In search of "normality": producing relevant counterfactual explanations for complex behaviors

Samuel Reyd, Ada Diaconescu, Jean-Louis Dessalles
Telecom Paris, LTCI, IPParis, Palaiseau, France
name.surname@telecom-paris.fr

***Abstract*—Complex systems often exhibit unexpected behaviors that call for explanation. A classical way to explain such behavior is to identify what would have had to differ from the observed situation for it not to occur. This is called a counterfactual explanation. Many such counterfactual explanations exist for a given situation, and a classical criterion to select among them is to prefer those whose counterfactual situation is more normal. This makes measuring normality a key step for selecting relevant explanations. We propose to use "Anchors", a classic XAI method, to identify low-dimensional normality zones: regions in which the expected behavior reliably holds. Our central contribution is a normality metric that scores a situation by the normalized volume of its normality zone. Our second contribution is a method that assembles such zones into a global model used to guide counterfactual explanation search toward more normal alternatives. We evaluate both contributions on a swarm simulation (agents self-organizing into a coherent moving group), explaining why particular parameter configurations fail to produce this collective behavior. The proposed metric better captures identified properties of normality than standard anomaly-detection baselines. Normality-guided search yields simpler explanations, fewer low-normality counterfactuals, and substantially lower computational cost.**

***Index Terms*—Relevant Counterfactual Explanations, Actual Causation, Normality Metric**

I. Introduction

Complex adaptive systems are large collections of interacting agents whose collective behavior cannot be predicted from the individual behaviors. They routinely produce outcomes that are difficult to anticipate. When such a system departs from its expected behavior, a natural question follows: why did this happen? Answering it is a prerequisite for diagnosing, repairing, or trusting the system.

A common way to answer such a question is to use counterfactual explanations: we explain an outcome by identifying what would have had to differ, in this specific instance, for it not to occur [1], [2], [3]. For instance, consider a pedestrian who never checks for traffic before crossing the road. One day, they stop to tie their shoelace. By coincidence, this delay puts them at the crossing at the exact moment a car passes, and they are hit. An infinite number of interventions could have prevented the accident in this specific situation, but two of them stand out. On the one hand, not having stopped to tie the lace would have made the pedestrian cross a few moments earlier and avoid the car. On the other hand, having checked for traffic before stepping out would have allowed them to let the car pass before crossing. Both interventions prevent the outcome in this particular situation.

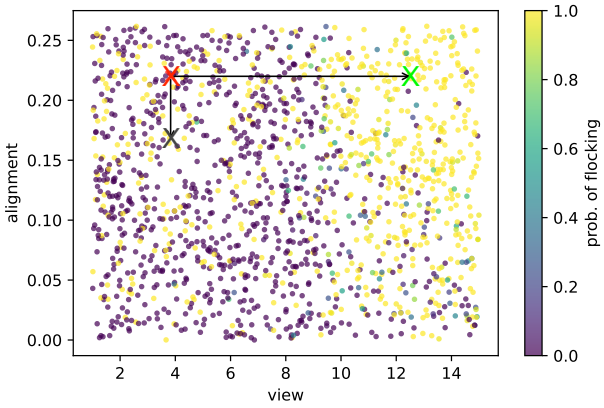
Halpern and Hitchcock [4] propose a framework to grade those two explanations. In their framework, only normal counterfactual situations can produce relevant explanations. Normality can be assessed through several accounts. In our example, not having stopped to tie the lace is the more usual alternative, so someone using a typicality-based account of normality would favor this explanation. However, checking for traffic is what a pedestrian ought to do, so someone using a normative account of normality would prefer this explanation. Halpern and Hitchcock thus incorporate world knowledge to assess the relevance of the explanation, going beyond the purely mechanistic base of counterfactual reasoning. However, they do not propose a general computable way to measure normality. This is the gap we address.

Anomaly detection is the natural computational approach to normality assessment [5], [6]. These methods operationalize normality as robustness. Given a situation represented by a vector that describes it on multiple dimensions, its normality is assessed by whether it sits in a dense, stable region of a reference distribution. We follow this framing. For instance, checking before crossing is a more robust intervention than relying on a timing-based coincidence.

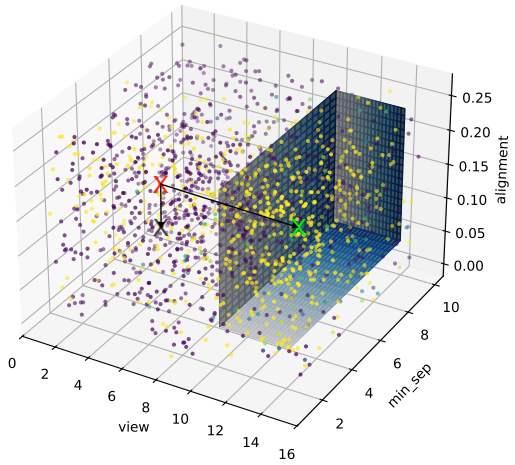
However, anomaly detection methods aggregate information across all available dimensions. Since explanations are ultimately intended for human understanding, a normality measure that diverges from human judgment undermines the relevance of the explanation itself. Norm theory [7] shows that human normality judgments work by comparing a situation to normal alternatives, obtained by varying only a small set of relevant dimensions. Consider judging whether it is normal that a person carries an umbrella: you compare to similar days varying the weather, not the person's age or the color of their coat. Anomaly detection does not make this distinction. This misalignment grows with the number of dimensions. What is needed is robustness restricted to a few relevant dimensions.

This paper addresses this problem in the context of complex systems, where the expected behavior is a collective phenomenon emerging from the interaction of many agents. We use "Anchors" [8], a classic explainability method that delivers the required combination. Given a situation represented by a vector of dimensions, it finds the largest sparse high-precision rule that constrains only a subset of dimensions while leaving the others free. Such a rule delineates a low-dimensional normality zone in which the expected behavior reliably holds.

Figure 1a illustrates the idea in a flocking simulation (a



(a) Example of interventions that restore the expected behavior (flocking). Starting from a point (red cross) in the parameter configuration space where flocking is improbable (purple points), interventions (black arrows) change the system configuration so that flocking occurs (yellow points). The intervention on the "view" parameter, i.e., the perception radius (green cross), is more "robust" than the intervention on the alignment (gray cross), as it creates a counterfactual within a larger region of "flocking" points.



(b) Example of a 3-dimensional "normality zone". It is defined by the rule "view $>$ 11.09 & min_sep. $<$ 8.55 & alignment $>$ 0.07". The surfaces represent these thresholds. At least 95% of the points inside this zone exhibit flocking. It includes the "robust" counterfactual point (green cross).

Fig. 1: Two projections of the space of parameters for a flocking simulation.

collective behavior in which agents following simple local rules self-organize into a coherent moving group). Starting from a configuration that does not produce flocking (red cross), both increasing the perception radius (green) and decreasing the alignment strength (gray) restore flocking. But the former does so within a large normality zone (Figure 1b), whereas the latter sits on a narrow margin. The extent of a normality zone along its constrained dimensions captures how far the relevant dimensions may vary before the behavior is lost: robustness made quantitative, over exactly the dimensions that matter.

Our central contribution is a normality metric that scores a situation by the normalized volume of its normality zone. This metric yields a computable notion of normality that is both cognitively motivated and directly applicable to complex systems. Our second contribution is a method for generating relevant counterfactual explanations under normality constraints. We assemble the normality zones of a set of reference situations into a global model of where the system behaves normally, and use it to steer the search towards counterfactual situations that our metric considers more normal. Normality thus acts as a filter on admissible counterfactuals, in the spirit of Halpern and Hitchcock [4], but in a form directly applicable to complex systems. Our work builds on several existing theories and methods; its specific novelty is the interpretation (and subsequent implementation) of "Anchors" to measure normality and constrain search for counterfactual explanations.

