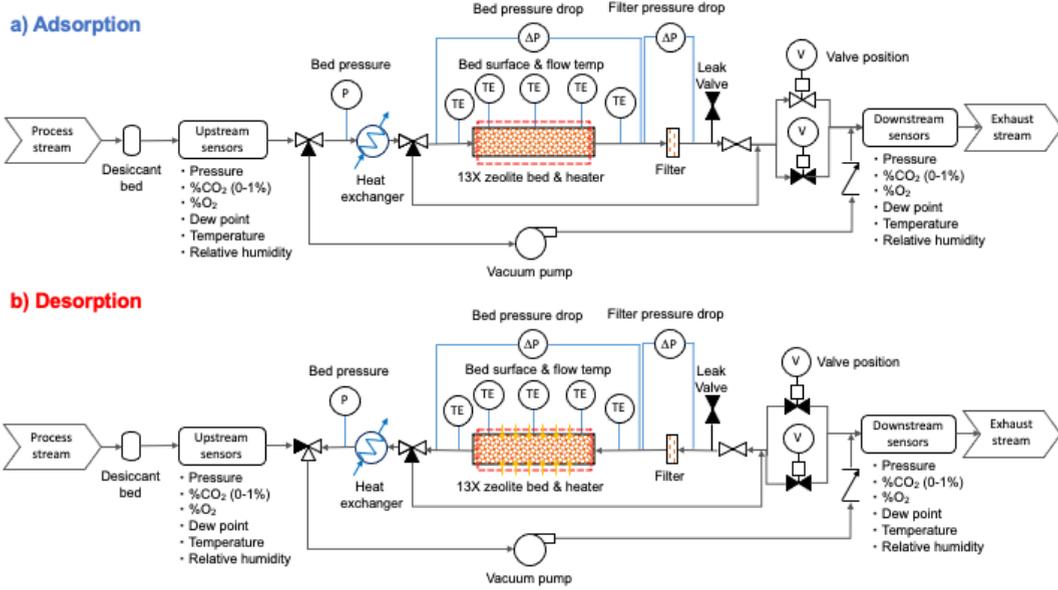


FIGURE 4. STEVE Piping and Instrumentation Diagram

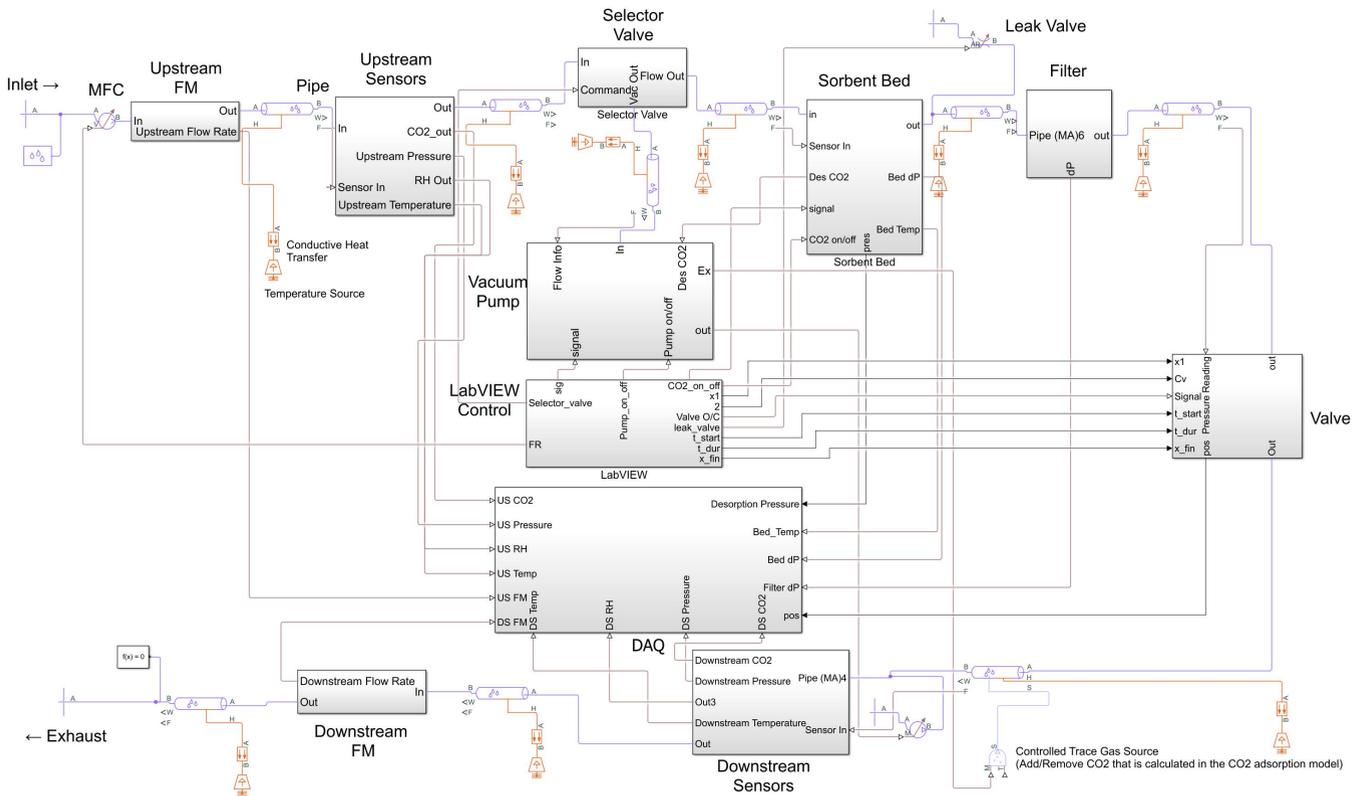


FIGURE 5. Overview of the Simulink Model.

TABLE 1. FE Results for Case 1.

Effect Formula	CaSTL Explanation Φ_i^*	$S(\Phi^*)$	$N(\Phi^*)$
$\varphi_e := \square_{[424,431]}(CO_2 \leq 87\%)$	$do(\square_{[310,428]}(Leak \geq 7.2 \times 10^{-8} \wedge Leak \leq 2.3 \times 10^{-7})) \rightsquigarrow \varphi_e$	0.54	0.68
$\varphi_e := \square_{[424,431]}(CO_2 \leq 71\%)$	$do(\square_{[308,425]}(Leak \geq 9.2 \times 10^{-8} \wedge Leak \leq 2.2 \times 10^{-7})) \rightsquigarrow \varphi_e$	0.63	0.72

