POSITION: ALL COUNTERFACTUALS ARE WRONG, AND FEW ARE USEFUL

Anonymous authors
Paper under double-blind review

ABSTRACT

Counterfactuals—statements about what *might* have happened under different circumstances—offer a natural way to phrase causal questions and have gained increasing prominence in contemporary machine learning with applications ranging from evaluating medical decisions to identifying discrimination. This position paper argues that, while counterfactuals are powerful philosophical constructs for describing and formalizing human causal reasoning, the computation of *individual* counterfactual trajectories is riddled with dangers. In many practical scenarios, individualized counterfactuals are not truly useful or even harmful, either because the required assumptions cannot be adequately justified or because counterfactuals are conceptually misaligned with what would actually be informative for the task at hand.

1 INTRODUCTION

Counterfactual reasoning plays a fundamental role in everyday cognition and scientific discovery. Its mathematical formalization has led to a rich body of work. This raises the practical question motivating this paper: *when is computing a counterfactual actually useful?*

Counterfactuals answer "*what if?*" questions (Pearl, 2012; Starr, 2019) by describing how reality would have unfolded had the past been modified in some way. Our focus is on Pearl's structural causal models (SCMs) and their growing use for computing *individual* counterfactual trajectories in modern applications.

In this position paper, we argue that the practical utility of computing individual counterfactual trajectories is limited at best and harmful at worst. This is not because counterfactuals are theoretically flawed, but because they are often applied without the appropriate level of care. Our position therefore goes beyond "know your assumptions": for individual counterfactuals, the relevant assumptions are often unclear, untestable, or incompatible with standard modeling choices.

In the face of an increasing number of publications on counterfactuals in AI-based ethics and causal machine learning, we highlight the following three problems:

1. Claim A: Counterfactuals are often used or proposed to be used where other forms of reasoning are more appropriate.
2. Claim B: Counterfactuals depend on strong assumptions that are rarely justified in practice.
3. Claim C: Counterfactuals are highly sensitive to modeling choices, yet these choices are often made with purposes other than counterfactual estimation in mind.

We immediately clarify the following:

1. We do not argue that counterfactuals are conceptually flawed in general. Our claim is that their use often requires more care and justification than is commonly assumed.
2. We focus on *individual-level* counterfactual trajectories in a Pearl-style structural causal framework. Related issues may also arise in Rubin-style or group-level settings, but these are not our main concern.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

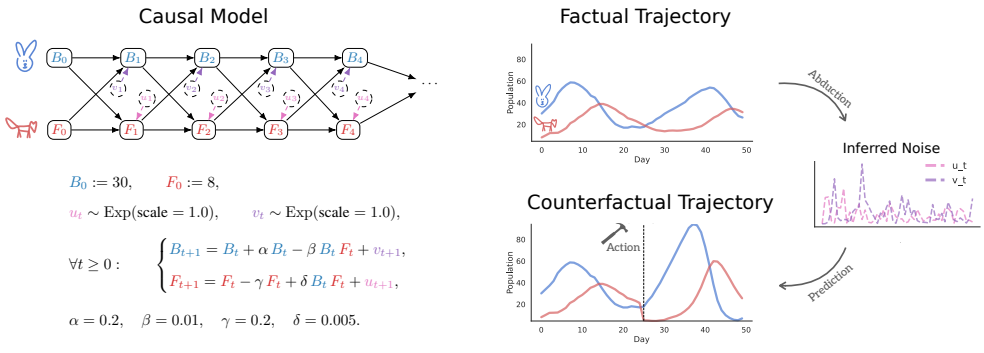

Figure 1: Consider a version of a predator–prey model represented as an SCM with exponential noise, where time is discretized in days and shown either as an abstract DAG (top left) or as a set of structural equations (bottom left). We can sample a single factual trajectory (top right). Treating this trajectory as the observation, we *abduct* (i.e., infer) the latent noise variables. Using the inferred noise, we then construct a counterfactual by simulating a hypothetical world to answer a *"what if?"* question: What would have happened had the fox population been reduced to 5 on day 25?

Motivation. This work is less concerned with the philosophical foundations of counterfactuals than with their proposed use in machine learning and related domains. In particular, **we observe a growing popularity of counterfactual modeling in time-series settings and in combination with neural-network–based inference.** In these contexts, earlier concerns about model specification and interpretability become more acute, motivating a renewed examination of when—and in what sense—counterfactual trajectories can be considered useful. These problems are also self-perpetuating. For prediction, poor modeling eventually surfaces as poor benchmark performance. Counterfactual claims have no observable counterpart, so mistaken assumptions produce no error signal: a method can be wrong about individual cases while appearing entirely healthy. This is particularly concerning as such methods are increasingly proposed for high-stakes decisions.

Relevance. More broadly, counterfactual reasoning increasingly underpins research and emerging tools in medicine, law, public policy, and AI, where it is used to explain outcomes (Melnychuk et al., 2022; Bica et al., 2020; Richens et al., 2022), assign responsibility (Triantafyllou et al., 2022; Andreas et al., 2023), determine liability (Bavli, 2020; Triantafyllou et al., 2022), assess discrimination (Alvarez & Ruggieri, 2023; Carey & Wu, 2022), and justify interventions (Garriga et al., 2024; Bergström et al., 2025; Hızlı et al., 2022). We provide a list of representative time-series applications in Appendix B.

Context. The debate about counterfactuals in science is not new and is perhaps best encapsulated by Dawid's critique (Dawid, 2000) and the responses it elicited from many pioneers of the field (Pearl, 2000; Rubin, 2000; Robins & Greenland, 2000). However, this earlier debate focused largely on whether counterfactuals are legitimate conceptual tools for statistical reasoning in principle. Dawid emphasized their untestability and argued that they are often unnecessary to settle decision problems. Pearl (2000), Robins, and others disagreed to varying degrees. Whether or not counterfactuals are strictly necessary to render concepts such as *risk* and *harm* amenable to statistical inference is still an ongoing debate, see (Dawid & Senn, 2023; Sarvet & Stensrud, 2025a; Mueller & Pearl, 2024; Sarvet & Stensrud, 2025b; Hájek, 2014; Beckers, 2025a). Rather than revisiting these philosophical disputes in full generality, we consider them only insofar as they matter for how counterfactuals are operationalized in contemporary machine learning and adjacent fields.

2 Counterfactuals in Structural Causal Models

Before detailing our criticism in Section 3, we recall some basic concepts. Assumptions about the data-generating process are the essence of causality. SCMs make these assumptions explicit and provide a language for specifying such processes (Pearl, 2012). Rather than giving the full formal definition here, we introduce the key ideas through a simple time-series example in Figure 1. A formal definition is provided in Appendix A.