We evaluate the approach on a swarm simulation inspired by Reynolds' model [9], explaining why particular parameter configurations fail to produce flocking. For the normality metric, we test whether it behaves as norm theory prescribes: depending on a few dimensions and remaining insensitive to irrelevant ones. We compare it against six standard anomaly-detection baselines. For explanation generation, we measure the simplicity, normality, and computational cost of the produced explanations. The proposed metric matches the intended properties of normality better than the baselines. Normality-guided search yields fewer and simpler explanations, fewer low-normality counterfactuals, and a substantial reduction in simulation calls, while revealing a tradeoff between normality and minimality. We present these results as an initial validation on a single experimental domain. Section VII discusses the structural conditions under which the approach is expected to transfer. The code for both contributions is publicly available[1].

II. Background and Related Work

This section presents the tools and theories this paper relies on, as well as alternative approaches. We first present how relevant counterfactual explanations can be accounted for. We then focus on the question of normality and how it can be measured in complex systems. Finally, we present the XAI method that constitutes the basis of our methodology.

A. Counterfactual explanations

Explanations of complex-system behaviors can take several forms, including mechanistic accounts based on the organization and interaction of system components [10], or sensitivity- or feature-based accounts that identify the variables most responsible for an outcome [11]. In this paper, we focus on counterfactual explanations.

Counterfactual explanations aim to identify what could be changed in a situation to alter a given outcome. To model

[1]https://github.com/SamuelReyd/CounterfactualNormality

interventions, we adopt the formalism of Structural Causal Models (SCMs) [1], [12], [2]. SCMs represent variables and the causal relations among them, as well as predicates (i.e., boolean functions) that depend on variable values. This section presents a simplified version of this formalism adapted to our experiments; the full formalism is presented in Appendix A.

We consider a set of causal variables $\mathcal{V} = \{X_1, \ldots, X_n\}$, where "causal" means that the variables are eligible for intervention in the explanation process. In our experiments, these variables are the simulation parameters (e.g., perception radius or alignment), but the same framework could also be applied to simulation variables (e.g., the position or direction of specific agents at different time steps). For each variable $X \in V$, let $\mathcal{R}(X)$ denote its range of admissible values. A situation is a complete assignment of values to all variables, represented as a vector $\mathbf{v} \in \mathcal{R}(\mathcal{V})$, where $\mathcal{R}(\mathcal{V}) = \prod_{X \in \mathcal{V}} \mathcal{R}(X)$ denotes the Cartesian product of all individual ranges. The symbol $\mathcal{R}$ denotes either the range of a variable or of a set of variables.

Let φ denote the target behavior (i.e., predicate) to be explained under a particular situation $\mathbf{v}$, e.g., the occurrence of flocking ($\varphi(\mathbf{v}) = 1$) or its absence ($\varphi(\mathbf{v}) = 0$) given the set of parameters $\mathbf{v}$. For any subset of variables $\mathbf{X} \subseteq \mathcal{V}$ and any alternative assignment $\mathbf{x}' \in \mathcal{R}(\mathbf{X})$, we write $\mathbf{v}[\mathbf{X} \leftarrow \mathbf{x}']$ for the counterfactual situation obtained by replacing in $\mathbf{v}$ the values of $\mathbf{X}$ by $\mathbf{x}'$. The corresponding counterfactual behavior is then evaluated through $\varphi(\mathbf{v}[\mathbf{X} \leftarrow \mathbf{x}'])$.

A counterfactual explanation identifies variables whose modification would alter the behavior of interest. We use the following simplified behavior-restoring version of actual causation, inspired by the HP definition [2]: a subset of variables $\mathbf{X} \subseteq \mathcal{V}$ is a cause of $\varphi(\mathbf{v}) = 0$ if two conditions hold. First, there must be values for the cause variables that "cancel" the consequence: $\exists \mathbf{x}' \in \mathcal{R}(\mathbf{X}) : \varphi(\mathbf{v}[\mathbf{X} \leftarrow \mathbf{x}']) = 1$. Second, $\mathbf{X}$ must be minimal in the sense of set inclusion, i.e., no strict subset of $\mathbf{X}$ satisfies the first condition. For instance, in Fig. 1a, the perception radius and alignment parameters represent separate causes for the absence of flocking, as each supports interventions (black arrows) that restore flocking.

Counterfactual explanations are also popular in explainable artificial intelligence (XAI), where causal variables are the inputs of the machine-learning model and the target predicates are its outputs. An explanation specifies how inputs would have to be modified to obtain different outputs [13]. While such methods could apply to our particular case study, they would not extend to cases featuring interactions among causal variables (see Appendix A for details). Therefore, we choose to rely on the actual causation framework instead, as it supports such extensions. While we do not rely on XAI approaches to identify the counterfactuals, we adopt another XAI method ("Anchors", subsec. II-D) to evaluate the normality of counterfactual situations.

B. Counterfactual relevance

The definition of actual causes (cf. above) allows for numerous causes to explain a situation. This is especially problematic in large complex systems, with many interdependent variables and feedback loops [14]. Extending the actual cause definition to support relevance ranking among causes becomes crucial.

Halpern and Hitchcock [4] propose using an explicit normality ordering $\preceq$ to address this issue. In their framework, $\preceq$ is a partial preorder over situations, with $\mathbf{v} \preceq \mathbf{v}'$ meaning that situation $\mathbf{v}' \in \mathcal{R}(\mathcal{V})$ is at least as normal as situation $\mathbf{v} \in \mathcal{R}(\mathcal{V})$. They propose a modified HP definition where the counterfactual situation, $\mathbf{v}' = \mathbf{v}[\mathbf{X} \leftarrow \mathbf{x}']$, has to be at least as normal as the actual situation, $\mathbf{v}$. More generally, candidate causes can be compared based on the normality of their corresponding counterfactual situations [4]: the best cause is the one supported by the most normal counterfactual.

Reyd et al. [15] propose a similar notion of comparative causation based on relevance criteria such as simplicity and recency. They also measured normality via the distance of the counterfactual to the actual situation. Alternatively, Houze et al. [16] measure the relevance of a cause via memorability. Miller [17] review criteria from social science to assess the relevance of explanations in XAI. They notably stressed that causation must be contrastive, which they later formalized using causal models [18].

The actual-causation literature defines normality via informal notions (e.g., conventions, normative expectations, and background assumptions), which makes them difficult to evaluate for comparative purposes. While several works did model such notions more formally in complex social systems [19], [20], [21], they do not specifically address normality measurements. The present paper aims to retain the role of normality emphasized by Halpern and Hitchcock's theory [4] while operationalizing it in a form suitable for abstract, high-dimensional parameter spaces.

C. Modeling and measuring normality

This subsection examines general properties of causal judgment from a cognitive perspective and existing computational proxies.

Norm theory proposes a cognitive account of how people produce judgments of normality in everyday situations. Kahneman and Miller [7] argue that the normality of a situation is evaluated by comparing it to nearby alternatives produced by changing a limited number of relevant factors. When an unwanted event occurs, people immediately imagine counterfactual situations by mutating limited sets of features. They assess the normality of the original event by evaluating how much the situation changed in the imagined alternatives.