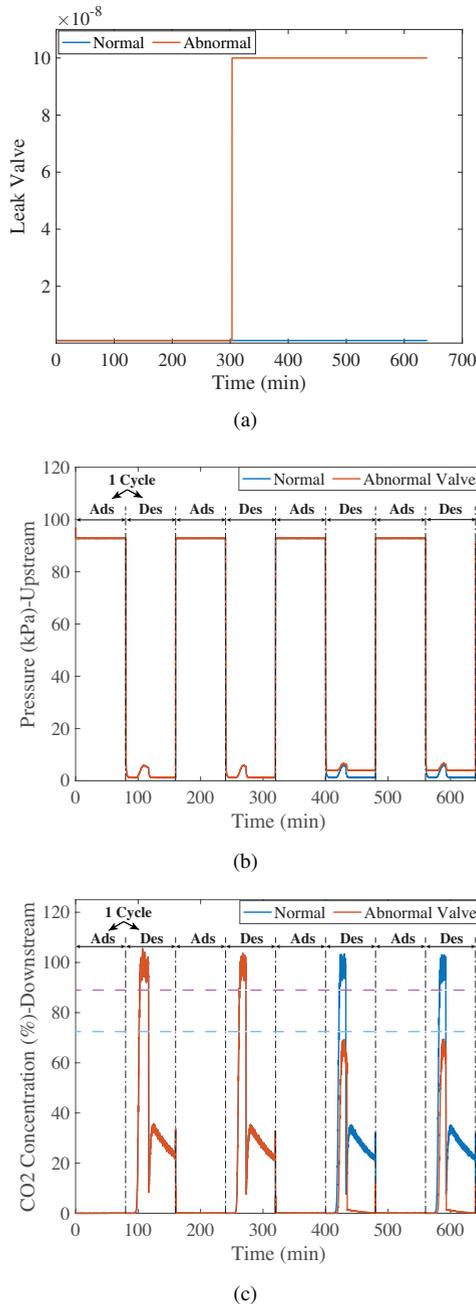


FIGURE 6. In (a)-(c), the blue (red) trajectory corresponds to a normal (abnormal) behavior of the system used in the case study. Please notice that in paper, we assume we have access to the abnormal behaviors as the fault trace π and the normal ones as π^* . Each cycle includes one CO_2 adsorption (ads) cycle and one desorption (des) cycle. (a) illustrates the “real” cause, an abrupt fault artificially injected; while the purple and blue dash lines in (c) illustrate two thresholds to define different effect formulas φ_e , respectively.

Temperature, CO_2 Concentration, etc., as shown in Fig. 6. (b), (c). In this case study, we made some simplifications and only analyzed the causal relationship between two components. Next, we follow our proposed framework as shown in Fig. 2 to infer a CaSTL to explain this fault qualitatively and quantitatively:

1. The developers of this ECLSS provide a causal graph of

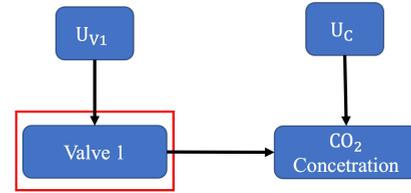


FIGURE 7. The causal graph pertaining to case 1, there is a leak happened to Valve 1. And this failure mode (red box) has an effect on CO_2 concentration.

this failure mode, as shown in Fig. 7.

2. Determine the abnormal behavior, the value of sensor CO_2 Concentration is below the threshold 87% (the purple dash line as shown in Fig. 6.(c)) at time bonds $[424min, 431min]$, as the effect for our problem. This can be written in the effect STL as $\varphi_e := \square_{[424,431]}(CO_2 \leq 87\%)$.
3. Determine the potential cause of this particular abnormal behavior could be a leak of “Valve 1”. This cause can be written in a cause STL as $\varphi_c := \square_{[t_1,t_2]}(Leak \geq \alpha \wedge Leak \leq \beta)$.
4. Given the abnormal and normal behaviors shown in Fig. 6. (a) and (c), we infer the optimal parameters in the CaSTL, $\Phi := do(\square_{[t_1,t_2]}(Leak \geq \alpha \wedge Leak \leq \beta)) \rightsquigarrow \square_{[424,431]}(CO_2 \leq 87\%)$, following steps listed in the blue box in Fig. 2 using Alg. 2.

Table 1 shows the formulas learned using our algorithm. It can be observed that $\Phi_1^* := do(\square_{[310,628]}(Leak \geq 8.2 \times 10^{-8} \wedge Leak \leq 2.3 \times 10^{-7})) \rightsquigarrow \varphi_e$ not only precisely locates the time bond $t \in [310min, 628min]$ but also approximately locates the leakage range of the valve, $Leak \in [8.2 \times 10^{-8}, 2.3 \times 10^{-7}]$. Compare this with the real cause shown in Fig. 6. (a), we can observe that the learned formula correctly identifies the temporal and spatial properties of the injected fault. Next, we change the threshold from 81% to 71% (the blue dash line as shown in Fig. 6.(c)) in the effect formula $\varphi_e := \square_{[424,431]}(CO_2 \leq 71\%)$ and then infer the parameters in the cause formula, $\varphi_c := \square_{[t_1,t_2]}(Leak \geq \alpha \wedge Leak \leq \beta)$. The corresponding results are shown in Table 1. According to the results, we can observe that for different effects, even for cause formulas have the same form, the parameters inferred in the cause formulas are different, and the corresponding **Sufficiency** and **Necessity** are also different. This is because the cause formula on the one hand should capture the difference between abnormal and normal behavior (as shown in Fig. 6), and provide the best explanation for different outcomes on the other hand, that is maximizing N and S .