Structural Causal Models. An SCM specifies a set of system nodes (e.g., predator and prey populations), and typically each system node is accompanied by a noise node (here, u and v for the system nodes F and B, respectively). Each noise node follows a specified probability distribution. The value of each system node is determined by a deterministic updating rule based on its parent nodes. The parent–child relationships are defined by a DAG. To model time-series data, the SCM is unfolded in time via discretization, yielding a separate instance of each variable at every time step (Oberst & Sontag, 2019; Aalaila et al., 2025). A single trajectory is obtained by sampling a realization for each noise node and then iteratively applying the updating rules to determine the values of the system nodes.

Computing Counterfactuals. Given observational data and a specified SCM, counterfactual reasoning asks how the observed outcome would have unfolded under a different hypothetical model (Pearl et al., 2016). In the predator–prey example, this could mean asking what would have happened, had the fox population been set to a different value at a particular time, while keeping all other background conditions unchanged.

In SCMs, counterfactuals are computed using a three-step procedure that crucially relies on knowledge of the underlying SCM. First, the observed trajectory is treated as given and used to infer the latent noise that would have generated it (*abduction*). This can either uniquely identify the values of the noise variables (as in this case) or, more generally, result in a posterior distribution over latent noise values. Next, the model is modified by applying an intervention, such as fixing the fox population at a chosen time (*action*). In our example, this means setting $F_{24} = 5$ and removing incoming edges of the node. Finally, the modified model is simulated forward using the same inferred noise to obtain the hypothetical outcome (*prediction*).

The guiding idea is that counterfactuals describe a minimally altered world: the intervention changes only the targeted mechanism, while all other relevant circumstances are carried over from the factual world.

Ex-Ante vs. Counterfactual Predictions. It is important for the remainder of this manuscript to distinguish between *ex-ante* predictions and counterfactual predictions. *Ex-ante* predictions are forecasts made at a given time using only the information available at that moment (e.g., a doctor choosing a treatment) and are therefore directly relevant for decision-making under uncertainty. Counterfactual predictions, by contrast, are constructed retrospectively and condition on knowledge of how the world actually unfolded after the decision point. In Pearl's framework, this information is encoded in the inferred noise variables.

In other words:

Counterfactuals use information revealed by what did happen to constrain our understanding of what would have happened.

Individual vs. Group-Based Counterfactuals. We are concerned with *individual* counterfactuals, which ask how a specific realized unit would have responded under a different intervention. For example, if a patient dies from COVID-19, asking whether that patient would have survived under better ICU treatment requires fine-grained assumptions about that individual's treatment response and circumstances. A *group-level* counterfactual instead asks how outcomes would have changed across an entire population or subpopulation, where individual variation can be averaged out.Clearly, this distinction is fluid: what is considered a subpopulation in one model may be considered a unit in another. We make it nevertheless to stress that the level of abstraction at which a counterfactual is considered is part of what needs to be justified, and we will argue that such a justification is generally easier to provide at the group-level.

Actual Causality. The preceding distinction between *individual* and *group-level* counterfactuals is closely related to the distinction between *actual* and *general* causality in the philosophical literature (Halpern, 2016). *Actual causality* asks why a specific event occurred in a particular instance (e.g., *Did this cigarette cause this patient's lung cancer?*), whereas *general causality* concerns population-level or structural relationships (e.g., *Smoking increases the risk of lung cancer.*) Counterfactuals lie at the intersection of these notions: they use a general causal model to answer questions about a specific realized case by holding the relevant background circumstances fixed.

3 COUNTERFACTUALS: A CRITIQUE

This section presents the core problems underlying our critique of quantitative counterfactual reasoning. We group these concerns into three broader claims.

3.1 CLAIM A: COUNTERFACTUALS OFTEN ADDRESS THE WRONG QUESTIONS

Our first claim is that counterfactuals are often used where they add little practical value. In many settings, the real goal is to make decisions under uncertainty and judge actions based on the information available at the time. These tasks are often better addressed by forecasts, decision-based reasoning, or group-level causal analysis than by reconstructing individual alternative histories.

A.1 Counterfactuals Add Little Beyond Predictions. Because counterfactual claims lack an empirical sanity check, it is important to justify why they are computed and what they add. In many cases, it is unclear why an individual-level counterfactual trajectory should be more informative than an appropriate *ex-ante* prediction or decision-theoretic quantity available at decision time. A related difficulty is that the decomposition into "structure" and "noise" is often underdetermined by the data (Marx et al., 2020; D'Amour et al., 2022). More broadly, this creates a paradoxical situation: even when a counterfactual can be computed, it is often unclear which question it is meant to answer.

Turning retrospective counterfactuals into forward-looking decision criteria is a general difficulty. Bica et al. (2020), for example, estimate individual counterfactual treatment trajectories for a treatment choice that is a prospective problem. This is clearest for harm, where Mueller & Pearl (2024) propose counterfactual harm as a decision criterion and Sarvet & Stensrud (2025b) show that the best bounds computable from data yield the same decisions as interventional harm.

A.2 Counterfactuals Misjudge Accountability. A patient dies in a hospital and their physician is accused of having chosen the wrong treatment. How should that decision be evaluated?

Petrocelli (2013) argues that counterfactual reasoning is prevalent but often *misused* in this type of medical setting. In particular, after adverse outcomes, they argue that counterfactual comparisons promote outcome bias by focusing on what would have happened under alternative choices rather than on the information available at the time. Counterfactual reasoning can be useful *to a degree*: had the patient passed away under all possible treatment choices, this does speak in favor of exonerating the physician. In general, however, clinical decisions should be judged *ex ante*, based on risk assessments and expected outcomes. In other words, the physician's decision should be evaluated based on the information available to them at the time, not on retrospective counterfactual reconstructions. This argument is also echoed by Dawid & Senn (2023) who argue that appealing to individual-level counterfactuals adds little beyond standard decision-theoretic quantities.

Our claim here concerns the evaluation of the decision, not the separate question of whether the treatment caused the death. The latter is a genuinely counterfactual question, and counterfactual reasoning is the right tool for it, even where the counterfactual cannot realistically be computed. Notions such as harm sit in between, since they combine actual causation with a normative comparison.

Another aspect of accountability is counterfactual fairness, which is concerned with whether individuals would have been treated differently had they belonged to a different social category (Kusner et al., 2017). Chou et al. (2022) argue that many approaches labeled as "counterfactual fairness" are counterfactual only in name, not in a causal sense. In particular, they do not rely on well-defined causal models with explicitly specified background conditions. However, the conflict between *ex-ante* and counterfactual predictions has been discussed in the fairness literature. Zuo et al. (2024) argue that fairness criteria based on retrospective counterfactuals can conflict with forward-looking, decision-time objectives.

A.3 Counterfactuals Create Impossible Worlds. Interventions—that is, modifying an SCM into a hypothetical one—are a crucial step in computing counterfactuals. However, an intervention can break coherence: it may impose a change without specifying the upstream adjustments needed for the hypothetical world to make sense.