The anomaly detection literature has produced numerous computational methods based on similar notions. Here, abnormal points are detected by comparison with nearby points or with a representation of their distribution [5], [6]. These approaches assume the availability of a dataset of reference normal points and provide scores to assess the normality of a query point.

kNN distance [22] estimates the normality of a query point from its distance to its k nearest neighbors in the reference set, with larger distances indicating lower normality. Local Outlier Factor (LOF) [23] compares the local density around a point

to that of its neighbors, so that observations lying in regions that are significantly sparser than their surroundings receive lower normality scores. Gaussian Mixture Model (GMM) [24] represents the distribution of normal observations as a weighted combination of several Gaussian components. The normality of a point is its likelihood under the fitted mixture model. One-class Support Vector Machines (OC-SVM) [25] estimate a decision region containing most of the normal points and quantifies normality through the position of a point relative to this learned boundary. Isolation Forest (IF) [26] characterizes anomalies through their susceptibility to recursive partitioning: normality is computed as the average path length over an ensemble of random trees. Finally, Kernel Density Estimation (KDE) [27] provides an estimate of the distribution of normal observations. Normality is then evaluated through the estimated density at the query point. These methods have also been adapted to address challenges of complex adaptive systems, such as awareness [28] or continual changes [29].

These approaches are relevant to our normality-assessment problem because they provide operational scores over continuous dimensional spaces. Yet, they typically aggregate information across all provided dimensions, whereas norm theory [7] suggests that normality judgments depend on restricted sets of salient variables. Hence, anomaly detection provides only partial proxies for explanatory normality in complex systems.

D. Anchor explanations

This subsection introduces a classic XAI method called "Anchors" [8] that models low-dimensional, high-precision rules in multi-parameter spaces. This method will later be the foundation of our approach to measuring normality and identifying "high-normality" zones.

In XAI, surrogate approaches aim to explain the behavior of a complex machine learning model around a specific input by approximating it locally with a simpler model. LIME [30] is a prominent example that approximates the machine learning model around an input by a sparse linear model. This surrogate model shows that only some dimensions are relevant to the machine learning output. For instance, by approximating an image classification model with linear regression that predicts whether an image contains a frog based on how green some important pixels are, LIME identifies which pixels of the input are relevant. Alternatively, "Anchors" [8] seeks simple rules that constrain the input so as to ensure the output.

Such black-box approaches only consider the input-output behavior of machine learning models. They can easily be adapted to our experiments, which simulate a system given a set of parameters (inputs) and detect a target phenomenon (output).

We now present the "Anchors" method using our notation, where inputs are the causal variables (i.e., simulation parameters) and the output is a Boolean indicating whether the expected phenomenon is observed. Formally, let $\varphi : \mathcal{R}(\mathcal{V}) \to \{0, 1\}$ denote the predictor of the target behavior on parameter settings $\mathbf{v} \in \mathcal{R}(\mathcal{V})$. Given a specific $\mathbf{v}^*$, "Anchors" identifies which values of $\mathbf{v}^*$ are responsible for the observed $\varphi(\mathbf{v}^*)$.

In general, an anchor used to explain the output $\varphi(\mathbf{v}^*)$ is an interpretable predicate (i.e., a Boolean function), or rule, that constrains the input space: $A : \mathcal{R}(\mathcal{V}) \to \{0, 1\}$. The anchor A for $\mathbf{v}^*$ is defined based on the following conditions.

First, the rule must hold for the target input, i.e., $A(\mathbf{v}^*) = 1$.

Second, the precision of the anchor must exceed a certain threshold τ given as a parameter to the method. To define the precision, the method relies on a distribution $\mathcal{D}$ obtained via perturbations of $\mathbf{v}^*$. If $\mathbf{z}$ is a set of samples that follows distribution $\mathcal{D}$ restricted to $A(\mathbf{z}) = 1$, the precision is the proportion that preserves the observed output: $\text{prec}(A) = \mathbb{E}[\mathbb{1}_{\varphi(\mathbf{z})=\varphi(\mathbf{v}^*)}]$, where $\mathbb{1}_{\varphi(\mathbf{z})=\varphi(\mathbf{v}^*)}$ is the indicator function that returns 1 when $\varphi(\mathbf{z}) = \varphi(\mathbf{v}^*)$ and 0 otherwise.

Third, the final anchor is the one that has the maximum coverage. If $\mathbf{x}$ is a set of samples that follows $\mathcal{D}$, the coverage is the proportion that satisfies the anchor: $\text{cov}(A) = \mathbb{E}[A(\mathbf{x})]$.

Overall, the anchor that explains the prediction of φ for input $\mathbf{v}^*$ with precision τ is the one that solves the following optimization problem:

$$\max_{A} \text{cov}(A) \quad \text{s.t.} \quad A(\mathbf{v}^*) = 1 \quad \text{and} \quad \text{prec}(A) \geq \tau \qquad (1)$$

We use the approximate algorithm and its implementation from the original paper [8] to identify the anchor associated with a given situation (i.e., the input to explain). Fig. 1b shows an anchor identified based on the point represented with a green cross, where the rule is based on thresholds that are represented with planes.

We adopt "Anchors" in our proposal as it combines two properties that are directly relevant to our notion of normality. First, anchors are sparse by construction, as they only include a subset of parameters in the rule, while leaving the others unconstrained. Second, they identify regions where modifying a parameter does not alter the expected behavior.

III. Contributions

Our contributions include a normality metric and an explanation generation method. We first present the setup and notation for grounding our contributions in III-A. We then present our first contribution in III-B. Finally, we present our second contribution in two steps: 1) modeling global "normality zones" (III-C); and 2) two alternatives for generating explanations, based on: a) contrast to the nearest zone (III-D1); and b) constraining cause search to the nearest zone (III-D2).

A. Problem setup and notation

We consider a system controlled by a finite set of real-valued parameters $\mathcal{V} = \{X_1, \ldots, X_n\}$. Each parameter $X \in \mathcal{V}$ takes values in a prescribed range $\mathcal{R}(X) = [r_X^{\min}, r_X^{\max}] \subset \mathbb{R}$. A parameter setting (instance) is denoted by $\mathbf{v} \in \mathcal{R}(\mathcal{V})$. Importantly, while this paper focuses on simulation parameters (e.g., attraction or repulsion strengths), our method can also be applied to runtime state variables (e.g., position and velocity of each agent, at each time t).

We assume that we have access to a simulation of the system and a detector of the expected behavior. As the simulation may be stochastic, a parameter setting $\mathbf{v}$ is associated with a probability $p(\mathbf{v}) = \mathbb{P}(\text{expected behavior}|\mathbf{v})$. This induces a binary predictor $\varphi(\mathbf{v}) = \mathbb{1}_{p(\mathbf{v})\geq 1/2}$, which indicates whether the expected behavior is more likely than not under $\mathbf{v}$. The method also presupposes access to a reference sample of points $S^+ \subset \mathcal{R}(\mathcal{V})$ that exhibit the expected behavior.

Our approach is designed to complement an external actual-cause identification procedure (or solver). Given an abnormal point to explain $\mathbf{v}^e$ and a set of intervention domains, such a procedure is supposed to return a set of causes $\mathbf{C} \subseteq \mathcal{V}$ together with a supporting intervention $\mathbf{C} \leftarrow \mathbf{c}$ yielding a behavior-restoring counterfactual $\mathbf{v}^c = \mathbf{v}^e[\mathbf{C} \leftarrow \mathbf{c}]$.

B. Local anchor-based normality score

We first define a normality score $N(\mathbf{v})$ used to rank behavior-restoring counterfactuals $\mathbf{v}^c$. This score is computed based on anchors when $\varphi(\mathbf{v}) = 1$; otherwise, it is null.

