Counterfactual Fault Responsibility in a Dynamic Structural Causal Model of a DC Motor

Emir Mujić
Institute of Software Engineering and Artificial Intelligence, Graz University of Technology, Austria

Franz Wotawa[1]
Institute of Software Engineering and Artificial Intelligence, Graz University of Technology, Austria

Abstract

Model-based diagnosis can identify fault hypotheses that are compatible with observed abnormal behaviour. Still, it does not necessarily determine which fault is responsible for a particular symptom when several faults are active simultaneously. We propose a counterfactual framework for post-diagnostic fault responsibility analysis in dynamical systems. The physical process is represented as a time-indexed structural causal model, while faults are encoded as interventions on individual system mechanisms. Responsibility is evaluated by comparing the factual system trajectory with counterfactual trajectories in which candidate faults are removed while the operating conditions and control structure are kept fixed. We instantiate the framework on an electro-thermo-mechanical model of a DC motor and evaluate it in two simulated multi-fault scenarios. The results show that counterfactual fault removal can separate the contributions of different faults to individual symptoms and distinguish severity contribution from necessity for a failure event. The proposed approach therefore complements conventional diagnosis with symptom-level causal explanations of faulty system behaviour.

2012 ACM Subject Classification Computing methodologies → Causal reasoning and diagnostics

Keywords and phrases Counterfactual Reasoning, Fault Responsibility, Structural Causal Models, Dynamic Systems, Model-Based Diagnosis

Digital Object Identifier 10.4230/OASIcs.DX.2026.16

Category Short Paper

Funding This research was funded in whole or in part by the Austrian Science Fund (FWF) 10.55776/COE12.

1 Introduction

Fault diagnosis in dynamical systems commonly addresses two related tasks: detecting that the system has departed from nominal behaviour and isolating the fault or set of faults that is compatible with the observed measurements. Standard fault detection and isolation methods typically achieve this through residuals, analytical redundancy relations, observers, or structural relations between system variables. These methods are well suited to determining whether a fault is detectable or distinguishable from other faults. However, once several faults are simultaneously active, identifying the active fault set does not necessarily explain which fault is responsible for a particular symptom or failure event. This distinction is important because compatibility, detectability, and causal responsibility are different notions. A fault may produce a nonzero residual or contribute to abnormal behaviour without being necessary for a specified event. Conversely, a fault may be largely hidden in the measured output by feedback control while still being responsible for increased actuator effort, thermal

[1] Authors appear in the alphabetic order.

stress, or the crossing of a safety threshold. Residual magnitudes alone therefore do not generally establish whether a failure would still have occurred had one of the active faults been absent.

We address this problem using counterfactual reasoning. Rather than asking only whether a fault hypothesis is consistent with the observations, we ask whether the observed symptom would still have occurred under an intervention that removes the fault while preserving the remaining operating conditions. This requires a model in which faults modify the physical mechanisms that generate the system trajectory. Structural causal models (SCMs) provide such a representation through modular structural assignments and explicit interventions.

In this work, a dynamical system is represented as a time-indexed structural causal model derived from its physical equations. Electrical and mechanical faults are encoded as interventions on individual mechanisms rather than as direct changes to measured variables. Starting from a factual multi-fault trajectory, we generate counterfactual trajectories in which candidate faults are removed individually while the initial conditions, reference task, load profile, and controller remain fixed. The resulting comparisons provide a post-diagnostic explanation of fault responsibility. They can distinguish whether a fault contributes to the magnitude of a symptom, is mainly responsible for one symptom rather than another, or is necessary for a failure event to occur. The proposed method therefore complements conventional FDI. FDI supplies candidate fault mechanisms, while counterfactual analysis explains their symptom-specific consequences after diagnosis.

We demonstrate the approach using a DC motor model. The model contains electrical, mechanical, and thermal feedback mechanisms and therefore provides a suitable setting for studying fault propagation under closed-loop control. Two simulated multi-fault experiments are used to evaluate whether counterfactual fault removal can separate responsibility for angle-synchronous disturbances and overheating.

The main contributions of this paper are (1) the formulation of a physical DC motor model as a dynamic SCM, (2) the representation of electrical and mechanical faults as interventions on physical mechanisms, (3) a counterfactual fault-removal procedure for assigning symptom-specific responsibility.

2 Related Work

In Reiter's theory of diagnosis [13], a diagnosis is a set of components assumed to be abnormal such that the system description is consistent with the observations. De Kleer and Williams extended this framework to multiple-fault diagnosis through the General Diagnostic Engine [5], while later work studied minimal and kernel diagnoses and structured system descriptions [4, 3]. These approaches provide rigorous methods for generating candidate fault explanations, but primarily ask which abnormality assumptions are consistent with the observations rather than whether a fault was responsible for a particular failure event.

Abductive diagnosis places greater emphasis on explanation by selecting hypotheses that imply the observed behaviour. Poole proposed an early general framework for abductive explanation and prediction [12], while Console and Torasso integrated models of correct and faulty behaviour [1]. Friedrich, Gottlob, and Nejdl further connected diagnosis with decisions and therapy [6]. Cordier et al. [2] relate AI-based model-based diagnosis to control-based fault detection and isolation (FDI). Both traditions identify faults through violations of healthy-system relations, although AI diagnosis commonly uses symbolic component models and logical conflicts, whereas FDI typically uses numerical dynamic models and residuals.

Structural methods extend FDI to larger dynamical systems by reasoning about the connections between variables and equations, sometimes without requiring a complete numerical model. They have been applied to fault detectability and isolability, sensor placement, and residual-generator selection [14, 15]. In contrast, our work assumes that relevant fault mechanisms are available and uses counterfactual simulation to determine which fault is responsible for a specified symptom.