Consider, for instance, a case where Alice won an election against Bob and one later asks, "What would have happened if Bob had won?" Simply declaring Bob the winner produces an inconsistent (or at least underspecified) world state. Given the observed votes, preferences, and campaign dynamics, Alice's victory was itself a consequence of the causal state of the system, and the consequences of Bob's hypothetical victory might strongly depend on how exactly this victory was achieved. Meaningful causal questions can only be asked if the concerned alternative scenario is sufficiently specific.

Likewise, Kasirzadeh & Smart (2021) argue that incoherent world states are a problem in ethical machine learning. Many fairness methods counterfactually manipulate traits such as race or gender, implicitly assuming these attributes can be changed while all other aspects of a person remain fixed, which is socially unrealistic. Hu & Kohler-Hausmann (2020) go further and argue that such traits are too tightly bound up with the outcomes to be treated as separate causes that could be changed on their own, so there is no clear target for the intervention. This shows that when an intervention creates an incoherent hypothetical world, the resulting counterfactual may answer no meaningful practical question at all. A partial remedy is offered by *backtracking* counterfactuals (Von Kügelgen et al., 2023), which trace the change back to different background conditions rather than imposing it by intervention, and so keep the hypothetical world coherent.

3.2 Claim B: Counterfactuals Rely on Strong Assumptions That Are Rarely Justified

Our second claim is that individual counterfactuals require much stronger assumptions than ordinary prediction or average effect estimation. These assumptions are often difficult to justify empirically, especially for individual-level conclusions. We highlight four recurring issues.

B.1 Counterfactuals Raise the Burden of Justification. Classical predictive models can be evaluated empirically and derive much of their utility from the fact that their predictions can be tested. In practice, modeling choices are often justified in hindsight by the ability of a model to make accurate predictions about observable phenomena, such as weather or stock markets. Counterfactual statements are empirically untestable by definition, which was the core of Dawid's fundamental criticism (Dawid, 2000). While we do not subscribe to the viewpoint that their untestability makes counterfactuals meaningless *in principle*, the lack of empirical testing substantially raises the bar for what is considered an adequate justification of assumptions. This bar is raised even further when the domain of application is sensitive or when the claims of a study are highly consequential.

B.2 Counterfactuals Rely on Unclear Noise Semantics. A central difficulty in counterfactual reasoning arises from how *noise* is understood, interpreted, and ultimately used. In SCMs, noise is intended to represent "everything outside the model", but this quickly turns it into a vague and underspecified construct. Noise conveniently bundles together many distinct influences: unmodeled variables, unknown factors, physical randomness, measurement error, and approximations. While such aggregation is usually harmless for predictive tasks, it becomes a fundamental problem for counterfactual reasoning.

Some of our most common assumptions for modeling noise are the way they are because we are usually not interested in properties of individual samples. This is best illustrated by the common assumption of Gaussian noise distributions. Noise terms are often assumed to follow a normal distribution because we believe that there are many small independent factors influencing our random variables and that, *on average*, these are approximated well by a Gaussian distribution due to the central limit theorem (see also **C.1** and Figure 2). However, this is no longer true on the individual level. For individual outcomes, a large number of unmeasured factors that were all folded into one single Gaussian noise term might have mattered. Or, in other words, the counterfactual outcomes of two patients with an identical noise value may ultimately be very different.

This issue is especially acute in neural counterfactual models. In deep structural causal models (Pawlowski et al., 2020), abduction infers a latent vector with a normalizing flow, and counterfactuals are generated by changing causal inputs while holding this vector fixed. Yet the flow is optimized for distributional fit, not for separating stable individual attributes from scanner artifacts or transient physiological state. Holding the vector fixed therefore preserves a learned manifold position, not necessarily a causally meaningful description of the individual.

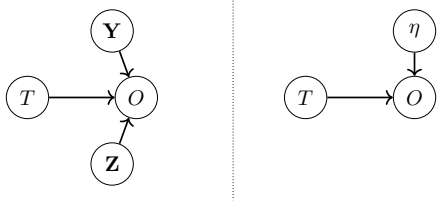

Figure 2: Two models of the effect of treatment dosage (T) on cancer cell count (O). Realistically, the outcome of patient i depends on many (often unobserved or even unknown) factors $\mathbf{Y} = (Y_1, \ldots, Y_M)$ and $\mathbf{Z} = (Z_1, \ldots, Z_N)$ that influence both the effect of the treatment and the final outcome $O^i = \lambda(Y_1, \ldots, Y_M) \cdot T + \sum_{k=1}^{N} Z_k$ (**left**).
On the population level, it can be reasonable to ignore these factors and to replace $\lambda(Y_1, \ldots, Y_M)$ by the constant $\mathbb{E}[\lambda(Y_1, \ldots, Y_M)] = \lambda$ and $\sum_{k=1}^{N} Z_k$ by a Gaussian η. This results in a simple linear model $O = \lambda \cdot T + \eta$ from which patient outcomes are sampled (**right**). This model simplification is often unavoidable and does not change average treatment effects. It does change individual counterfactuals.

B.3 Counterfactuals Rely on Unjustified Cross-World Links. The unclear semantics of noise also create ambiguity when noise is used to link the factual and counterfactual worlds. In Pearl-style counterfactuals, noise variables are the sole mechanism for transferring information from the factual to the counterfactual world. If their meaning is unclear, this link becomes problematic.

In particular, reusing noise that represents genuine realized randomness is philosophically problematic. Some noise components encode chance events—such as quantum or microscopic fluctuations—that occurred in the factual world and have no reason to repeat under a different hypothetical choice. Reusing such noise in a counterfactual world implicitly assumes that inherently random processes would unfold in exactly the same way after the decision is changed (Dawid, 2000).

An illustration of this is given by Mlodozeniec et al. (2025). Consider a patient who received a drug and ask what would have happened had they taken a different drug instead. Some noise variables are reasonable to share across worlds because they encode person-specific attributes such as age, weight, or general health, which are expected to remain the same for the same individual. Other noise variables represent realization-specific factors that should not be shared like manufacturing variability in the active substance of a drug tablet. In the factual world, a patient may have taken a tablet whose actual dose slightly deviates from the advertised amount. Reusing the same noise when considering a different tablet implicitly assumes that the new tablet would have the same manufacturing deviation.

Thus, in many cases, carrying over the exact noise from the factual to the counterfactual world does not preserve meaningful background conditions but instead creates an artificial link between unrelated physical realizations. Likewise, standard SCM practice often folds qualitatively different sources of variation into a single noise term and counterfactual rollouts implicitly choose which of these components are carried across worlds, but this choice is rarely made explicit or justified. In philosophy, Hájek (2014) pushes this line further and argues that genuine indeterminism leaves most "would" counterfactuals outright false, since any chance of a different outcome means it "might not" have happened; fixing the noise makes such claims determinate by construction rather than by evidence.