Let $\mathbf{v} \in \mathcal{R}(\mathcal{V})$ be a situation such that $\varphi(\mathbf{v}) = 1$. We apply the "Anchors" algorithm to $\mathbf{v}$ using the predictor φ. The output of the algorithm is an anchor A such that $A(\mathbf{v}) = 1$ with a maximal coverage $\text{cov}(A)$ under a minimal authorized precision $\text{prec}(A)$ (i.e., solving Eq. 1, subsec II-D). In our continuous bounded settings, an anchor A is a rule built as a conjunction of threshold predicates on a subset of the variables in $\mathcal{V}$. It can be represented as a Cartesian product of intervals:

$$A = \prod_{X \in \mathcal{V}} [h_X^{\min}, h_X^{\max}] \tag{2}$$

where unconstrained variables are assigned their full range, i.e., $[h_X^{\min}, h_X^{\max}] = [r_X^{\min}, r_X^{\max}]$. Given our interpretation, an anchor A is no longer referred to as a Boolean function but as a subspace of the parameter space, i.e., $A \subseteq \mathcal{R}(\mathcal{V})$. Hence, to write that a situation $\mathbf{v}$ "satisfies" A, we will write $\mathbf{v} \in A$ instead of the previous $A(\mathbf{v}) = 1$.

Importantly, the perturbation distribution $\mathcal{D}$ used for learning the anchors is inherited from the original method [8] and its implementation; it is not a design choice of our work.

To define the normality score $N(\mathbf{v})$, we rely on the anchor's dimension-normalized volume $\rho(A)$, i.e., the geometric mean of the normalized interval widths:

$$\rho(A) = \Big(\prod_{X \in \mathcal{V}} \frac{h_X^{\max} - h_X^{\min}}{r_X^{\max} - r_X^{\min}} \Big)^{\frac{1}{|\mathcal{V}|}} \tag{3}$$

Unconstrained variables contribute a factor equal to 1, whereas constrained variables contribute according to the relative size of their admissible interval. Larger anchors receive higher scores, reflecting the idea that a counterfactual is more normal when the expected behavior is preserved under broader variations of the relevant parameters.

The normality score N for a situation $\mathbf{v} \in \mathcal{R}(\mathcal{V})$ is:

$$N(\mathbf{v}) = \begin{cases} 0 & \text{if} \quad \varphi(\mathbf{v}) = 0 \\ \rho(A) & \text{if} \quad \varphi(\mathbf{v}) = 1 \end{cases} \tag{4}$$

where A is the selected anchor containing $\mathbf{v}$.

Our normality metric intuitively aligns well with the results of norm theory [7]. The constrained dimensions are selected by the anchor algorithm as the ones relevant to the expected phenomenon. The interval sizes then quantify which modifications of the selected dimensions alter the expected phenomenon and which do not. We can notably suppose that this metric only relies on a limited set of dimensions and is insensitive to irrelevant ones, which we empirically verify in Sec. VI-A.

C. Global modeling of high-normality regions

The score defined above evaluates the normality of a single behavior-restoring counterfactual. We now present a methodology to construct a set of high-normality zones that we later use to improve the generation of explanations.

For each $\mathbf{v} \in S^+$, we compute a local anchor satisfying the precision requirement introduced above. Each such anchor is viewed as a candidate local region of high normality.

Because different points may yield identical or overlapping anchors, these candidates are then post-processed. First, duplicate anchors are removed. Second, pairs of anchors are merged whenever their union still satisfies the precision requirement. The result of merging anchors A_1 and A_2 is defined by $A_M = \prod_{X \in \mathcal{V}} [\min(h_{X,1}^{\min}, h_{X,2}^{\min}), \max(h_{X,1}^{\max}, h_{X,2}^{\max})]$, where $h_{X,1}$ and $h_{X,2}$ are the thresholds for A_1 and A_2.

By construction, such a merge can only preserve or increase the volume and coverage of the candidate region; precision is the only criterion that may invalidate it. This merging step is repeated to obtain larger admissible regions. Finally, we only keep maximal anchors. Maximality is measured via a set inclusion ordering: we write $A_1 \subseteq A_2$ when, for every $X \in \mathcal{V} : [h_{X,1}^{\min}, h_{X,1}^{\max}] \subseteq [h_{X,2}^{\min}, h_{X,2}^{\max}]$.

The global normality model is defined as the family $\mathcal{A}$ of maximal anchors, optionally filtered by a minimum volume criterion. Using several anchors rather than a single one is essential in practice, since the regions supporting normal behavior may be disconnected or multi-modal. Therefore, the family $\mathcal{A}$ provides a compact approximation of the high-normality structure of the parameter space, which is subsequently used to guide the explanation generation process.

D. Explanation generation under normality constraints

We now use the family $\mathcal{A}$ of high-normality anchors to generate explanations for an abnormal situation (i.e., set of parameter values) $\mathbf{v}^e \in \mathcal{R}(\mathcal{V})$ such that $\varphi(\mathbf{v}^e) = 0$. For any anchor $A = \prod_{X \in \mathcal{V}} [h_X^{\min}, h_X^{\max}] \in \mathcal{A}$, the contrast of $\mathbf{v}^e$ with respect to A is defined as the set of variables that take values outside the corresponding anchor intervals:

$$\mathbb{C}_A(\mathbf{v}^e) = \{X \in \mathcal{V} : \mathbf{v}_X^e \notin [h_X^{\min}, h_X^{\max}]\} \tag{5}$$

This contrast identifies the parameters on which the abnormal point differs from the candidate normal region represented by A. In particular, variables unconstrained by the anchor are never contrasted, since their anchor interval coincides with the full range by definition.

To select the anchor used for an explanation, we associate with each pair $(\mathbf{v}^e, A)$ a "cost" obtained from the contrast. For

each variable $X \in \mathbb{C}_A(\mathbf{v}^e)$, we define the cost of associating $\mathbf{v}^e$ with anchor A on dimension X:

$$\text{cost}_X(\mathbf{v}^e, A) = \begin{cases} (h_X^{\min} - \mathbf{v}_X^e)/(r_X^{\max} - r_X^{\min}), \text{ if } \mathbf{v}_X^e < h_X^{\min} \\ (\mathbf{v}_X^e - h_X^{\max})/(r_X^{\max} - r_X^{\min}), \text{ if } \mathbf{v}_X^e > h_X^{\max} \end{cases} \quad (6)$$

The cost for associating $\mathbf{v}^e$ to anchor A is then: $\text{cost}(\mathbf{v}^e, A) = \sum_{X \in \mathbb{C}_A(\mathbf{v}^e)} \text{cost}_X(\mathbf{v}^e, A)$.

This quantity combines the number of contrasted variables with the magnitude of their violations, normalized by their corresponding ranges. We then select the nearest anchor $\widehat{A}(\mathbf{v}^e) = \text{argmin}_{A \in \mathcal{A}} \text{cost}(\mathbf{v}^e, A)$, and define the contrasted feature set of $\mathbf{v}^e$ as $\mathbb{C}_{\widehat{A}(\mathbf{v}^e)}(\mathbf{v}^e)$, which we denote as $\mathbb{C}(\mathbf{v}^e)$.

When the intended anchor precision is lower than 100%, a situation that does not exhibit the expected behavior may lie within an anchor. In such a case, it would be irrelevant to assign a "nearest anchor" that would not be the one in which it lies. When this occurs, we consider that we do not return an explanation for the parameter setting $\mathbf{v}^e$. In practice, this reflects a tradeoff between broader anchors and stricter precision: an anchor with a lower precision may yield larger normality regions and provide simpler explanations, but may induce more points that cannot be explained.