Causal models provide a natural representation of faults as changes to system mechanisms. Pearl's structural causal model framework formalizes interventions and counterfactuals [10], while Halpern and Pearl define actual causality in terms of whether an event remains necessary under suitable counterfactual changes [8]. This perspective is directly relevant to fault responsibility analysis, as shown in the work of Weiblich et al. [16], where they construct counterfactual trajectories and score them to indicate whether the failure was removed. We employ a similar idea, however using an SCM constructed .

3 Methods

In this section, we describe the dynamic SCM used for the DC motor. A DC motor is a useful example of a causal dynamical system because its main mechanisms are known from physics: voltage affects armature current, current generates torque, torque changes angular velocity and rotor angle, and the resulting current and motion generate heat. Following the view of causal dynamical systems in [11], we model the plant dynamics as deterministic structural equations and add noise only at the measurement level.

As is the practice in causal reasoning, we make an assumption on the world and the model around the level of abstraction needed for the task. We do not explicitly model microscopic transient effects inside the motor, but rather focus on second-order dynamics. We also assume that a model of the motor is available, as well as the knowledge of how certain faults affect the motor.

The structure of the causal graph and variable ordering is defined by the physical direction of the mechanisms when considering a motor operating regime. The input voltage affects the armature current through the electrical dynamics. The current affects the produced torque, which affects the angular velocity and rotor angle. The angular velocity also causes back-EMF and opposes changes in the current, creating a primary loop in the dynamics. The current also generates heat through resistive losses, and angular velocity generates heat through friction, causing the winding temperature to change the armature resistance. The winding temperature then affects the armature resistance, closing an electro-thermal feedback path in the causal graph.

3.1 Structural Causal Models

A Structural Causal Model (SCM) [10, 8, 7] represents a system in terms of causal mechanisms for each variable. The central property of an SCM is modularity. Each structural assignment represents an autonomous mechanism that can be modified without changing other mechanisms in the model. An intervention in an SCM $\mathrm{do}(X_k = x)$, replaces the original structural assignment for variable X_k by the constant assignment $X_k := x$. This distinguishes intervention from conditioning: observing $X_k = x$ does not change the data-generating mechanism, whereas $do(X_k = x)$ does.

SCMs support counterfactual reasoning. A counterfactual asks what would have happened under a different intervention, given the same factual situation. Conceptually, this is evaluated by first inferring the relevant background conditions from the factual observation, then modifying the intervened mechanism, and finally evaluating the modified model under the same background conditions. In this paper, we use a domain-specific counterfactual fault removal criterion to distinguish severity contribution from necessity for a symptom event.

To correctly define a counterfactual on a DC motor, we use an SCM. Since the DC motor is a dynamical system, we use a time-indexed SCM. Instead of a classic static assignment where any variable X_i in an SCM is defined by a function of its parents Pa_i and exogenous inputs U_i, $X_i := f_i(\mathrm{Pa}_i, U_i)$, the model specifies how the variable X_i evolves from one time step to the next:

$$X^i_{t+\Delta t} := f(X^i_t, \mathrm{Pa}^i_t, U^i_t), \tag{1}$$

where now the variable is dependent on its previous state as well as the time-indexed states of the parents and exogenous variables. This connects SCMs with classical state-space and differential-equation models; the equations describe physical time evolution, while the causal interpretation specifies which mechanisms affect how and which each variable is affected.

3.2 A causal dynamical model of a DC motor

As our plant model, we develop a custom model of a DC electric motor. The model is a dynamic, nonlinear, multivariate system that falls into the category of dynamic structural causal models with noise added to the measured values, as detailed in [11]. As a starting point, we use well-known descriptions of the basic DC motor behaviour:

$$L\frac{di_a(t)}{dt} = u_{in}(t) - R_a(t)i_a(t) - k\omega(t), \tag{2}$$

$$J\frac{d\omega(t)}{dt} = ki_a(t) - \tau(\omega) - M_L(t). \tag{3}$$

where $u_{in}[V]$ is the input voltage, $R_a[\Omega]$ is the coil armature resistance, $i_a[A]$ is the armature current, $k[Vs]$ [2] is the back-EMF constant, $\omega[rad/s]$ is the angular velocity, and $L[H]$ is the inductance. $J[kgm^2]$ is the moment of inertia of the rotor, $\tau[Nm]$ is the friction torque and $M_L[Nm]$ is the load torque.

We model the friction $\tau(\omega)$ as a Stribeck curve. It consists of three parts: the startup, mixed, and running friction. General mechanical friction is given by the following formula as opposing torque:

$$\tau(\omega) = \left(\tau_C + (\tau_{max} - \tau_C) \cdot e^{-\left(\frac{|\omega|}{\omega_C}\right)^2}\right) \cdot \tanh\left(\frac{\omega}{\epsilon}\right) + k_r \cdot \omega. \tag{4}$$

In Equation 4, τ_C is the Coulomb friction, τ_{max} the maximum friction value, ω is the angular velocity, ω_C is a value of ω at which we switch from the startup to mixed friction, k_r is the viscous friction coefficient, and ϵ is a small value used to saturate the $\tanh(\cdot)$ function to make sure that the friction is always opposing rotation.

The thermal part of the model is added because the winding resistance is not constant, but changes with the winding temperature. This allows the model to account for overheating. The temperature-dependent armature resistance is modelled as

$$R_a(t) = R_0\left(1 + \alpha_{Cu}(T_w(t) - T_{ref})\right), \tag{5}$$

[2] The back-EMF constant k in both the electrical and mechanical equations is equivalent in SI units: $[\frac{Nm}{A}] = \frac{[J]}{[\frac{C}{s}]} = \frac{[J]}{[C]}[s] = [Vs]$.

where R_0 is the nominal armature resistance, $\alpha_{Cu}[1/K]$ is the copper temperature coefficient and T_{ref} is the reference temperature. The copper losses are modelled as $P_{Cu}(t) = i_a^2(t)R_a(t)$.