2) Multiple Faults

In this case, we inject a fault mode, a leak, into Valve 1 and another fault, valve stiction, into Valve 2, which is the “Valve” model in Fig. 5. These two valves nominally open and close during adsorption and desorption, respectively. The corresponding causal graph is shown in Fig. 8. The corresponding anomalies can be observed in a sequences sensors

TABLE 2. FE Results for Case 2.

Effect Formula	CaSTL Explanation Φ_i^*	$S(\Phi^*)$	$N(\Phi^*)$
$\varphi_e := \square_{[424,431]}(CO_2 \leq 71\%)$	$do(\square_{[307,423]}(Leak \geq 9.4 \times 10^{-8} \wedge Leak \leq 2.4 \times 10^{-7})) \rightsquigarrow \varphi_e$	0.66	0.72
$\varphi_e := \square_{[424,431]}(CO_2 \leq 71\%)$	$do(\square_{[215.3,417.7]}(Degree \leq 17.2)) \rightsquigarrow \varphi_e$	0.02	0.03

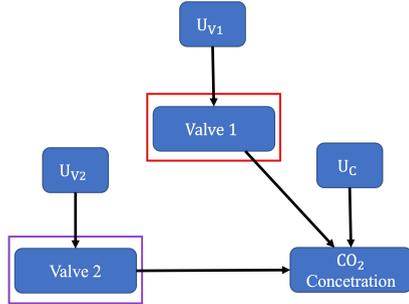
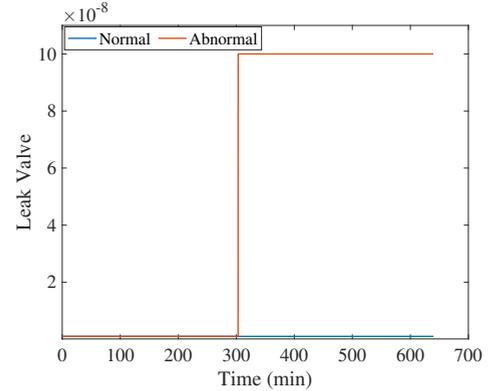


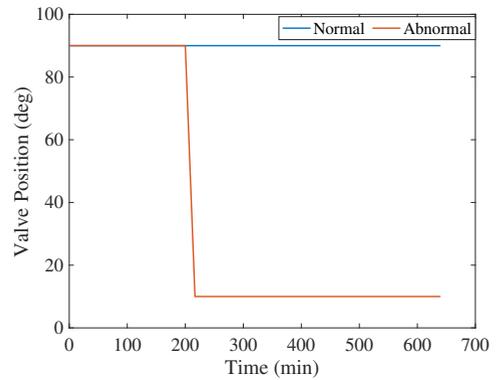
FIGURE 8. The causal graph pertaining to case 2, there is a leak (red box) happened to Valve 1. And there is another stiction (purple box) happened to Valve 2. Both of these two faults have effects on CO_2 concentration.

as shown in Fig. 9. (a)-(c). We define the abnormal behavior as $\varphi_e := \square_{[424,431]}(CO_2 \leq 71\%)$ for this case. And we set the causes are related to (1) the leak of Valve 1, written as $\varphi_{c_1} := \square_{[t_1,t_2]}(Leak \geq \alpha \wedge Leak \leq \beta)$; and (2) the stiction of Valve 2, written as $\varphi_{c_2} := \square_{[t_3,t_4]}(Degree \leq \gamma)$. Table 2 shows the learned formulas and the corresponding Sufficiency and Necessity. We can observe that both the Sufficiency and Necessity of cause related to Valve 2 are smaller than those related to Valve 1. This means Valve 1 has a stronger causal effect than Valve 2 for this specific abnormal behavior. From this, we are informed that different causes have different degrees of influence for the same effect.