This issue becomes especially acute for counterfactual trajectories over time. In flow-based counterfactual time-series models such as Wu et al. (2025a), latent noise variables are inferred from observed trajectories and reused under modified interventions. These learned encodings have no clear correspondence to real-world factors, and the model does not separate patient-specific attributes from realization-specific randomness. As a result, it is unclear which aspects of the factual trajectory are carried into the counterfactual world, or whether this is defensible.

B.4 Counterfactuals Rely on Fragile Modularity Assumptions. Even when an intervention is conceptually coherent, standard counterfactual reasoning typically relies on an *invariance* (modularity) assumption: background sources of randomness and the interpretation of noise remain un-

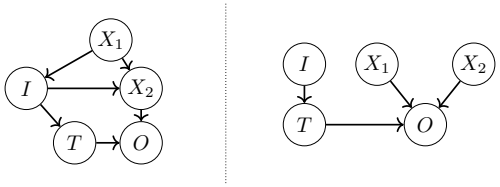

Figure 3: An intervention on treatment T modeled with an intervention node I. Realistically, the intervention I will depend on enabling background conditions X_1 and might have small effects on other covariates X_2 that affect individual outcomes (**left**) and thus counterfactuals. In the population mean (**right**), deleting the edges $X_1 \to I$ and $I \to X_2$ may be justified by the fact that the unintended effects of the interventions are close to 0 on average.

changed when a variable is forcibly modified. This is the do-intervention idealization, often described as a "minimal surgery" that sets one variable while leaving the remaining structural equations intact, and is modeled after idealized randomized controlled trials in which treatment assignment is exogenous. Cartwright (2007) argues that real-world mechanisms rarely satisfy this strong modularity assumption: intervening on one variable often disrupts the surrounding physical or social context that generates uncertainty. Along similar lines, Levy (2015) cautions that even "surgical" interventions can trigger unintended ripple effects, especially as causal chains lengthen.

In our view, domain researchers are often aware that do-interventions provide an idealized picture of real-world actions, particularly at the individual level, even when they produce a coherent hypothetical world overall (unlike the case in **A.3**). Medical drugs may have rare side effects, gene perturbations may have off-target effects, and policy changes may alter behavior; moreover, background conditions enabling the intervention can also affect other covariates (cf. Figure 3). Such deviations may be negligible *on average* for group-level estimands, but they become consequential when inferring *individual* counterfactual trajectories.

3.3 Claim C: Counterfactuals Are Highly Sensitive to Modeling Choices

Our third claim is that counterfactual outputs often reflect modeling choices as much as the underlying phenomenon. Assumptions that have little effect on prediction or average effects can substantially alter individual counterfactual trajectories. As a result, different yet equally plausible modeling choices can lead to drastically different conclusions.

C.1 Counterfactuals Depend on Models Built for Other Goals. A trade-off that any modeler has to deal with is that between *internal validity* (the accuracy of the model for the available samples) and *external validity* (the ability of the model to generalize beyond the modeled samples). Many practices in ML research such as sample splits into training, validation, and test data are explicitly designed to avoid overfitting to the available samples and hence guarantee sufficient external validity. Modeling choices such as linearity of mechanisms and additivity of noise terms are usually not made because they are believed to be true down to the individual level. Consider, for instance, a model for the effect of the treatment dosage of a drug on a cancer cell count. Modelers are usually aware that a multitude of unmeasured (e.g., genetic) factors have an effect on the treatment response for individual patients and that the mechanism by which the treatment affects the response is a complicated function of all these factors (even ignoring the issue of hidden confounding for now). The justification for modeling this problem with a simple linear equation is that such a model is considered a good approximation of the effect *on average*, not for the individual, and that a linear model will generalize better beyond our, perhaps small, cohort of patients (cf. Figure 2).

A similar tension appears in counterfactual regression methods such as CFRNet (Shalit et al., 2017). CFRNet estimates individual treatment effects by learning representations that both predict factual outcomes and reduce imbalance between treated and control groups. This is a sound response to selection bias, but it also exposes the trade-off at issue here: the representation is optimized for factual prediction and group-level comparability, not for preserving all individual-specific variation relevant to a single counterfactual trajectory. Indeed, Shalit et al. note that smoothing or regularization choices helpful for average treatment effects may harm individual treatment-effect estimation.

Thus, even models explicitly designed for individual causal prediction rely on objectives whose regularization structure can conflict with single-unit counterfactual fidelity.

Similarly, machine learning models or loss functions are typically designed in ways that avoid overfitting, for instance, by using regularization. While this is good practice for prediction (we want to predict as well as possible *on average*) and effect estimation (we want an *average* effect over some group of patients), it makes counterfactual estimation problematic: we may have regularized away details that matter on the individual level. We argue that this issue can be mediated to an extent by shifting the focus of counterfactual modeling away from individual trajectories and towards group-level properties. Rather than asking *would patient X have survived if...?*, counterfactual modeling should focus on questions such as *would more patients within this group have survived if...?* that are more aligned with our modeling choices and less prone to individual particularities.

C.2 Counterfactuals Require Detailed Specification. Causal models are, by design, abstractions of the real world. This abstraction is a strength when reasoning about causal structure, as it avoids the need for full physical simulation. However, when computing counterfactuals, abstraction may become a liability.

Peters et al. (2017, Example 6.19) illustrate that distinct SCMs can share the same graph, observational distribution, and interventional distributions while yielding different counterfactuals. Thus, observable and interventional agreement does not determine counterfactual behavior. This point is foundational to Pearl's own framework (Pearl, 2009, Ch. 7 and 9) and is now formalized as the *Causal Hierarchy Theorem* (Bareinboim et al., 2022), according to which the layers of the hierarchy almost never collapse.

The same issue appears in machine-learning settings, and the standard way of dealing with it is itself part of the problem. To resolve the ambiguity for sequential decision processes, Oberst & Sontag (2019) add an extra requirement, *counterfactual stability*, which the Gumbel-Max SCM provably satisfies. In Pearl's framework the causal model is the basic object and counterfactuals follow from it, but here the model leaves them open and an extra assumption is brought in to fill the gap. That assumption does not follow from the causal model, and it is not the only one that would work: Lorberbom et al. (2021) show in a sepsis simulator that a whole family of valid counterfactual mechanisms exists, that they change treatment-effect variance, and that they can even be optimized. Gumbel-Max is therefore better seen as a modeling choice than as a result of the framework, and much later work has adopted it without questioning the assumption behind it. For instance, Beckers (2025b) shows that recent methods for generating counterfactuals from language models (e.g., (Chatzi et al., 2024)) build on the same construction, and that the Gumbel-Max SCM itself imposes further commitments that go beyond counterfactual stability and cannot be traced back to anything observed.