The winding temperature is then described as

$$C_w \frac{dT_w(t)}{dt} = P_{Cu}(t) - \frac{T_w(t) - T_h(t)}{R_{wh}}, \tag{6}$$

where $C_w[J/K]$ is the winding thermal capacitance, $T_w(t)[K]$ is the winding temperature, $T_h[K]$ is the housing temperature, and $R_{wh}[K/W]$ is the thermal resistance between the winding and the housing. The housing temperature is modelled as

$$C_h \frac{dT_h(t)}{dt} = P_h(t) + \frac{T_w(t) - T_h(t)}{R_{wh}} - \frac{T_h(t) - T_a}{R_{ha}}, \tag{7}$$

where C_h is the housing thermal capacitance, R_{ha} is the thermal resistance between the housing and ambient, and T_a is the ambient temperature. The term $P_h(t)$ represents the heat generated by mechanical losses. Power lost to mechanical movement is proportional to viscous friction, so we approximate it by a scaling factor $P_h(t) = \eta_h k_r \omega^2(t)$, where η_h is the scaling factor.

The full electro-thermo-mechanical model is given by the state vector:

$$x(t) = \begin{bmatrix} i_a(t) & \omega(t) & \theta(t) & T_w(t) & T_h(t) \end{bmatrix}^T, \tag{8}$$

equations describing the causal mechanisms for the state endogenous variables written in Equations 2, 3, 6, 7 and the expression $\omega(t) = \dot{\theta}(t)$. To complete the causal model, we define the exogenous variables in a vector $u(t) = \begin{bmatrix} u_{el}(t) & M(t) & T_{\text{ambient}}(t) \end{bmatrix}^T$. $u_{el}(t)$ is the effective input voltage, $M(t)$ is the outside torque required for the task and $T_{\text{ambient}}(t)$ is the ambient temperature. A graph of the dynamic SCM can be seen in Figure 1. The figure shows the model's causal relations, but leaves out some mechanisms since the graph would be too crowded. A graph like the one in Figure 1 is meant to easily show the causal relationships and help with determining how interventions (and with that counterfactual) would propagate through the model.

Each fault is represented as an intervention on one structural mechanism of the motor model. Thus, a fault does not directly assign a value to an observed variable. Instead, it modifies the equation that generates the affected variable. The observed symptoms then arise through the propagation of this mechanism change through the dynamical system. The electrical faults in the model are defined as follows. A resistance fault increases (FEF1) the effective armature resistance $R_0 \to R_0 + \Delta R$. This affects both the electrical voltage equation and the copper losses. A brush or commutator voltage drop fault (FEF2) introduces an additional voltage loss to the input $u_{in}(t) = u_{el}(t) - u_{drop}(t)$. A localized commutator fault (FEF3) introduces a voltage drop only at a specific rotor angle $u_{drop}(t) = \Delta V_s \cdot g(\theta(t))$.

Here, $g(\theta)$ is an angular gate function:

$$g(\theta) = \begin{cases} 1, & |\text{wrap}_\pi(\theta - \theta_0)| \le \delta, \\ 0, & \text{otherwise.} \end{cases} \tag{9}$$

This means that the fault is only active when the rotor passes through a defective angular sector. A demagnetization fault (FEF4) is modelled by reducing the motor constant $k \to k - \Delta k$. Since k appears in both the electrical and mechanical equations, this fault simultaneously changes the back-EMF and the produced motor torque.

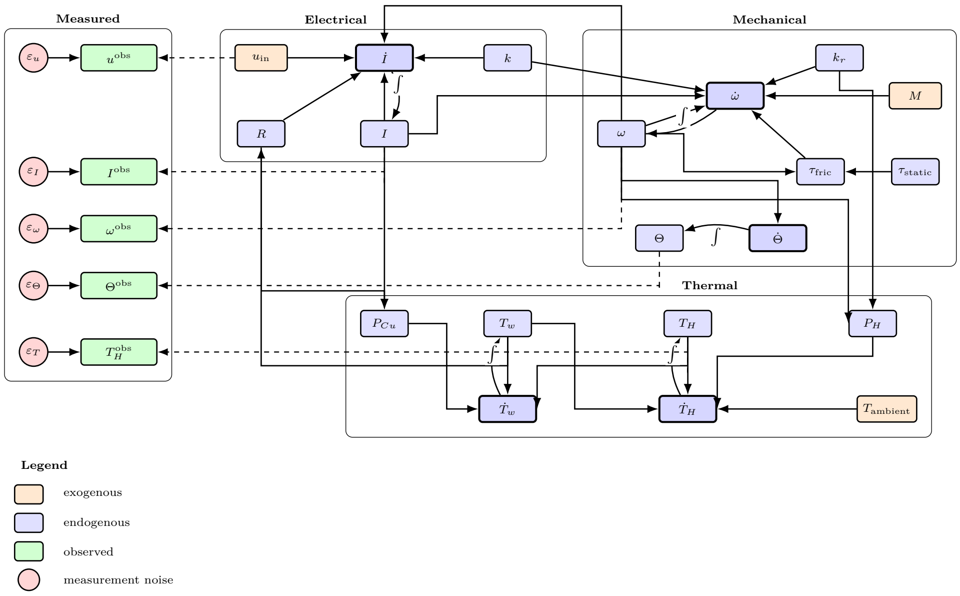

Figure 1 A dynamic SCM representing a DC electric motor model. For clarity, this graph does not contain all variables of the full model.

The mechanical faults are modelled similarly. Increased viscous friction fault (FMF1) modifies the viscous friction coefficient $k_r \to k_r + \Delta k_r$. An increased Coulomb friction (FMF2) fault modifies the Coulomb friction level: $\tau_C \to \tau_C + \Delta\tau_C$. An inertia fault (FMF3) changes the effective inertia of the system $J \to J + \Delta J$. Finally, a periodic mechanical disturbance (FMF4) is represented as an angle-dependent torque $M_p(t) = \Delta M_p \cdot g(\theta(t))$.