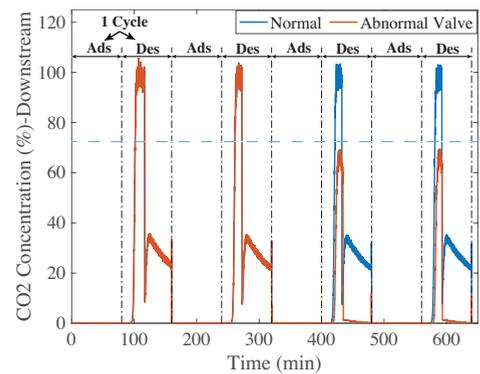
To further validate the advantages of our proposed method, we compare it with two fault diagnosis (explanation) methods. For the data-driven fault diagnosis method, we choose the approach proposed in [39], which is able to extract and classify temporal properties of different fault modes based on the k-means classification method. For the other method, we choose an STL inferring approach proposed in [21], which is designed for learning an STL to capture the temporal properties of a fault mode. This method can be used as a post-hoc explanation to a classification approach, e.g., here we use it to explain the classification results generated by the data-driven fault diagnosis method [39]. In the comparison, we construct a dataset which includes the value of the valve, the position of the valve and CO_2 concentration corresponding the fault mode 1: a leak failure, fault mode 2: a stiction of a valve and the normal state. For fault mode 1, we generate 50 traces by randomly injecting a leak greater 8×10^{-8} after 300 min. For fault mode 2, we generate 50 traces by randomly injecting a leak greater 8×10^{-7} after 200 min and commanding valve position to partially open to 5° after 200 min. Finally, we generate 50 traces for the normal condition. We randomly select 80% of the traces for training a cluster [39] to differentiate these three modes and then learning



(a)



(b)



(c)

FIGURE 9. The blue (red) trajectory corresponds to a normal (abnormal) behavior of the system used in the second case study. (a) illustrates a leak injected into Valve 1 and (b) illustrates a stiction injected into Valve 2. The blue dash lines in (c) illustrate the threshold to define effect formulas φ_e which describes the abnormal behavior related to CO_2 Concentration.

correspond STLs [21] to describe properties of each fault

TABLE 3. Results of two data-driven methods.

Fault Mode	Accuracy		Learned Formula
	Method 1 [39]	Method 2 [21]	
Leak Valve (Valve 1)	100%	100%	$\Box_{[308,535]}(Leak \geq 8.13 \times 10^{-8}) \wedge \Box_{[600,640]}(CO_2 \leq 52\%)$
Stiction Valve (Valve 2)	100%	100%	$\Box_{[212,576]}(Leak \geq 8.13 \times 10^{-8}) \wedge \Box_{[270.4,501]}(Degree \leq 70.4) \wedge \Box_{[452,489]}(CO_2 \leq 55\%)$

modes while the remaining signals are used for testing. The results are shown in Table 3

Here we would love to discuss the main difference between our work from other existing methods. Firstly, multiple temporal logic inference algorithms [18]–[21] focus on inferring STLs that can capture the proprieties to differentiate the abnormal behaviors from the normal ones. These STLs are somehow interpretable for human users; however, the learned proprieties, e.g., the learned formulas in Table 3, do not have any causal implications. Similarly, many components in an ECLSS statistically correlate with each other. Therefore, it is quite possible that a purely data-driven method may wrongly identify a cause of an effect. In addition, existing model-based fault diagnosis methods may be able to identify the source of the fault, e.g., a valve leak. However, they cannot quantify the relative strengths of the different explanations (and diagnosis results). On the other hand, our method can quantify them, as already shown in Table 1 and 2. These, we believe, demonstrate the power of our method.