For temporal point processes, Noorbakhsh & Rodriguez (2022) adopt one particular simulation scheme as the basis for generating counterfactual trajectories. While this choice is plausible, Großmann et al. (2024) show that equally defensible design choices can still produce fundamentally different counterfactual predictions. More broadly, Soto (2023) notes that many natural-language counterfactuals do not map uniquely onto interventions or model variables. The same ambiguity arises one level earlier, in the choice of variables itself. A macro-level intervention can often be realized in many different micro-level ways (Spirtes & Scheines, 2004), and equally defensible variable definitions can support different causal structures (Eberhardt, 2016); such realizations may agree interventionally while differing counterfactually, an issue studied formally under causal abstraction (Rubenstein et al., 2017; Beckers & Halpern, 2019). At the population level these differences average out, but an individual counterfactual conditions on which realization actually occurred, so here they do not.

C.3 Counterfactuals Amplify Epistemic Uncertainty. In practice, the SCM is rarely known: both structure and parameters are typically inferred from finite, noisy data. This induces epistemic uncertainty that is often modest for prediction yet consequential for counterfactual rollouts, because small deviations in learned mechanisms can produce large differences in counterfactual trajectories. Our claim here concerns this amplification, not the size of the estimation error itself. Aalaila et al. (2025) show that even when parameters are inferred with high accuracy and sampled from narrow posteriors centered around the true values, counterfactual trajectories can diverge rapidly

under minimal interventions while predictions remain accurate. Estimation error does shrink as more data become available, whereas the counterfactual non-identifiability does not (Nasr-Esfahany et al., 2023; Galhotra & Halpern, 2024), so a residual difference between plausible models remains in either case. Counterfactual conclusions can therefore be highly sensitive to differences between models that are otherwise hard to distinguish.

C.4 Counterfactuals Enter Unstable Regions. Even with good knowledge of the model parameters, an additional problem remains. As first emphasized by King & Zeng (2006), counterfactual interventions may move the system beyond the support of observed data, forcing extrapolation into weakly constrained regions. In dynamical systems, this is especially severe because interventions can shift the system away from the regime in which it was observed and calibrated. A model that reproduces the observed trajectory well may still generate counterfactual rollouts in unstable or chaotic regions, where small perturbations are rapidly amplified. As a result, minor changes in interventions or parameter values can yield qualitatively different outcomes. Standard validation on observed data does not guarantee that intervened trajectories remain in a stable, empirically supported regime. As shown by Aalaila et al. (2025), minimal interventions in the Lorenz and Rössler systems can induce rapid divergence of counterfactual trajectories while prediction remains accurate. In such regimes, observed data provide little guidance, and modeling choices become the dominant determinant of counterfactual outcomes. In neural counterfactual models, this concern is harder to diagnose. Seedat et al. (2022) generate counterfactual outcomes by evolving learned neural controlled differential equations under hypothetical treatment schedules. The resulting trajectory is governed by learned vector fields, but the paper does not adequately justify the stability or support of rollouts outside the observed treatment regime.

4 Practical Recommendations

Before computing a counterfactual, researchers should first ask whether the question actually requires one. For forecasting, policy selection, or prospective decisions, *ex-ante* prediction, expected utility, or population-level causal effects may be more appropriate. When individual trajectories are weakly identified or highly sensitive, group-level conclusions are often more robust and more useful. The assumptions behind the counterfactual should be stated clearly, including assumptions about causal structure, interventions, invariance, and latent variables. If latent noise, hidden states, or learned representations are shared across factual and counterfactual worlds, it should be explained what they represent and why they should remain fixed. In temporal settings, conclusions should also be checked for sensitivity to small changes, unstable dynamics, and extrapolation beyond observed data. Finally, although individual counterfactuals cannot be observed directly, parts of the underlying assumptions can sometimes be tested using interventional data, natural experiments, held-out environments, or later regime changes.

5 Conclusion

Returning to our motivating question—*when are counterfactuals actually useful?*—we have argued that counterfactual reasoning places substantially stronger demands on modeling assumptions than prediction alone. While many scientific models remain useful despite being approximate or even incorrect, this tolerance rarely extends to counterfactuals, whose interpretation depends on detailed and often untestable assumptions about causal mechanisms, noise, and interventions.

These limitations already arise for static counterfactuals, intensify in time-series settings, and become particularly acute when counterfactual trajectories are generated by neural networks, where opacity, underspecification, and instability can outweigh their intended explanatory value.

At the same time, counterfactuals are not without merit: in narrowly defined and carefully justified settings—such as model diagnostics for algorithmic decision-making tools—their use can be appropriate. Our take-away message is therefore one of caution: counterfactuals are powerful modeling constructs, but their practical usefulness is often more fragile than is implicitly assumed in practice.

AI USE STATEMENT

We used generative AI tools to improve wording and readability and to cross-check our literature search for potentially overlooked relevant work. All suggestions were reviewed by the authors, who take responsibility for the final content.

ETHICS STATEMENT

Counterfactual methods are increasingly used in high-stakes settings, including medical decision-making and fairness evaluation involving protected attributes. This work aims to make researchers and practitioners more aware of the assumptions and risks involved in such uses.

REFERENCES

Yahya Aalaila, Gerrit Großmann, Sumantrak Mukherjee, Jonas Wahl, and Sebastian Vollmer. When counterfactual reasoning fails: Chaos and real-world complexity. *arXiv preprint arXiv:2503.23820*, 2025.

Farbod Alinezhad, Jianfei Cao, Gary J Young, and Brady Post. Causal diffusion models for counterfactual outcome distributions in longitudinal data. *arXiv preprint arXiv:2604.12992*, 2026.

Jose Manuel Alvarez and Salvatore Ruggieri. Counterfactual situation testing: Uncovering discrimination under fairness given the difference. In *Proceedings of the 3rd ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization*, pp. 1–11, 2023.

Holger Andreas, Matthias Armgardt, and Mario Gunther. Counterfactuals for causal responsibility in legal contexts. *Artificial intelligence and law*, 31(1):115–132, 2023.

Elias Bareinboim, Juan D Correa, Duligur Ibeling, and Thomas Icard. On pearl's hierarchy and the foundations of causal inference. In *Probabilistic and causal inference: the works of judea pearl*, pp. 507–556. 2022.

Hillel J Bavli. Cause and effect in antidiscrimination law. *Iowa L. Rev.*, 106:483, 2020.

Sander Beckers. Causal counterfactuals reconsidered. *arXiv preprint arXiv:2512.12804*, 2025a.

Sander Beckers. Large language models as nondeterministic causal models. *arXiv preprint arXiv:2509.22297*, 2025b.

Sander Beckers and Joseph Y Halpern. Abstracting causal models. In *Proceedings of the aaai conference on artificial intelligence*, volume 33, pp. 2678–2685, 2019.