Finally, we write the SCM model in a compact form. At every time step, the previous state x_t, the input voltage u_t, the load torque $M_{L,t}$ and the active fault vector F_t determine the next state $x_{t+\Delta t}$:

$$x_{t+\Delta t} = f_{SCM}\left(x_t, u_t, M_{L,t}, F_t\right). \tag{10}$$

The measured values are then generated from the true simulated states by adding measurement noise: $y_t = h(x_t) + \varepsilon_t$. The variable y_t denotes the measured variables, $h(\cdot)$ is the measurement function, and ε_t is measurement noise. The measurement function $h(\cdot)$ represents a sampling algorithm from simulation sampling time to sensor sampling. To mimic reality, we keep the states of the motor hidden and only observe certain variables that would be easily measurable with sensors, i.e., the voltage, current, angle, rotational velocity, and the housing temperature.

The motor is simulated by numerically integrating the structural equations. For this, we use a fourth-order Runge–Kutta method. During one integration step, the input voltage, load torque, and active fault configuration are kept constant. Faults can therefore be scheduled by changing the corresponding fault variable at a chosen time.

3.3 Counterfactual fault responsibility analysis

An SCM can also be used to explain which fault mechanisms are responsible for specific observed symptoms. This is particularly useful when multiple faults are active at the same time. In such cases, identifying the set of active faults does not necessarily explain which fault caused which part of the faulty behaviour. We use counterfactual fault removal to assign responsibility to individual mechanisms.

Let the factual fault configuration set be denoted by $F = \{F_1, F_2, \ldots, F_k\}$,

where multiple fault mechanisms may be active, that is $F_i > 0, i = 1, \ldots, k$. This is the set of all faults we predefine to be possible or probable in a system. The factual trajectory is generated by simulating the SCM under a fault configuration containing at least a single active fault F^{fact}:

$$x_{t+\Delta t}^{fact} = f_{SCM}\left(x_t^{fact}, u_t, M_{L,t}, do(F = F^{fact})\right). \tag{11}$$

This represents a real scenario where a system is detected to be faulty by an arbitrary measure, but it is unknown what exactly the root cause is; that is, we know that there is at least a single fault active $F_j > 0$, but don't know which fault exactly is to blame such that $F_j \in F$.

To evaluate the responsibility of any particular fault $F_i \in F$, we construct a counterfactual world in which this fault is removed while the remaining background conditions are kept fixed. This corresponds to a counterfactual trajectory with a single fault F_i removed:

$$x_{t+\Delta t}^{-i} = f_{SCM}\left(x_t^{-i}, u_t, M_{L,t}, do(F_i = 0, F_{\setminus i} = F_{\setminus i}^{fact})\right). \tag{12}$$

Here, the same initial state and torque profile are used as in the factual case; the controller input is recomputed in each counterfactual branch since the system is closed in a control loop.

Responsibility score is then measured by comparing the symptom measure between the observed and the counterfactual trajectories. Let $S_s(x)$ denote a scalar measure for any symptom s. The responsibility of fault F_i for symptom s is defined as

$$R_{i \to s} = S_s(x^{fact}) - S_s(x^{-i}). \tag{13}$$

A large positive value means that removing fault F_i reduced the symptom s and therefore that F_i is responsible for that symptom. A value close to zero means that the symptom is largely unchanged when the fault is removed.

This analysis allows the diagnosis to be refined from a fault-level statement to a symptom-level explanation. For example, a localized electrical defect may be responsible mainly for angle-synchronous current and tracking ripple, while increased Coulomb friction may be responsible mainly for persistent high current and thermal load. Counterfactual removal therefore explains how different fault mechanisms contribute to different aspects of the observed faulty behaviour.

4 Experiments

We simulate the model parameters found in a datasheet for an industrial DC motor, MAXON RE40 (model number 148867) motor [9].

We evaluate the counterfactual responsibility procedure on two simulated fault-assignment experiments. The purpose of these experiments is not to identify an unknown fault, but to explain which fault mechanism best explains the observed symptom. Each experiment follows the same procedure: simulate the faulty branch, remove each candidate fault by intervention, compute symptom-specific scores, and assign contribution or necessity by comparing factual and counterfactual branches.

All experiments use the previously described SCM, simulated with an integration step of $\Delta t = 10^{-4}$ s. The faults are present from the start of the simulation, so the system never evolves under nominal conditions. The motor is controlled by the same closed-loop position controller in every factual and counterfactual branch. The full control loop consists of a PI controller for the positional error, $e(t) = \theta_{ref}(t) - \theta(t)$, voltage saturation of 24 V, and a small feedforward term to compensate for the motor's own inertia.

The electrical and mechanical quantities are sampled at $f_{\text{el}} = 10\,\text{kHz}$ and $f_{\text{mech}} = 1\,\text{kHz}$, respectively, with added Gaussian noise. The same mechanical sampling is used for the controller. Position is measured with a 10-bit relative encoder. Noise is added to all measurements with standard deviations of $\sigma_u = 0.05\,\text{V}$, $\sigma_{i_a} = 0.02\,\text{A}$, $\sigma_\omega = 0.5\,\text{rad/s}$, $\sigma_\theta = 0.002\,\text{rad}$ and $\sigma_{T_h} = 1\,°\text{C}$. The values of both the sampling frequencies and measurement noise are chosen as plausible for sensors accompanying a DC motor in use.

Unlike the real setting where the initial states of the system would come from an abduction step (filtering), we use the simulated values directly as we have access to them. To verify that the method works with a mismatch between the simulated motor and the model, we run $N = 100$ tests with random parameter values uniformly sampled around nominal values (from [9]) by 25%.