VII. CONCLUSIONS

In order to explain to humans the anomalies occurring in CPSs so that FDD tasks can be performed with human advantages, this paper proposed a temporal logic called causal signal temporal logic. This logic has the ability to reason about the causation between a fault (cause) and the corresponding abnormal behaviors (effect). We also presented an algorithm for inferencing this logic based on a simulation model. The performance comparison with other formal interpretation methods on our developed ECLSS Simulink model verifies the advantage of our method in quantifying causality. In the future, we will keep developing methodologies that combine the strengths of formal methods and causality theory for explainable anomaly response for CPSs.

REFERENCES

- [1] R. V. Yohanandhan, R. M. Elavarasan, P. Manoharan, and L. Mihet-Popa, "Cyber-physical power system (cpps): A review on modeling, simulation, and analysis with cyber security applications," *IEEE Access*, vol. 8, pp. 151 019–151 064, 2020.
- [2] S. Pasandideh, P. Pereira, and L. Gomes, "Cyber-physical-social systems: Taxonomy, challenges, and opportunities," *IEEE Access*, vol. 10, pp. 42 404–42 419, 2022.
- [3] E. A. Lee, "Cyber physical systems: Design challenges," in *2008 11th IEEE international symposium on object and component-oriented real-time distributed computing (ISORC)*. IEEE, 2008, pp. 363–369.
- [4] L. Hu, N. Xie, Z. Kuang, and K. Zhao, "Review of cyber-physical system architecture," in *2012 IEEE 15th international symposium on object/component/service-oriented real-time distributed computing workshops*. IEEE, 2012, pp. 25–30.
- [5] J.-R. Jiang, "An improved cyber-physical systems architecture for industry 4.0 smart factories," *Advances in Mechanical Engineering*, vol. 10, no. 6, p. 1687814018784192, 2018.