Fanny Bergström, Felix Günther, and Tom Britton. A counterfactual analysis quantifying the covid-19 vaccination impact in sweden. *Vaccine*, 52:126870, 2025.

Ioana Bica, Ahmed M Alaa, James Jordon, and Mihaela van der Schaar. Estimating counterfactual treatment outcomes over time through adversarially balanced representations. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=BJg866NFvB.

Alycia N Carey and Xintao Wu. The causal fairness field guide: Perspectives from social and formal sciences. *Frontiers in big Data*, 5:892837, 2022.

Nancy Cartwright. *Hunting causes and using them: Approaches in philosophy and economics*. Cambridge University Press, 2007.

Ivi Chatzi, Nina Corvelo Benz, Eleni Straitouri, Stratis Tsirtsis, and Manuel Gomez-Rodriguez. Counterfactual token generation in large language models. *arXiv preprint arXiv:2409.17027*, 2024.

Yu-Liang Chou, Catarina Moreira, Peter Bruza, Chun Ouyang, and Joaquim Jorge. Counterfactuals and causability in explainable artificial intelligence: Theory, algorithms, and applications. *Information Fusion*, 81:59–83, 2022.

Alexander D'Amour, Katherine Heller, Dan Moldovan, Ben Adlam, Babak Alipanahi, Alex Beutel, Christina Chen, Jonathan Deaton, Jacob Eisenstein, Matthew D Hoffman, et al. Underspecification presents challenges for credibility in modern machine learning. *Journal of Machine Learning Research*, 23(226):1–61, 2022.

A. P. Dawid. Causal inference without counterfactuals. *Journal of the American Statistical Association*, 95(452):407–424, 2000. doi: 10.1080/01621459.2000.10474210.

A Philip Dawid and Stephen Senn. Personalised decision-making without counterfactuals. *arXiv preprint arXiv:2301.11976*, 2023.

Edward De Brouwer. Deep counterfactual estimation with categorical background variables. *Advances in Neural Information Processing Systems*, 35:35213–35225, 2022.

Frederick Eberhardt. Green and grue causal variables. *Synthese*, 193(4):1029–1046, 2016.

Sainyam Galhotra and Joseph Y Halpern. Intervention and conditioning in causal bayesian networks. *Advances in Neural Information Processing Systems*, 37:89019–89041, 2024.

Tomas Garriga, Gerard Sanz, Eduard Serrahima de Cambra, and Axel Brando. A novel application of SCMs to time series counterfactual estimation in the pharmaceutical industry. In *NeurIPS 2024 Causal Representation Learning Workshop*, 2024. URL https://openreview.net/forum?id=zh8WGoQYRt.

Tomas Garriga, Gerard Sanz, Eduard Serrahima de Cambra, and Axel Brando. CEPAE: Conditional entropy-penalized autoencoders for time series counterfactuals. *Transactions on Machine Learning Research*, 2026. ISSN 2835-8856. URL https://openreview.net/forum?id=X6lrzqOtQo.

Gerrit Großmann, Sumantrak Mukherjee, and Sebastian Vollmer. Peculiarities of counterfactual point process generation. In *Proceedings of the 1st ACM SIGSPATIAL International Workshop on Spatiotemporal Causal Analysis*, pp. 11–22, 2024.

Alan Hájek. Most counterfactuals are false. 2014.

Joseph Y. Halpern. *Actual Causality*. MIT Press, 2016. ISBN 0262035022, 9780262035026.

Konstantin Hess and Stefan Feuerriegel. Stabilized neural prediction of potential outcomes in continuous time. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=aN57tSd5Us.

Çağlar Hızlı, S. T. John, Anne Tuulikki Juuti, Tuure Tapani Saarinen, Kirsi Hannele Pietiläinen, and Pekka Marttinen. Joint point process model for counterfactual treatment-outcome trajectories under policy interventions. In *NeurIPS 2022 Workshop on Learning from Time Series for Health*, 2022. URL https://openreview.net/forum?id=G-vMO8qUts.

Richard James Holleyman, Sharmani Barnard, Clarissa Bauer-Staeb, Andrew Hughes, Samantha Dunn, Sebastian Fox, John N Newton, Justine Fitzpatrick, Zachary Waller, David John Deehan, et al. Adjusting expected deaths for mortality displacement during the covid-19 pandemic: a model based counterfactual approach at the level of individuals. *BMC medical research methodology*, 23(1):241, 2023.

Lily Hu and Issa Kohler-Hausmann. What's sex got to do with fair machine learning? *arXiv preprint arXiv:2006.01770*, 2020.

Zilin Jing, Chao Yang, and Shuang Li. Counterfactual optimization of treatment policies based on temporal point process. In *ICML 3rd Workshop on Interpretable Machine Learning in Healthcare (IMLH)*, 2023. URL https://openreview.net/forum?id=MJq0lr8EJ4.

Atoosa Kasirzadeh and Andrew Smart. The use and misuse of counterfactuals in ethical machine learning. In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*, pp. 228–236, 2021.

Gary King and Langche Zeng. The dangers of extreme counterfactuals. *Political analysis*, 14(2): 131–159, 2006.

Matt J Kusner, Joshua Loftus, Chris Russell, and Ricardo Silva. Counterfactual fairness. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL https://proceedings.neurips.cc/paper_files/paper/2017/file/a486cd07e4ac3d270571622f4f316ec5-Paper.pdf.

Jack S Levy. Counterfactuals, causal inference, and historical analysis. *Security Studies*, 24(3): 378–402, 2015.

Jiangmeng Li, Bin Qin, Qirui Ji, Yi Li, Wenwen Qiang, Jianwen Cao, and Fanjiang Xu. Teleporter theory: A general and simple approach for modeling cross-world counterfactual causality. *arXiv preprint arXiv:2406.11501*, 2024.

Bryan Lim. Forecasting treatment responses over time using recurrent marginal structural networks. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018. URL https://proceedings.neurips.cc/paper_files/paper/2018/file/56e6a93212e4482d99c84a639d254b67-Paper.pdf.

Guy Lorberbom, Daniel D Johnson, Chris J Maddison, Daniel Tarlow, and Tamir Hazan. Learning generalized gumbel-max causal mechanisms. *Advances in Neural Information Processing Systems*, 34:26792–26803, 2021.

Aurghya Maiti, Drago Plecko, and Elias Bareinboim. Counterfactual identification under monotonicity constraints. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 26841–26850, 2025.

Charles Marx, Flavio Calmon, and Berk Ustun. Predictive multiplicity in classification. In *International conference on machine learning*, pp. 6765–6774. PMLR, 2020.

Valentyn Melnychuk, Dennis Frauen, and Stefan Feuerriegel. Causal transformer for estimating counterfactual outcomes. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato (eds.), *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 15293–15329. PMLR, 17–23 Jul 2022. URL https://proceedings.mlr.press/v162/melnychuk22a.html.