4.1 Angle-synchronous ripple

The first experiment tests the method on two simultaneously active faults: FEF3, an angle-gated localized electrical voltage drop, and FMF2, an increase in Coulomb friction. Hypothetical scenario represents a damaged brush-commutator contact for $\delta = 10\,°$ where the voltage drops by 2 V, and worn bearings increasing the load torque by $\Delta\tau_C = 0.1\,\text{Nm}$ at angular velocity around 0 rad/s.

The reference angle repeatedly passes through the damaged angular sector, making the localized electrical defect observable. At the same time, the trajectory contains low-speed reversals and loaded crawling phases. The motor is simulated for 120 s.

The simulated world faults FEF3 = 1 and FMF2 = 1 active is compared to three different counterfactual worlds: $do(\text{FEF3} = 0)$, $do(\text{FMF2} = 0)$ and $do(\text{FEF3} = 0, \text{FMF2} = 0)$.

These counterfactual worlds are chosen to show whether the method can correctly assign the blame and separate the effects of faulty behaviour from individual faults and their combined effect. In a more realistic setting, all possible faults would be included as possible, or some subset chosen by a heuristic. The expected assignment is that FEF3 is responsible mainly for angle-synchronous current and tracking ripple, while FMF2 is responsible mainly for persistent high current and overheating.

To evaluate the angle-localized symptom, we define a diagnostic angular sector around the defective commutator position:

$$g_{\text{diag}}(\theta) = \mathbf{1}\left(|\text{wrap}_\pi(\theta - \theta_0)| \le \delta_{\text{diag}}\right), \tag{14}$$

with $\theta_0 = 0$ and $\delta_{\text{diag}} = 15°$.

This diagnostic sector is slightly wider than the physical fault sector to capture the dynamic response around the defective angular region. Given the information from the encoder, it's possible to estimate at which angles the faults appear and therefore estimate the diagnostic sector δ_{diag}.

For each counterfactual trajectory, we compute a set of scores, each pointing toward a symptom. To calculate the scores, we only use the measurements we plausibly could have in the system already. We record tracking RMSE, RMS and peak current, RMS and peak voltage, voltage-saturation time (above $0.98u_{max}$) and peak temperatures. The thermal symptom is evaluated using the winding and housing temperature. A winding-overheating event is declared if $\max_t T_w(t) \ge T_{w,\text{lim}}$. A housing-temperature event is defined analogously using $T_{h,\lim} = 70°\mathrm{C}$. To measure not only whether the threshold is crossed but also how severe the violation is, we use the thermal excess cost.

$$J_{\mathrm{thermal}} = \frac{1}{N}\sum_k \left(\max(0, T_w(k) - T_{w,\lim})\right)^2 + 0.25\frac{1}{N}\sum_k \left(\max(0, T_h(k) - T_{h,\lim})\right)^2, \quad (15)$$

which grows quadratically if they exceed the limits and is zero if the respective temperatures stay in the safe region below. We also say that overheating of the winding is four times as harsh, hence the 0.25 factor in the expression.

To evaluate the localized electrical fault symptom, we compute the differences in the angle tracking values and current in the damaged commutator sector (denoted by $\in$) and outside of it ($\notin$). A diagnostic angular gate at the damaged commutator is defined in Equation 14. To describe the angle-synchronous tracking ripple we write $R_\theta = \mathrm{RMSE}_{e_\theta,\in} - \mathrm{RMSE}_{e_\theta,\notin}$ and for the angle-synchronous current ripple, $R_I = \mathrm{RMS}_{I,\in} - \mathrm{RMS}_{I,\notin}$

An angle-ripple event is detected if either of these quantities exceeds its threshold of 0.05 rad and 0.75 A respectively. The corresponding angle-ripple cost is

$$J_{\mathrm{angle}} = \left(\frac{\max(0, R_\theta)}{R_{\theta,\lim}}\right)^2 + \left(\frac{\max(0, R_I)}{R_{I,\lim}}\right)^2. \quad (16)$$

Finally, the overall severity score is a weighted sum of mean tracking error, mean current, voltage saturation, thermal excess, and angle-synchronous ripple:

$$J_{\mathrm{total}} = w_{\mathrm{track}}J_{\mathrm{track}} + w_I J_I + w_{\mathrm{sat}}J_{\mathrm{sat}} + w_{\mathrm{thermal}}J_{\mathrm{thermal}} + J_{\mathrm{angle}}. \quad (17)$$

In our experiments, all weights were set to 1. The individual score components are chosen to match the expected physical effects of the faults we model and are aware of the effects. The localized electrical fault FEF3 is angle-dependent, so its main symptom should appear as an increase in current or tracking error when the rotor passes through the damaged angular sector. Increased Coulomb friction FMF2 primarily increases the required motor current and therefore the winding and housing temperatures. Voltage saturation is included because severe faults may force the controller to operate at its actuator limit.

Counterfactual responsibility is evaluated by comparing these scores to the SCM factual branch. For a removed fault F_i, the reduction in a symptom score is $\Delta J_s^{-i} = J_s^{\mathrm{fact}} - J_s^{-i}$.

A large positive value means that removing F_i reduces symptom s, and therefore that the fault contributes to that symptom. If an event occurs in the factual branch but disappears under $do(F_i = 0)$, then F_i is treated as necessary for that event. This separates event prevention from severity reduction: a fault may be responsible for making a symptom worse even if it is not individually necessary for the event threshold to be crossed.