- [6] A. Ahmadi, A. H. Sodhro, C. Cherifi, V. Cheutet, and Y. Ouzrout, "Evolution of 3c cyber-physical systems architecture for industry 4.0," in *Service Orientation in Holonic and Multi-Agent Manufacturing: Proceedings of SOHOMA 2018*. Springer, 2019, pp. 448–459.
- [7] M. Sampayo and P. Peças, "Cpsd2: A new approach for cyber-physical systems design and development," *Journal of Industrial Information Integration*, vol. 28, p. 100348, 2022.
- [8] M. A. R. Garcia, R. Rojas, L. Gualtieri, E. Rauch, and D. Matt, "A human-in-the-loop cyber-physical system for collaborative assembly in smart manufacturing," *Procedia CIRP*, vol. 81, pp. 600–605, 2019.
- [9] B. Wang, H. Zhou, G. Yang, X. Li, and H. Yang, "Human digital twin (hdt) driven human-cyber-physical systems: key technologies and applications," *Chinese Journal of Mechanical Engineering*, vol. 35, no. 1, p. 11, 2022.
- [10] H. Liu and L. Wang, "Remote human-robot collaboration: A cyber-physical system application for hazard manufacturing environment," *Journal of manufacturing systems*, vol. 54, pp. 24–34, 2020.
- [11] J. Boi-Ukeme, C. Ruiz-Martin, and G. Wainer, "Real-time fault detection and diagnosis of cps faults in devs," in *2020 IEEE 6th International Conference on Dependability in Sensor, Cloud and Big Data Systems and Application (DependSys)*. IEEE, 2020, pp. 57–64.
- [12] C. Alippi, S. Ntalampiras, and M. Roveri, "Model-free fault detection and isolation in large-scale cyber-physical systems," *IEEE Transactions on Emerging Topics in Computational Intelligence*, vol. 1, no. 1, pp. 61–71, 2016.
- [13] L. Piardi, P. Costa, A. Oliveira, and P. Leitão, "Collaborative fault detection and diagnosis architecture for industrial cyber-physical systems," in *2022 IEEE International Conference on Industrial Technology (ICT)*. IEEE, 2022, pp. 1–6.
- [14] S.-C. Wu and A. H. Vera, "Capability considerations for enhancing safety on long duration crewed missions: Insights from a technical interchange meeting on autonomous crew operations," *Journal of Space Safety Engineering*, vol. 7, no. 1, pp. 78–82, 2020.
- [15] S. P. Eshima, J. Nabity, and R. Moroshima, "Analysis of fault propagation of environmental control and life support system for self-awareness," in *ASCEND 2020*, 2020, p. 4012.
- [16] C. Baier, C. Dubsloff, F. Funke, S. Jantsch, R. Majumdar, J. Piribauer, and R. Ziemek, "From verification to causality-based explanations," *arXiv preprint arXiv:2105.09533*, 2021.
- [17] C. Baier and J.-P. Katoen, *Principles of model checking*. MIT press, 2008.
- [18] Z. Kong, A. Jones, and C. Belta, "Temporal logics for learning and detection of anomalous behavior," *IEEE Transactions on Automatic Control*, vol. 62, no. 3, pp. 1210–1222, 2016.
- [19] E. Bartocci, N. Manjunath, L. Mariani, C. Mateis, and D. Ničković, "Cpsdebug: Automatic failure explanation in cps models," *International Journal on Software Tools for Technology Transfer*, vol. 23, no. 5, pp. 783–796, 2021.
- [20] G. Bombara and C. Belta, "Offline and online learning of signal temporal logic formulae using decision trees," *ACM Transactions on Cyber-Physical Systems*, vol. 5, no. 3, pp. 1–23, 2021.
- [21] Z. Deng and Z. Kong, "Interpretable fault diagnosis for cyberphysical systems: A learning perspective," *Computer*, vol. 54, no. 9, pp. 30–38, 2021.
- [22] J. Pearl, *Causality*. Cambridge university press, 2009.
- [23] S. Kleinberg, "A logic for causal inference in time series with discrete and continuous variables," in *Twenty-Second International Joint Conference on Artificial Intelligence*, 2011.
- [24] J. Y. Halpern, *Actual causality*. MIT Press, 2016.
- [25] T. Miller, "Explanation in artificial intelligence: Insights from the social sciences," *Artificial intelligence*, vol. 267, pp. 1–38, 2019.
- [26] G. Manco, E. Ritacco, P. Rullo, L. Gallucci, W. Astill, D. Kimber, and M. Antonelli, "Fault detection and explanation through big data analysis on sensor streams," *Expert Systems with Applications*, vol. 87, pp. 141–156, 2017.
- [27] E. Eells, *Probabilistic causality*. Cambridge University Press, 1991, vol. 1.

[28] A. Rashid and O. Hasan, "Formal analysis of the continuous dynamics of cyber-physical systems using theorem proving," *Journal of Systems Architecture*, vol. 112, p. 101850, 2021.

[29] J. Runge, S. Bathiany, E. Bollt, G. Camps-Valls, D. Coumou, E. Deyle, C. Glymour, M. Kretschmer, M. D. Mahecha, J. Muñoz-Marí et al., "Inferring causation from time series in earth system sciences," *Nature communications*, vol. 10, no. 1, p. 2553, 2019.

[30] J. M. Mooij, D. Janzing, and B. Schölkopf, "From ordinary differential equations to structural causal models: the deterministic case," *arXiv preprint arXiv:1304.7920*, 2013.

[31] A. Donzé, "On signal temporal logic," in *International Conference on Runtime Verification*. Springer, 2013, pp. 382–383.

[32] A. Donzé and O. Maler, "Robust satisfaction of temporal logic over real-valued signals," in *International Conference on Formal Modeling and Analysis of Timed Systems*. Springer, 2010, pp. 92–106.

[33] N. Mehdipour, C.-I. Vasile, and C. Belta, "Arithmetic-geometric mean robustness for control from signal temporal logic specifications," in *2019 American Control Conference (ACC)*. IEEE, 2019, pp. 1690–1695.

[34] C. Molnar, *Interpretable machine learning*. Lulu.com, 2020.