Bruno Kacper Mlodozeniec, David Krueger, and Richard E Turner. Position: Probabilistic modelling is sufficient for causal inference. In *Forty-second International Conference on Machine Learning Position Paper Track*, 2025.

Scott Mueller and Judea Pearl. Perspective on 'harm'in personalized medicine–an alternative perspective. *American Journal of Epidemiology*, 2024.

Arash Nasr-Esfahany, Mohammad Alizadeh, and Devavrat Shah. Counterfactual identifiability of bijective causal models. In *International conference on machine learning*, pp. 25733–25754. PMLR, 2023.

Kimia Noorbakhsh and Manuel Rodriguez. Counterfactual temporal point processes. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh (eds.), *Advances in Neural Information Processing Systems*, volume 35, pp. 24810–24823. Curran Associates, Inc., 2022.

Michael Oberst and David Sontag. Counterfactual off-policy evaluation with gumbel-max structural causal models. In *International Conference on Machine Learning*, pp. 4881–4890. PMLR, 2019.

Nick Pawlowski, Daniel Coelho de Castro, and Ben Glocker. Deep structural causal models for tractable counterfactual inference. *Advances in neural information processing systems*, 33:857–869, 2020.

Judea Pearl. Causal inference without counterfactuals: Comment. *Journal of the American Statistical Association*, 95(450):428–431, 2000.

Judea Pearl. *Causality*. Cambridge university press, 2009.

Judea Pearl. The do-calculus revisited. *arXiv preprint arXiv:1210.4852*, 2012.

Judea Pearl, Madelyn Glymour, and Nicholas P Jewell. *Causal inference in statistics: A primer*. John Wiley & Sons, 2016.

Jonas Peters, Dominik Janzing, and Bernhard Schölkopf. *Elements of causal inference: foundations and learning algorithms*. The MIT press, 2017.

John V Petrocelli. Pitfalls of counterfactual thinking in medical practice: Preventing errors by using more functional reference points. *Journal of Public Health Research*, 2(3):jphr–2013, 2013.

Jonathan Richens, Rory Beard, and Daniel H. Thompson. Counterfactual harm. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh (eds.), *Advances in Neural Information Processing Systems*, volume 35, pp. 36350–36365. Curran Associates, Inc., 2022.

James M Robins and Sander Greenland. Causal inference without counterfactuals: comment. *Journal of the American Statistical Association*, 95(450):431–435, 2000.

Paul K Rubenstein, Sebastian Weichwald, Stephan Bongers, Joris M Mooij, Dominik Janzing, Moritz Grosse-Wentrup, and Bernhard Schölkopf. Causal consistency of structural equation models. *arXiv preprint arXiv:1707.00819*, 2017.

Donald B Rubin. Causal inference without counterfactuals: comment. *Journal of the American Statistical Association*, 95(450):435–438, 2000.

Aaron L Sarvet and Mats J Stensrud. Perspectives on "harm" in personalized medicine. *American Journal of Epidemiology*, 194(6):1743–1748, 2025a.

Aaron L Sarvet and Mats J Stensrud. Rejoinder to "'harm'in personalized medicine—an alternative perspective". *American Journal of Epidemiology*, 194(6):1752–1755, 2025b.

Nabeel Seedat, Fergus Imrie, Alexis Bellot, Zhaozhi Qian, and Mihaela van der Schaar. Continuous-time modeling of counterfactual outcomes using neural controlled differential equations. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato (eds.), *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 19497–19521. PMLR, 17–23 Jul 2022. URL https://proceedings.mlr.press/v162/seedat22b.html.

Uri Shalit, Fredrik D Johansson, and David Sontag. Estimating individual treatment effect: generalization bounds and algorithms. In *International conference on machine learning*, pp. 3076–3085. PMLR, 2017.

Martín Soto. Why are counterfactuals elusive? https://www.lesswrong.com/posts/gwdwukkc8NfpyPitw/why-are-counterfactuals-elusive-2, March 2023. Accessed: December 21, 2025.

Peter Spirtes and Richard Scheines. Causal inference of ambiguous manipulations. *Philosophy of Science*, 71(5):833–845, 2004.

William Starr. Counterfactuals. In Edward N. Zalta (ed.), *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, fall 2019 edition, 2019. URL https://plato.stanford.edu/entries/counterfactuals/.

Eleni Straitouri, Suhas Thejaswi, and Manuel Gomez Rodriguez. Controlling counterfactual harm in decision support systems based on prediction sets. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang (eds.), *Advances in Neural Information Processing Systems*, volume 37, pp. 129443–129479. Curran Associates, Inc., 2024. doi: 10.52202/079017-4113.

Stelios Triantafyllou, Adish Singla, and Goran Radanovic. Actual causality and responsibility attribution in decentralized partially observable markov decision processes. In *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society*, pp. 739–752, 2022.

Sean Vaskov, Wilko Schwarting, and Chris Baker. Do no harm: A counterfactual approach to safe reinforcement learning. In Alessandro Abate, Mark Cannon, Kostas Margellos, and Antonis Papachristodoulou (eds.), *Proceedings of the 6th Annual Learning for Dynamics & Control Conference*, volume 242 of *Proceedings of Machine Learning Research*, pp. 1675–1687. PMLR, 15–17 Jul 2024. URL https://proceedings.mlr.press/v242/vaskov24a.html.

Julius Von Kügelgen, Abdirisak Mohamed, and Sander Beckers. Backtracking counterfactuals. In *Conference on Causal Learning and Reasoning*, pp. 177–196. PMLR, 2023.

Dongze Wu, Feng Qiu, and Yao Xie. Doflow: Causal generative flows for interventional and counterfactual time-series prediction. *arXiv preprint arXiv:2511.02137*, 2025a.

Peng Wu, Haoxuan Li, Chunyuan Zheng, Yan Zeng, Jiawei Chen, Yang Liu, Ruocheng Guo, and Kun Zhang. Learning counterfactual outcomes under rank preservation. *arXiv preprint arXiv:2502.06398*, 2025b.

Shenghao Wu, Wenbin Zhou, Minshuo Chen, and Shixiang Zhu. Counterfactual generative models for time-varying treatments. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pp. 3402–3413, 2024.

Zhiqun Zuo, Tian Xie, Xuwei Tan, Xueru Zhang, and Mohammad Mahdi Khalili. Lookahead counterfactual fairness. *arXiv preprint arXiv:2412.01065*, 2024.

A Formal Definition of Structural Causal Models

We briefly recall the standard definition of a Structural Causal Model (SCM) following Pearl (2009).

An SCM is a triple $\mathcal{M} = (\mathbf{U}, \mathbf{V}, \mathbf{F})$, where:

- $\mathbf{U}$ is a set of exogenous (noise) variables that represent external or unmodeled influences,
- $\mathbf{V}$ is a set of endogenous (system) variables whose values are determined within the model,
- $\mathbf{F}$ is a collection of structural functions $\{f_i\}$, one for each variable $V_i \in \mathbf{V}$.