Next, we present two explanation generation methods based on the notion of contrast to the nearest anchor $\mathbb{C}(\mathbf{v}^e)$.

1) Contrast-only explanations: A first explanation strategy is obtained directly from the contrast. The explanation returned for $\mathbf{v}^e$ is the set $E_{\text{contrast}}(\mathbf{v}^e) = \mathbb{C}(\mathbf{v}^e)$.

Similar to actual causes, this method outputs a set of variables that support a behavior-restoring intervention. However, it provides the selected anchor as supporting evidence instead of the intervention itself. The anchor indicates that modifying the contrasted variables so as to enter the anchor yields a region in which the expected behavior is restored with a probability that is at least equal to the anchor's precision. This contrast-based strategy imposes counterfactuals with high-normality regions, but it does not guarantee minimality, as the contrast with the selected anchor may contain variables that are unnecessary for restoring the behavior.

2) Constrained actual causes: A second strategy consists in rerunning actual-cause identification under the restrictions induced by the selected anchor. Only contrasted variables $X \in \mathbb{C}(\mathbf{v}^e)$ are varied, within the interval specified by the anchor, i.e., $\widehat{\mathcal{R}}(X) = [h_X^{\min}, h_X^{\max}]$. Then, the actual-cause solver is applied within this restricted search space. As in standard actual-cause identification, the output is a cause set, together with a supporting intervention. Yet, the search is biased toward counterfactuals that remain close to a high-normality anchor.

These two strategies induce an inherent explanatory tradeoff. The contrast-only method directly uses the anchor as support, hence favoring explanations backed by highly normal regions, at the risk of including superfluous variables. The constrained actual-cause method recovers a stricter form of minimality by selecting only a subset of the contrasted variables. However, the supporting counterfactual may not remain inside the anchor itself, so its normality may be lower than that of the anchor used to guide the search.

IV. Experiments

This section presents the simulation used to evaluate our contributions (IV-A), the experimental settings (IV-B), and implementation details of our contributions (IV-C).

A. Swarm simulation and expected behavior

We evaluate the proposed approach via a classic swarm simulation [9]. This system provides a suitable testbed for the present study because it involves several continuous parameters, stochastic outcomes, and multiple distinct interventions that can restore the expected collective behavior (i.e., flocking).

At each step, each agent updates its heading according to three standard rules: *alignment* with neighbors; *attraction* toward their average position (cohesion); and *repulsion* from excessively close neighbors (separation). In addition, agent dynamics depend on their velocity, their perception radius, and a minimum separation distance. In our implementation, each rule produces a target heading correction. The applied correction is capped by a parameter-specific maximum angle, yielding three continuous factors that control the relative strength of alignment, cohesion, and separation.

B. Experimental settings

We initialize 25 agents using Poisson disc sampling within a 150×150-unit area centered in the simulation space, with a minimum pairwise distance of 8.4 units, and with headings drawn uniformly from a ±18° cone pointing in the same direction. Each simulation lasts for 50 steps. This duration is intentionally short: our goal is not to characterize all asymptotic flocking regimes, but to assess whether a given parameter configuration robustly and quickly produces coherent collective behavior. Short runs prevent slow-converging configurations from appearing falsely normal. Stochastic variability over individual runs is absorbed by averaging each label over 100 independent simulations per parameter setting, rather than by extending run length. To account for stochasticity, each heading update is additionally corrupted by Gaussian noise.

We consider a six-dimensional parameter space $\mathbf{v} \in \mathcal{R}(\mathcal{V})$:

- speed (ranging within 2-8 units per frame);
- perception radius (ranging within 1-15 units);
- minimum separation distance (1-10 units);
- alignment, cohesion and separation (within 0-15°, each).

Each situation $\mathbf{v} \in \mathcal{R}(\mathcal{V})$ takes a value for each parameter, from the given ranges. We sample 1500 such parameter settings uniformly over these ranges to form the sample set $\mathcal{S}$ used throughout the experiments. Each setting is evaluated through 100 repeated simulation runs. For each run, the target flocking behavior occurs when fewer than five agents lie outside a radius of 30 units from the center of the group at the end of the simulation. For each sampled parameter setting, we estimate the flocking probability $p(\mathbf{v})$, as the fraction of positive runs over the 100 simulations.

We differentiate two subsets within our sample set $\mathcal{S}$, depending on their flocking probability: $\mathcal{S}^+ = \{\mathbf{v} \in \mathcal{S} : p(\mathbf{v}) > 0.99\}$ for highly-positive points, employed to create the normality zones; and $\mathcal{S}^- = \{\mathbf{v} \in \mathcal{S} : p(\mathbf{v}) < 0.01\}$ for highly-negative points, employed to evaluate generated explanations.

C. Implementation

Our normality metric is implemented in Python and reuses the public "Anchors" implementation [8].

The first step of the proposed explanation method, i.e., determining the normality regions $\mathcal{A}$ in our domain space (subsec. III-C), is implemented in Python and also reuses the "Anchors" public implementation [8]. We use the highly-positive points in $\mathcal{S}^+$ as inputs for anchor construction, as they correspond to parameter settings that robustly exhibit the expected behavior. We consider three precision regimes for the global family of anchors $\mathcal{A}$: loose (90%), medium (95%), and tight (100%). Anchors with normalized volumes $\rho(A)$ smaller than 0.6 are discarded. In practice, no anchor from the tight regime survives this volume filter, so only the loose and medium families are actually employed.

To implement the second step of our explanation method (subsec. III-D), we provide two alternatives. First, for the contrast-only variant (subsec. III-D1), we provide a Python implementation. Second, for the constrained actual-causes variant (subsec. III-D2), we use the `actualcauses`[2] library [31] to identify the causes. This library provides approximate search procedures adapted to stochastic systems and returns both cause sets and supporting interventions.

V. Evaluation

This section presents the evaluation criteria for our two contributions. For the first one, we present the baseline normality measures (V-A) and the comparison criteria for evaluating our normality measure (V-B). For the second one, we present four comparison criteria for evaluating our method relative to the naive one (V-C).

A. Baseline normality metrics

To evaluate the proposed normality score, we compare it with six standard measures from the anomaly- and novelty-detection literature: Local Outlier Factor (LOF) [23], Gaussian Mixture Models (GMM) [24], inverse positive-kNN distance [22], Kernel Density Estimation (KDE) [27], Isolation Forests (IF) [26], and One-Class SVM (OC-SVM) [25]. All these methods are used as scalar normality scores, with larger values corresponding to more normal points. In particular, for the kNN baseline, we use the inverse of the average distance to the nearest positive neighbors. We use implementations from the `scikit-learn`[3] python library. Importantly, we normalize each parameter before using these baseline methods.

[2]https://pypi.org/project/actualcauses/

[3]https://scikit-learn.org/

B. Evaluation of the normality measure

We evaluate the proposed normality score against the six baseline measures above. Whenever random choices are involved, the same sampled instances are reused across measures for fair comparison. We use three criteria derived from norm theory [7], together with a complementary correlation analysis:

1) Pairwise recovery: we assess whether differences in normality can be recovered when changing restricted dimensions. To this end, we select 500 random pairs $(\mathbf{v}_1, \mathbf{v}_2)$ of anchor-covered points. For each pair, we start from $\mathbf{v}_1$ and progressively replace its dimensions by the corresponding dimensions of $\mathbf{v}_2$, following a sampled permutation of the variables. After each replacement, we recompute the normality score and record the smallest number of modified variables required to obtain exactly the score of $\mathbf{v}_2$. This minimum is approximated over 20 random permutations per pair, and the same permutations are used for all metrics. We report the average over pairs. Lower values indicate that normality comparisons can be captured through less feature change.