[35] V. Raman, A. Donzé, M. Maasoumy, R. M. Murray, A. Sangiovanni-Vincentelli, and S. A. Seshia, "Model predictive control with signal temporal logic specifications," in *53rd IEEE Conference on Decision and Control*. IEEE, 2014, pp. 81–87.

[36] P. I. Frazier, "Bayesian optimization," in *Recent advances in optimization and modeling of contemporary problems*. Informs, 2018, pp. 255–278.

[37] M. Seeger, "Gaussian processes for machine learning," *International journal of neural systems*, vol. 14, no. 02, pp. 69–106, 2004.

[38] N. Srinivas, A. Krause, S. M. Kakade, and M. W. Seeger, "Information-theoretic regret bounds for gaussian process optimization in the bandit setting," *IEEE transactions on information theory*, vol. 58, no. 5, pp. 3250–3265, 2012.

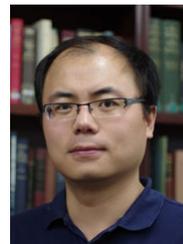
[39] L.-M. Wang and Y.-M. Shao, "Crack fault classification for planetary gearbox based on feature selection technique and k-means clustering method," *Chinese Journal of Mechanical Engineering*, vol. 31, no. 1, pp. 1–11, 2018.



JAMES NABITY received his Doctor's degree in mechanical engineering from the University of Colorado, Boulder, CO, USA, in 2007. He is currently an Associate Professor in the Smead Aerospace Engineering Sciences department with research focus in the field of Bioastronautics — the study and support of life in space. His current research advances environmental control and life support (ECLS) technologies with focus on ionic liquid membranes for atmosphere revitalization and CO2 capture, use of ionic liquid solvents to extract minerals and oxygen from regolith, explore the effects of space radiation on habitat layout and crew performance, develop bioregenerative systems, and investigate heat transport and fluid flow in microgravity. Prior to joining CU in 2013, he was a Principal Engineer for TDA Research and before that an engineer at the Naval Air Warfare Center (NAWC) advancing the development of propulsion systems, burners and combustors, and ECLS technologies for spacecraft and submarines. He is a NAWC Technical Fellow (1996) for contributions to combustion, an AIAA Associate Fellow (2016) and AIAA Life Sciences & Systems Technical Program Chair for the International Conference on Environmental Systems.



ZIQUAN DENG (Student Member, IEEE) received the bachelor's and master's degrees in mechanical engineering and aerospace engineering from the Harbin Institute of Technology, Harbin, China, in 2016 and 2018, respectively, the master's degree in mechanical engineering from the Washington University in Saint Louis, Saint Louis, MO, USA, in 2020. He is currently a Ph.D. student in the Department of Mechanical and Aerospace Engineering, University of California, Davis, California, USA. His research interests include formal methods and explainable Cyber-Physical Systems/Artificial intelligence.



ZHAODAN KONG (Member, IEEE) received the bachelor's and master's degrees in astronautics and mechanics from the Harbin Institute of Technology, Harbin, China, in 2004 and 2006, respectively, and the Ph.D. degree in aerospace engineering with a minor in cognitive science from the University of Minnesota, Twin Cities, MN, USA, in 2011. He is currently an Associate Professor in mechanical and aerospace engineering with the University of California (UC), Davis, CA, USA.

His current research interests include control theory, machine learning, formal methods, and their applications to human-machine systems, cyber-physical systems, and neural engineering. Before joining UC Davis in 2015, he was a Postdoctoral Researcher at the Laboratory for Intelligent Mechatronic Systems and the Hybrid and Networked Systems Lab, Boston University, Boston, MA, USA.

...



SAMUEL P. ESHIMA received his Bachelor's degree in mechanical engineering from Kanazawa University in 2018 and Master's degree in aerospace engineering sciences from the University of Colorado, Boulder, CO, USA, in 2020. He is currently a Ph.D. candidate in the Department of Smead Aerospace Engineering Sciences, University of Colorado, Boulder, Colorado, USA. His research is on sensor suite optimization for ECLSS anomaly detection and diagnostics.