4.2 Overheating responsibility

The second experiment evaluates responsibility for winding overheating in the presence of partial demagnetization (FEF4) and increased viscous friction (FMF1). This combination represents an aged motor with weakened permanent magnets and worn bearings. The factual configuration is $\mathrm{FEF4} = 1, \mathrm{FMF1} = 1$, with $\Delta k = 0.05k_0$ and $\Delta k_r = 30k_{r0}$. The motor is simulated for 1200 s to capture its slower thermal dynamics. The reference and load profiles are chosen to excite viscous friction through sustained moderate-to-high-speed operation. The motor follows a constant-speed trajectory while the load torque increases in steps from 120 mNm to the nominal value of 177 mNm.

The factual world is compared to five counterfactual worlds: do(FEF2 = 0), do(FEF4 = 0), do(FMF2 = 0), do(FMF1 = 0) and do(FEF4 = 0, FMF1 = 0). FEF2 and FMF2 are inactive in the factual world and are therefore included as negative controls. FEF4 and FMF1 are the candidate mechanisms for explaining the observed overheating.

We use the tracking, current, voltage-saturation, and thermal scores defined in Section 4.1, omitting the angle-ripple term. The primary event is winding overheating:

$$E_{\text{thermal}} = \mathbf{1}\left[\max_k T_w(k) \geq T_{w,\text{lim}}\right], \tag{18}$$

where $T_{w,\text{lim}} = 90°\text{C}$. Thermal severity is quantified using J_{thermal} from Equation 15, and the total score is the equally weighted sum of the remaining score components.

A fault contributes to thermal severity if its removal decreases the peak winding temperature by more than 5°C, a heuristically selected threshold. The reduction is measured as

$$\Delta T_{w,\max}^{-i} = \max_k T_w^{\text{fact}}(k) - \max_k T_w^{-i}(k).$$

A stronger condition is necessity: fault F_i is necessary for overheating when the event occurs in the factual trajectory but disappears after its removal,

$$E_{\text{thermal}}^{\text{fact}} = 1 \quad \text{and} \quad E_{\text{thermal}}^{-i} = 0. \tag{19}$$

5 Results

5.1 Fault blame assignment - Angle-synchronous ripple

Figure 2 shows the position tracking error, armature current, and winding temperature for the factual and counterfactual trajectories. The upper panel reveals angle-localized disturbances in branches where FEF3 remains active. These disturbances are strongly suppressed when FEF3 is removed, supporting the interpretation that FEF3 is responsible for the angle-synchronous tracking symptom.

The middle panel shows that branches with FMF2 active require a persistently higher armature current, while FEF3 additionally produces localized current spikes. Removing FMF2 lowers the current magnitude throughout the experiment, whereas removing FEF3 primarily suppresses the angle-synchronous spikes.

The winding-temperature trajectories in the lower panel show that FMF2 is necessary for the overheating event. Every branch in which FMF2 is removed remains below the winding-temperature limit, while removing FEF3 alone does not prevent overheating.

In Table 1, we show numerical results of our experiments, the values of score functions from Section 4.1, and winding temperature values for each simulated world. From the table, we can conclude that intervention $do(\text{FEF3} = 0)$ always removes the angle ripple event, while the intervention $do(\text{FMF2} = 0)$ always removes the overheating event. The results indicate that the effects of the two faults are interaction-dependent. Removing one fault does not necessarily reduce every cost component, because the intervention changes the system trajectory and, consequently, the conditions under which the remaining fault propagates. For example, removing only FMF2 eliminates the thermal event but increases (J_{ang}). This increase is associated with a change in the angle-dependent current behaviour: the RMS current inside the affected angular sector becomes higher than outside it. This suggests that, in the absence of FMF2, the effect of FEF3 on the current becomes more pronounced within

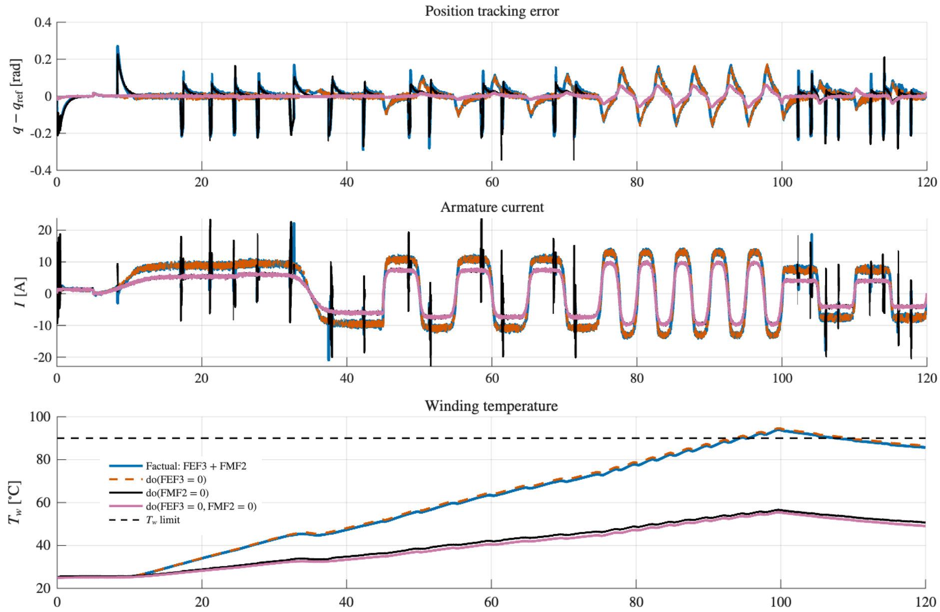

Figure 2 Counterfactual results for the FEF3+FMF2 scenario: position tracking error, armature current, and winding temperature. Removing FEF3 suppresses angle-synchronous disturbances, while removing FMF2 reduces current demand and prevents overheating.

the affected angular sector. We therefore do not interpret the total score alone as a global responsibility measure. Instead, responsibility is assigned symptom-wise: FMF2 is necessary for the thermal event, while FEF3 is necessary for the angle-ripple event.