2) Dimension sensitivity: we evaluate one-dimensional sensitivity. For every anchor-covered point and for every parameter, we vary that parameter alone over evenly spaced values spanning its full range while keeping all other parameters fixed. A parameter is counted as sensitive if at least one such perturbation changes the normality score of the point. The score reported for a metric is the average number of individually sensitive parameters across the evaluated points. Lower values indicate that the metric depends on smaller sets of individually effective variables, which are identified as relevant to normality.

3) Robustness to irrelevant dimensions: we directly assess robustness to irrelevant dimensions. We augment the reference sample with 10 additional binary dimensions with random values. We then compare each metric's output on the original and augmented point. To ensure fair comparison of the metrics, the raw scores are rank-normalized, i.e., the score of each point is replaced by the inverse of its rank according to the evaluated metric. We then compute the normalized difference between each rank-normalized score of each pair of points (regular and with irrelevant features added). If this score reaches 0, it means that all points are ranked the same by the metric, regardless of the additional irrelevant features.

Finally, norm theory [7] explicitly stresses that the normality of a situation is different from its occurrence probability. In our case, this means that a normality metric should not consider a high probability of flocking $p(\mathbf{v})$ as a criterion for a high normality score. While there is no explicit requirement that they remain uncorrelated, we still report the Pearson correlation between the normality scores and the probability of flocking. This is not an explicit evaluation criterion, but a low correlation ensures that the probability is not directly used as a proxy for normality, as warned against by norm theory.

C. Evaluation of explanation generation

To evaluate our second contribution, we compare the *naive explanation method* (i.e., generating causes with counterfactu-

als searched in the full parameter space) to four versions of our proposed explanation method. We use our two proposed variants: *contrast-only* and *constrained actual-causes*. For each variant, we compare the two regimes of the global normality model: *loose* (90% required precision) and *medium* (95% required precision). We compare the explanations generated by these five methods for 100 points, selected randomly from the highly-negative samples $\mathcal{S}^-$. We evaluate each generation method along four criteria:

1) Simplicity: we first measure the number of causes returned for each point, which captures the overall complexity of the explanation. The contrast-only methods obtain 1 whenever an explanation is produced and 0 when the point lies inside the selected anchor. Second, we report the average size of the returned cause sets for each point. These two quantities provide an approximation of explanatory simplicity at the level of the full explanation and of individual causes, respectively. This criterion is widely recognized as a core desideratum of effective explanations, as it directly reflects cognitive selectiveness since fewer causes reduce the cognitive load of understanding [17].

2) Normality: we evaluate the normality of the supporting counterfactuals. For the actual-cause methods, we compute the proposed normality score $N(\mathbf{v}^c)$ on the counterfactual settings returned by the external solver [31]. For the contrast-based methods, which do not return a counterfactual intervention, we use the normalized anchor volume $\rho(A)$ of the selected anchor as the supporting normality value. In addition, we report the fraction of low-normality explanations, using a threshold of 0.6. By construction, this fraction is zero for the contrast-based methods, since only anchors with $\rho(A) > 0.6$ are retained. As argued in Section II-B, normality is a key desideratum of a relevant explanation.

3) Minimality: we evaluate minimality relative to the naive actual-cause baseline. For a returned cause C, let $\mathcal{C}_{ref}$ denote the set of naive causes for the same point. If some reference cause $C_{ref} \in \mathcal{C}_{ref}$ is a strict subset of C, the minimality score of C is defined as $|C_{ref}|/|C|$; otherwise, it is set to 1. This metric penalizes causes that are strictly less minimal than those found by the unconstrained search. Minimality is one of the main desiderata from actual causation [12], [2].

4) Efficiency: we report the percentage of simulator calls saved relative to the naive actual-cause search. For the naive cause search (our reference) and for the constrained cause searches, we use the number of calls used by the external algorithm [31]. For the contrast methods, we sample points within the anchor by uniformly selecting interventions across all contrasted parameters simultaneously until we find one that restores flocking. This comparison does not include the offline construction of the anchor families.

VI. Results

A. Normality measures

Table I reports the performance of the proposed anchor-based normality score and of the six baseline measures on the criteria presented in Sec. V-B: pairwise recovery (*pairs*), dimension sensitivity (*dimension*), robustness to irrelevant dimensions (*irrelevant*), and the Pearson correlation with the empirical flocking probability (*p-corr*).

	pairs ↓	dimension ↓	irrelevant ↓	p-corr ↓
anchor (ours)	**1.31±1.17**	**4.06±0.32**	**0.00±0.00**	**0.13**
LOF	**3.76±2.58**	6.00±0.00	0.51±0.30	0.18
GMM	6.00±0.00	6.00±0.00	0.46±0.25	0.32
kNN	6.00±0.00	6.00±0.00	0.45±0.28	0.46
KDE	6.00±0.00	6.00±0.00	0.47±0.29	0.42
IF	6.00±0.00	6.00±0.00	0.34±0.28	0.31
OC-SVM	5.99±0.10	**3.00±0.00**	0.34±0.25	0.45

TABLE I: Comparison of various normality measures with respect to criteria from norm theory.

On the pairwise comparison criterion, the proposed metric performs best by a large margin, as LOF is the only one that is not essentially at the maximum value. This shows that, unlike most baselines, the proposed metric is recoverable through changes on a few dimensions, in line with the intended notion that normality comparisons should depend on restricted subsets of variables.

Only our metric and OC-SVM are not sensitive to all six dimensions. This indicates that the proposed metric reacts to changes in only some dimensions, while most standard baselines require only one.

The proposed score shows perfect invariance under the addition of the irrelevant dimensions, while all baselines are substantially more affected.

Finally, while all metrics exhibit relatively low correlation with the empirical flocking probability, the proposed metric is the least correlated. This shows that it does not rely on the probability of flocking. More generally, none of the compared metrics can be reduced to probability alone, which is consistent with the distinction emphasized by norm theory.

These results show that the proposed metric is the only one to perform strongly on all three target criteria: it is recoverable through low-dimensional pairwise changes, depends on only part of the parameters under one-dimensional perturbations, and is fully robust to irrelevant dimensions. These three criteria together constitute a sensitivity analysis of the metric to its input space based on its response to individual parameter variations and its robustness to irrelevant dimensions.