Each structural function specifies how a system variable is generated:

$$V_i = f_i(\mathrm{Pa}(V_i), U_i),$$

where $\mathrm{Pa}(V_i) \subseteq \mathbf{V}$ denotes the set of parent variables of V_i, and $U_i \in \mathbf{U}$ is the associated noise variable. The parent relationships implicitly define a directed acyclic graph (DAG) over the system variables.

The noise variables are jointly distributed according to some distribution $P(\mathbf{U})$. Together, the structural functions and noise distribution induce a joint distribution over $\mathbf{V}$.

Interventions are modeled by modifying one or more structural functions, for example replacing f_i with a constant assignment. Counterfactuals are defined by combining such interventions with inferred realizations of the noise variables, allowing one to compute hypothetical outcomes conditional on observed data.

A.1 Notes on Monotonicity

It is often considered desirable for counterfactual samples to satisfy *monotonicity* (Oberst & Sontag, 2019). Informally, monotonicity requires that if an intervention on a variable leads to a certain outcome, then a *stronger* intervention in the same direction should not produce a weaker or contradictory outcome. For example, increasing the infectivity of a disease while holding all other factors fixed should not lead to fewer cases.

Many causal relationships are inherently non-monotonic (e.g., physical activity improves health only up to a point). Even when monotonicity appears reasonable, relying on it can be misleading (Maiti et al., 2025). In particular, the assumption that enforcing monotonicity increases the credibility of counterfactual conclusions is generally unwarranted, and doing so can mask deeper model misspecification rather than ensuring meaningful counterfactual behavior.

B Applications of Counterfactual Modeling

Recent work has increasingly operationalized counterfactual reasoning as a mechanism for generating alternative temporal trajectories of complex systems, rather than static outcome contrasts. A prominent line of research develops structural or semi-structural models that enable counterfactual rollouts by inferring latent state or noise variables from observed histories and reusing them under modified interventions. This paradigm appears across deep structural causal models (Pawlowski et al., 2020), flow-based time-series models (Wu et al., 2025a), differential-equation–based dynamics (Seedat et al., 2022), causal transformers (Melnychuk et al., 2022), event-based point-process formulations (Noorbakhsh & Rodriguez, 2022), and autoencoder-based time-series counterfactual models such as CEPAE (Garriga et al., 2026). Related generative approaches produce counterfactual samples or outcome distributions under time-varying treatments, including conditional generative models (Wu et al., 2024) and diffusion-based models for longitudinal counterfactual outcomes (Alinezhad et al., 2026). Across these works, counterfactual trajectories are treated as meaningful alternative evolutions of a given system under modified intervention paths.

A parallel line of work targets static individual counterfactual outcomes rather than trajectories. Wu et al. (2025b) propose identifying an individual's counterfactual outcome as the quantile of the counterfactual outcome distribution that matches the quantile rank of the observed factual outcome in the factual distribution. This requires a rank preservation assumption: that each individual's factual and counterfactual outcomes occupy the same rank in their respective conditional distributions. The authors show this is strictly weaker than standard homogeneity and strict monotonicity assumptions, but it remains untestable from data: whether a specific patient's headache outcome occupies the same quantile rank under medication and non-medication cannot be verified empirically. Brouwer (De Brouwer, 2022) addresses a related identification problem by introducing categorical background variables under which counterfactual estimation can become identifiable; this illustrates how additional assumptions are often introduced to make individual-level counterfactual queries operational.

These counterfactual trajectories are not treated as abstract explanatory artifacts, but are explicitly deployed for decision comparison and policy evaluation. In longitudinal healthcare, sequential counterfactual models are used to assess alternative treatment regimes over time, supporting planning under time-varying confounding and adaptive interventions (Bica et al., 2020; Lim, 2018). Recent continuous-time methods such as SCIP-Net estimate conditional average potential outcomes from irregularly sampled patient trajectories while adjusting for time-varying confounding through stabilized inverse-propensity weighting (Hess & Feuerriegel, 2025). In industrial and economic settings, SCM-based time-series counterfactuals are applied to estimate the impact of discrete events, such as market entry or loss of exclusivity, by reconstructing post-event trajectories under hypothetical non-occurrence (Garriga et al., 2026). Similarly, point-process–based counterfactual models generate alternative event histories under modified policies (Hızlı et al., 2022), and recent extensions use counterfactual temporal point processes for treatment-rule optimization (Jing et al., 2023). These works illustrate a broader methodological trend: counterfactual rollouts are increasingly used not merely for explanation, but as inputs to treatment comparison, policy evaluation, and optimization.

At a broader scale, counterfactual time-series reasoning underpins retrospective population-level analyses in public health and epidemiology. Epidemic and survival models are used to construct counterfactual trajectories representing no-vaccination, delayed-intervention, or alternative-policy scenarios, yielding interpretable quantities such as lives saved or deaths displaced. These analyses are routinely presented as after-action assessments of real-world decisions, despite relying on counterfactual trajectories that cannot be empirically validated once the factual history has unfolded (Bergström et al., 2025; Holleyman et al., 2023).

Counterfactual reasoning is also increasingly embedded in normative and regulatory contexts. Formal notions such as counterfactual harm frame ethical evaluation by comparing realized outcomes to hypothetical alternatives generated by causal models (Richens et al., 2022), while extensions to decision support systems seek to bound or control such quantities using identifiability assumptions (Straitouri et al., 2024). Related ideas appear in safety-oriented sequential decision-making, where counterfactual rollouts are used to constrain behavior relative to baseline policies (Vaskov et al., 2024). Across these settings, counterfactual quantities function as standards against which systems are judged, constrained, or certified.

Finally, parallel advances in causal representation and graphical formalisms aim to systematize cross-world reasoning itself, enabling symbolic manipulation and bookkeeping of counterfactual variables across multiple hypothetical worlds. While such frameworks do not introduce new application domains, they reinforce the growing treatment of counterfactual trajectories as first-class objects within dynamic modeling pipelines (Li et al., 2024).

These works differ in how explicitly they discuss the limitations of their own assumptions. Shalit et al. (2017) note that choices which help average treatment effects can harm individual-level estimates. Oberst & Sontag (2019) state their assumptions clearly, but validate only in a synthetic setting where the model is known, while motivating the method with a real clinical review. Noorbakhsh & Rodriguez (2022) adopt one simulation scheme without discussing that equally defensible schemes yield different counterfactuals, Seedat et al. (2022) do not address whether rollouts stay stable outside the observed treatment regime, and Wu et al. (2025a) reuse learned noise across worlds without separating person-specific attributes from realization-specific randomness. We do not read these as oversights.