Table 1 Counterfactual fault-assignment results for the combined FEF3 and FMF2 experiment. The factual branch contains both the angle-gated electrical fault FEF3 and the increased Coulomb-friction fault FMF2. A dash indicates that the winding-temperature limit was not reached.

Intervention	Removed fault	max T_w [°C]	RMSE$_q$ [rad]	RMSE$_{q,\text{sec}}$ [rad]	J_T	J_{ang}	J_{tot}	T_w event	Angle event
Factual: FEF3 + FMF2	none	94.00	0.0606	0.1145	0.5090	1.585	83.631	yes	yes
do(FEF3 = 0)	FEF3	94.70	0.0430	0.0121	0.8496	0.000	83.472	yes	no
do(FMF2 = 0)	FMF2	56.63	0.0444	0.1186	0.0000	4.537	43.124	no	yes
do(FEF3 = 0, FMF2 = 0)	FEF3+FMF2	55.48	0.0163	0.0049	0.0000	0.000	36.038	no	no

In the randomized-parameter repetitions, the method assigned FEF3 as necessary for the angle-ripple event and FMF2 as necessary for the thermal event in all 100 runs.

5.2 Fault blame assignment - Overheating responsibility

Figure 3 shows the winding temperature and armature current for the factual branch and the counterfactual branch in which FMF1 is removed. With FMF1 active, the motor requires a persistently higher armature current throughout the loaded operating phases. After the intervention do(FMF1 = 0), the current demand is substantially reduced, indicating that increased viscous friction primarily appears as an increase in the actuation effort required to follow the same task.

The upper panel shows the thermal consequence of this increased current demand. In the factual branch, the winding temperature crosses the 90°C limit and reaches approximately 106°C. When FMF1 is removed, the winding temperature remains below the limit throughout the experiment. FMF1 is therefore necessary for the winding-overheating event under the considered operating conditions.

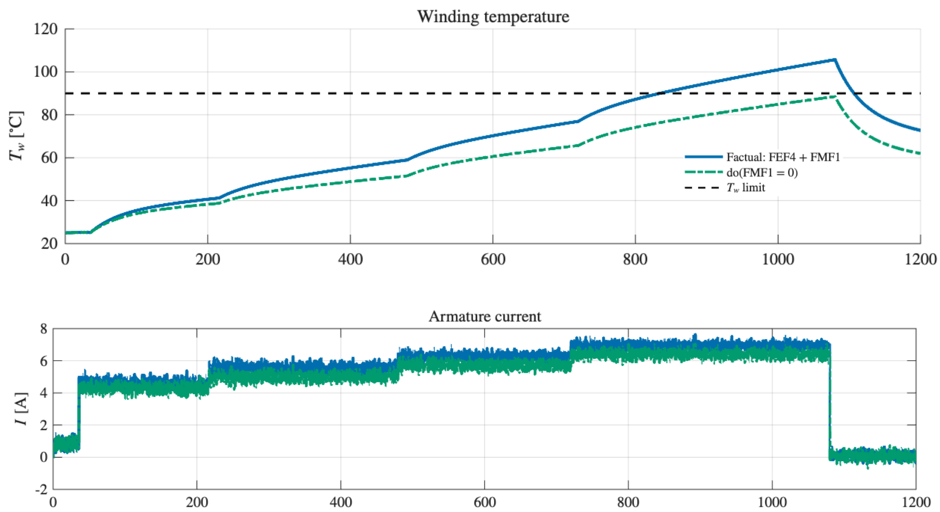

Figure 3 Counterfactual winding temperature and armature current for the FEF4+FMF1 scenario. Removing FMF1 reduces current demand and prevents overheating.

Table 2 reports the numerical results for all counterfactual branches. The counterfactual results indicate that the observed overheating is mainly attributable to FMF1, while FEF4 has a smaller but non-negligible contribution. Removing FEF4 reduces the peak winding temperature and decreases (J_T), but it does not prevent the winding-temperature event. In contrast, removing FMF1 reduces the peak winding temperature, eliminates the thermal cost, and prevents the overheating event. Removing both faults produces a further reduction in temperature and total cost, suggesting that FEF4 additionally increases the thermal severity under the conditions created by FMF1. The interventions on FEF2 and FMF2 serve as negative controls; their results are identical to the factual trajectory, as expected because the faults are not active in the simulation. This supports the specificity of the counterfactual fault-assignment procedure.

To evaluate robustness to parameter mismatch, we repeated the experiment using the randomized parameter offsets described in Section 4. The method correctly assigned responsibility to FMF1 in 84% of the 100 randomized repetitions. The failed repetitions occurred when parameter mismatch changed the model's thermal behaviour, and it became less clearly attributable to FMF1 under the chosen threshold.

6 Discussion and Limitations

The experiments demonstrate that counterfactual fault removal can separate the responsibilities of simultaneously active mechanisms. In the first experiment, removing FEF3 eliminates angle-synchronous ripple, whereas removing FMF2 prevents overheating. In the second,

Table 2 Counterfactual fault-assignment results for the combined FEF4 and FMF1 experiment. The factual branch contains both partial demagnetization FEF4 and increased viscous friction FMF1. A dash indicates that the winding-temperature limit was not reached.

Intervention	Removed fault	max T_w [°C]	max T_h [°C]	RMSE$_q$ [rad]	J_T	J_{tot}	T_w event	Prevents T_w event
Factual: FEF4 + FMF1	none	105.70	72.53	0.0057	19.4562	51.7240	yes	–
do(FEF2 = 0)	FEF2	105.70	72.53	0.0057	19.4562	51.7240	yes	no
do(FEF4 = 0)	FEF4	96.52	67.36	0.0057	1.6265	30.7509	yes	no
do(FMF2 = 0)	FMF2	105.70	72.53	0.0057	19.4562	51.7240	yes	no
do(FMF1 = 0)	FMF1	88.51	61.72	0.0056	0.0000	27.1003	no	yes
do(FEF4 = 0, FMF1 = 0)	FEF4+FMF1	81.30	57.63	0.0055	0.0000	24.4604	no	yes

removing FEF4 reduces thermal severity without preventing the event, while removing FMF1 prevents winding overheating. The method therefore distinguishes symptom responsibility, severity contribution, and event necessity even when the controller masks faults at the position level.