B. Qualitative explanation example

To illustrate our approach qualitatively, we present results for the naive explanation generation method and the different versions of our approach. We randomly select a run in which flocking does not occur. Table III shows the parameter values of this instance. Why didn't flocking occur? A naive actual cause search returns 2 causes: $\{\texttt{alignment}, \texttt{view}\}$, and $\{\texttt{cohesion}, \texttt{view}\}$, with normality scores of respectively 0.60 and 0.50. Table III shows the closest prototype (here selected from the loose regime). Running the constrained cause search returned only the first identified cause and took $\sim$5,000 model calls instead of $\sim$90,000. The final explanation

		Nb-causes	Cause-size ↓	Normality ↑	Low-norm-frac ↓	Minimality ↑	Calls saved (%) ↑
No anchor	Naive causes	2.0±2.0	3.46±2.9	0.58±0.1	0.48±0.4	1.0±0.0	0.0±0.0
Loose anchors	Contrast only	0.99±1.0	1.63±0.8	0.63±0.0	0.0±0.0	0.87±0.2	99.7±0.2
	Constrained causes	0.96±1.0	1.35±0.8	0.6±0.1	0.48±0.5	1.0±0.0	95.54±9.0
Medium anchors	Contrast only	0.99±1.0	1.84±0.8	0.63±0.0	0.0±0.0	0.79±0.2	99.67±0.3
	Constrained causes	1.09±1.1	1.57±1.1	0.6±0.1	0.29±0.4	1.0±0.0	92.71±17.8

TABLE II: Metrics for our causal explanations

	Parameter values	Prototype bounds	Match
speed	2.54	[-, 4.47]	✓
view	4.68	[10.61, -]	×
min-sep	8.87	[-, 8.98]	✓
alignment	1.72°	[6.30°, -]	×
cohesion	10.89°		✓
separation	1.72°		✓

TABLE III: Contrast of the illustrative instance to its closest prototype. The unconstrained ranges are not represented.

is therefore that flocking failed because the perception radius and alignment strength were too low.

C. Counterfactual explanations

Table II compares the five explanation-generation methods: naive actual-cause search over the full parameter domains (no anchors), contrast-only explanations derived from the loose and medium anchor families, and constrained actual-cause search guided by the same two anchor families. This comparison across both precision regimes constitutes a sensitivity analysis of the method to its main threshold parameter τ.

The first column shows that our normality-guided methods generate fewer causes: it focuses on limited but targeted explanations. However, points within anchors (false positives) make some points impossible to explain for our approach. Hence, the average sometimes falls below 1.

The naive method returns causes significantly larger than the contrast-based methods, which themselves returns slightly larger causes than the constrained causes (second column). Normality guided search filters out longer causes and removes superfluous variables introduced by raw contrast alone.

The proposed methods also improve the normality of the supporting counterfactuals (third column). However, the metric shows little variation, which reveals that the fraction of low-normality explanations is the more informative indicator: when using the medium regime, normality-guided search removes counterfactuals that would be weak supports for explanation.

Minimality (fifth column) reveals the main difference between the proposed explanation strategies. The constrained actual-cause methods remain as minimal as the baseline naive approach. Conversely, the raw contrast methods include variables that are not strictly necessary.

All normality-guided methods also produce substantial computational savings (sixth column). This shows that restricting the search to high-normality regions is not only explanatorily beneficial, but also practically useful for reducing the cost of counterfactual search.

As a reference, finding a cause for a single instance required $\sim$100,000 $\pm$ 31,000 simulation calls for the naive method, versus $\sim$5,000$\pm$14,000 and $\sim$7,500$\pm$20,000 for the loose and medium constrained searches, respectively. Building the corresponding normality models required $\sim$130,000, and $\sim$145,000 calls ($\sim$160,000 for the tight regime). The offline cost is thus recovered after only a few queries.

These results highlight two distinct tradeoffs. On the one hand, contrast-only explanations provide the strongest normality but sacrifice minimality. However, while constrained actual-cause search preserves minimality, it still substantially improves normality over the naive baseline. On the other hand, looser anchors tend to yield simpler and cheaper explanations than medium ones, but increase the proportion of abnormal points that cannot be explained by contrast.

Overall, all normality-guided methods outperform naive actual-cause search in explanatory simplicity and computational cost. The contrast-only approach greatly improves normality, while the constrained variant provides the best balance when minimality must be preserved.

VII. Discussion

A. Broader uses of the global normality model

Beyond their role in local counterfactual explanations, global normality zones can also be interpreted as a form of global explanation of system behaviors. Each anchor is a simple rule over a subset of parameters that characterizes a region in which the expected behavior is robustly observed. Because these rules are sparse and directly interpretable, the resulting family of anchors provides a readable description of how the system can operate normally. In that sense, the method can also support broader questions, such as "under which conditions does the system normally flock?" This perspective may be useful for offline parameter analysis for understanding the system's normal operating regimes, as illustrated in Appendix C.

B. Condition of use and generalization

1) Granularity of the metric: the proposed metric is region-based rather than fully point-sensitive: all points in the same anchor receive the same score, as normality is identified with the normalized volume of the anchor that supports the expected behavior around them. This is appropriate for the present use, where the score is mainly intended to validate, compare, or filter counterfactual supports and to construct global normality regions. However, it also limits the granularity of the metric, which does not provide a fine ranking among points inside

the same anchor. The method is therefore better suited for distinguishing between robust and brittle counterfactual supports than for producing a precise ordering among close alternatives.

2) Shape of the normality zone: a condition for the success of the proposed method is the existence of low-dimensional, high-precision anchors. When normal regions can be described by rules involving only a subset of the parameters, the method behaves as intended: it identifies restricted sets of relevant dimensions and ignores irrelevant ones. However, its benefits become limited in low-dimensional spaces, where all dimensions tend to be relevant. Additionally, it becomes more fragile in high-dimensional but "unstructured" spaces, where anchors are likely to become dense. For example, for raw image pixels, meaningful normality regions are unlikely to be expressible through sparse threshold rules. In such settings, causal representation learning [32] may be needed to recover structured variables before applying the method.

3) Scalability: high dimensionality (i.e., more than tens of variables) also raises computational difficulties. As the method relies on a representative reference sample to model normality regions globally, the number of required samples grows rapidly with the number of dimensions. We could pre-select variables before applying our method. Otherwise, the number of samples could be increased to maintain a global offline normality model, yet reducing the method's efficiency. Finally, anchors could be constructed online, i.e., only when an explanation is requested and only around candidate counterfactuals relevant to the point explained. Yet, this option would lose online efficiency and the method's global explanation capabilities.

4) Scope and transfer conditions: the results of this paper are established on a single experimental domain, which limits the empirical support for the claimed generality. We present this work as an initial validation. However, properties of the method make its transfer to other complex systems plausible. The method requires no knowledge of the system's internal dynamics, only a behavior detector and the ability to sample the parameter space. It also accommodates stochasticity through the precision threshold τ. However, we argued that transfer can only occur towards systems whose normal operating regions admit sparse threshold rules over a subset of interpretable parameters. We also discussed the conditions for transfer when the number of variables grows or diminishes.

Additionally, even within our evaluation scenario, the evaluation scope remains limited to a single behavior. Notably, we use a binary predicate φ that does not distinguish flocking modes; extending the predicate to richer behavioral distinctions is a natural direction for future work that requires no modification to the framework itself.

A more precise sensitivity analysis of the various parameters, beyond the one conducted on the precision threshold τ (which distinguishes the prototype regimes) and on the normality metric's input modifications, is also a natural direction for future work.

C. Dependence on the external actual-cause solver

The evaluation of explanation quality depends on an external actual-cause identification solver [14]. This solver efficiently searches for causes in stochastic models but does not optimize for the best supporting counterfactual in the sense suggested by graded causation: we therefore did not compare against the best possible normal counterfactuals, which naturally favors our approach. This limitation also reinforces the method's practical relevance: searching for actual causes is already difficult [31], and searching for the best counterfactuals would be harder still, so restricting the search to high-normality regions is not merely a theoretical refinement but a way of making the problem tractable. The approach is also modular: any actual-cause solver supporting restricted intervention domains could be combined with the same normality model.

D. Selecting the "nearest" anchor

We chose to associate only one anchor with each situation that does not exhibit the expected behavior. This choice forces the explanation to be selective. The main consequence is simplicity (i.e., size and number of causes generated). However, while simplicity is a good trait for a relevant explanation, it can also imply discarding important information. Contrasting the same point to several anchors could yield additional normal and relevant causes. In future work, we will consider ranking and filtering anchors to associate with a point, so as to extend the search for causes beyond the nearest anchor.