Counterfactual comparisons are performed under shared background conditions. The initial state, input trajectory, load torque, controller, and reference task remain fixed, so differences between trajectories can be attributed to the removed mechanism rather than to unrelated changes in operating conditions. This is particularly useful in closed-loop systems, where the controller may hide faults in some observed variables.

The randomized-parameter repetitions suggest that the method is not tied to one exact numerical realization. The FEF3/FMF2 assignment remained stable in all runs, whereas the FMF1 assignment succeeded in (84%) of the repetitions. This shows that counterfactual explanations depend on the accuracy of the model, as well as the intervention set and the chosen event thresholds.

The main limitation is that the experiments are simulation-based and use the same model class to generate and evaluate the counterfactual branches. Real motors contain unmodelled effects such as complex commutator behaviour, contact transients, magnetic saturation, temperature-dependent friction, changing loads, sensor delays, and actuator nonlinearities. The method also assumes that the relevant fault types and their effects on system mechanisms are known. In addition, the current experiments initialize counterfactual worlds using the true simulated state. In a real diagnostic setting, this state would need to be inferred from noisy observations, and estimation errors would affect the counterfactual trajectories and responsibility scores. Future work therefore includes a complete abductive pipeline in which hidden states and uncertain parameters are estimated from measured trajectories before evaluating counterfactual interventions.

7 Conclusion

We presented a dynamic structural causal model of a DC motor for counterfactual fault responsibility analysis. Faults are represented as interventions on physical mechanisms, allowing faulty behaviour to be explained through propagation in the electro-thermo-mechanical dynamics.

The experiments show that counterfactual fault removal can distinguish symptom responsibility, severity contribution, and necessity for a failure event. In the first experiment, FEF3 explains the angle-synchronous ripple while FMF2 explains overheating. In the second, FMF1 is necessary for winding overheating, while FEF4 mainly contributes to severity. Future work will focus on real-data validation, state and parameter estimation, and robustness under unmodelled fault mechanisms.

References

1 Luca Console and Pietro Torasso. Integrating models of correct behavior into abductive diagnosis. In *Proceedings of the European Conference on Artificial Intelligence (ECAI)*, pages 160–166, Stockholm, August 1990. Pitman Publishing.

2 M-O Cordier, Philippe Dague, François Lévy, Jacky Montmain, Marcel Staroswiecki, and Louise Travé-Massuyès. Conflicts versus analytical redundancy relations: a comparative analysis of the model based diagnosis approach from the artificial intelligence and automatic control perspectives. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, 34(5):2163–2177, 2004. doi:10.1109/TSMCB.2004.835010.

3 Adnan Darwiche. Model-based diagnosis using structured system descriptions. *Journal of Artificial Intelligence Research*, 8:165–222, 1998. doi:10.1613/JAIR.462.

4 Johan de Kleer, Alan K. Mackworth, and Raymond Reiter. Characterizing diagnosis and systems. *Artificial Intelligence*, 56, 1992. doi:10.1016/0004-3702(92)90027-U.

5 Johan de Kleer and Brian C. Williams. Diagnosing multiple faults. *Artificial Intelligence*, 32(1):97–130, 1987. doi:10.1016/0004-3702(87)90063-4.

6 Gerhard Friedrich, Georg Gottlob, and Wolfgang Nejdl. Hypothesis classification, abductive diagnosis and therapy. In *International Workshop on Expert Systems in Engineering*, pages 69–78. Springer, 1990. doi:10.1007/3-540-53104-1_32.

7 Joseph Y Halpern. Defaults and normality in causal structures. In *KR*, pages 198–208, 2008. URL: https://aaai.org/papers/KR08-020-defaults-and-normality-in-causal-structures/.

8 Joseph Y Halpern and Judea Pearl. Causes and explanations: A structural-model approach. part i: Causes. *The British journal for the philosophy of science*, 2005.

9 MAXON. Re40, part number 148867. https://www.maxongroup.de/maxon/view/product/motor/dcmotor/re/re40/148867. Accessed: 05-08-2026.

10 Judea Pearl. *Causality: Models, Reasoning, and Inference*. Cambridge university press, 2009.

11 Jonas Peters, Stefan Bauer, and Niklas Pfister. *Causal Models for Dynamical Systems*, pages 671–690. Association for Computing Machinery, New York, NY, USA, 1 edition, 2022. doi:10.1145/3501714.3501752.

12 David Poole. Explanation and prediction: an architecture for default and abductive reasoning. *Computational Intelligence*, 5(2):97–110, 1989.

13 Raymond Reiter. A theory of diagnosis from first principles. *Artificial Intelligence*, 32(1):57–95, 1987. doi:10.1016/0004-3702(87)90062-2.

14 M. Staroswiecki and G. Comtet-Varga. Analytical redundancy relations for fault detection and isolation in algebraic dynamic systems. *Automatica*, 37(5):687–699, 2001. doi:10.1016/S0005-1098(01)00005-X.

15 Christofer Sundström, Erik Frisk, and Lars Nielsen. Selecting and utilizing sequential residual generators in fdi applied to hybrid vehicles. *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, 44(2):172–185, 2013. doi:10.1109/TSMC.2013.2248147.

16 Juliane Weilbach, Sebastian Gerwinn, Karim Barsim, and Martin Fränzle. Counterfactual-based root cause analysis for dynamical systems. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 303–319. Springer, 2024.