VIII. Conclusion

Counterfactual explanations aim to identify changes that would prevent unexpected behavior, but not all such changes are equally relevant. We address this by introducing a computable notion of normality for counterfactual alternatives, framed as the similarity of adjacent configurations within a multi-dimensional low-dimensional space, where only a limited number of dimensions should matter for the comparison. Our first contribution is an anchor-based normality score that evaluates a behavior-restoring point through the normalized volume of the largest low-dimensional, high-precision zone supporting the expected behavior around it. Our second contribution is a method for generating counterfactual explanations under normality constraints from a global model of such zones, yielding two strategies: one based on contrast with the nearest normal zone, and one based on cause search restricted to it.

We evaluated the approach on a swarm simulation in which flocking was the expected collective behavior. The proposed metric better matches the intended properties of normality than standard anomaly detection baselines, and incorporating normality into counterfactual explanation yields simpler explanations, fewer low-normality supports, and substantially lower computational cost, though comparing the two strategies reveals an inherent tradeoff between normality and minimality. Overall, normality can be operationalized in a way that is both cognitively motivated and computationally useful for systems whose normal operating regions admit sparse, low-dimensional characterizations.

ACKNOWLEDGMENTS

Concerning the use of generative AI tools, we employed a LLM for polishing a first paper draft, which we then checked and rephrased manually into the final proposal.

REFERENCES

[1] J. Pearl, *Causality*, 2nd ed. Cambridge University Press, 2009.

[2] J. Y. Halpern, "A modification of the Halpern-Pearl definition of causality," in *Proceedings of the 24th International Conference on Artificial Intelligence*, ser. IJCAI'15. AAAI Press, Jul. 2015, pp. 3022–3033.

[3] R. Guidotti, "Counterfactual explanations and how to find them: Literature review and benchmarking," *Data Mining and Knowledge Discovery*, vol. 38, no. 5, pp. 2770–2824, 2024.

[4] J. Y. Halpern and C. Hitchcock, "Graded Causation and Defaults," *The British Journal for the Philosophy of Science*, vol. 66, no. 2, pp. 413–457, 2015.

[5] V. Chandola, A. Banerjee, and V. Kumar, "Anomaly detection: A survey," *ACM Comput. Surv.*, vol. 41, no. 3, pp. 15:1–15:58, 2009.

[6] M. Pimentel, D. Clifton, L. Clifton, and L. Tarassenko, "A review of novelty detection," *Signal Processing*, vol. 99, pp. 215–249, 2014.

[7] D. Kahneman and D. T. Miller, "Norm theory: Comparing reality to its alternatives," *Psychological Review*, vol. 93, no. 2, pp. 136–153, 1986.

[8] M. T. Ribeiro, S. Singh, and C. Guestrin, "Anchors: High-Precision Model-Agnostic Explanations," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 32, no. 1, Apr. 2018.

[9] C. W. Reynolds, "Flocks, herds and schools: A distributed behavioral model," in *Proceedings of the 14th Annual Conference on Computer Graphics and Interactive Techniques*, ser. SIGGRAPH '87. Association for Computing Machinery, 1987, pp. 25–34.

[10] W. Bechtel and A. Abrahamsen, "Explanation: A mechanist alternative," *Studies in History and Philosophy of Science Part C: Studies in History and Philosophy of Biological and Biomedical Sciences*, vol. 36, no. 2, pp. 421–441, 2005.

[11] B. Iooss and P. Lemaître, "A review on global sensitivity analysis methods," *Uncertainty management in simulation-optimization of complex systems: algorithms and applications*, pp. 101–122, 2015.

[12] J. Y. Halpern and J. Pearl, "Causes and explanations: A structural-model approach: Part i: Causes," in *Proceedings of the Seventeenth Conference on Uncertainty in Artificial Intelligence*, ser. UAI'01. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., Aug. 2001, pp. 194–202.

[13] S. Wachter, B. Mittelstadt, and C. Russell. (2018) Counterfactual Explanations without Opening the Black Box: Automated Decisions and the GDPR.

[14] S. Reyd, A. Diaconescu, J.-L. Dessalles, and L. Esterle, "A Roadmap for Causality Research in Complex Adaptive Systems," in *2024 IEEE International Conference on Autonomic Computing and Self-Organizing Systems Companion (ACSOS-C)*, 2024, pp. 35–40.

[15] S. Reyd, A. Diaconescu, and J.-L. Dessalles, "Finding Relevant Causes in Complex Systems: A Generic Method Adaptable to Users and Contexts," in *2025 IEEE International Conference on Autonomic Computing and Self-Organizing Systems (ACSOS)*, 2025, pp. 100–111.

[16] E. Houze, J.-L. Dessalles, A. Diaconescu, and D. Menga, "What Should I Notice? Using Algorithmic Information Theory to Evaluate the Memorability of Events in Smart Homes," *Entropy*, vol. 24, no. 3, p. 346, 2022.

[17] T. Miller, "Explanation in artificial intelligence: Insights from the social sciences," *Artificial Intelligence*, vol. 267, pp. 1–38, 2019.

[18] ——, "Contrastive explanation: A structural-model approach," *The Knowledge Engineering Review*, vol. 36, p. e14, Jan. 2021.

[19] N. Lloyd and P. Lewis, "A Socio-Technical Perspective on Fostering Sustainable Community Development via Social Capital and Social Cohesion," in *2024 IEEE International Conference on Autonomic Computing and Self-Organizing Systems Companion (ACSOS-C)*, 2024, pp. 61–66.

[20] P. Salmani and P. R. Lewis, "Self-Evaluation can Help Agents Meet Social Expectations," in *2025 IEEE International Conference on Autonomic Computing and Self-Organizing Systems Companion (ACSOS-C)*, Sep. 2025, pp. 1–7.

[21] C. M. Barnes, A. Ekárt, and P. R. Lewis, "Social Action in Socially Situated Agents," in *2019 IEEE 13th International Conference on Self-Adaptive and Self-Organizing Systems (SASO)*, 2019, pp. 97–106.

[22] S. Ramaswamy, R. Rastogi, and K. Shim, "Efficient algorithms for mining outliers from large data sets," in *International Conference on Management of Data*, ser. SIGMOD '00. Association for Computing Machinery, 2000, pp. 427–438.

[23] M. M. Breunig, H.-P. Kriegel, R. T. Ng, and J. Sander, "LOF: Identifying density-based local outliers," in *Proceedings of the 2000 ACM SIGMOD International Conference on Management of Data*, ser. SIGMOD '00. Association for Computing Machinery, 2000, pp. 93–104.

[24] G. J. McLachlan, S. X. Lee, and S. I. Rathnayake, *Finite mixture models*. Annual Reviews, 2019, vol. 6, no. 1.

[25] B. Schölkopf, R. C. Williamson, A. Smola, J. Shawe-Taylor, and J. Platt, "Support Vector Method for Novelty Detection," in *Advances in Neural Information Processing Systems*, vol. 12. MIT Press, 1999.

[26] F. T. Liu, K. M. Ting, and Z.-H. Zhou, "Isolation Forest," in *2008 Eighth IEEE International Conference on Data Mining*, 2008, pp. 413–422.

[27] E. Parzen, "On Estimation of a Probability Density Function and Mode," *The Annals of Mathematical Statistics*, vol. 33, no. 3, pp. 1065–1076, 1962.

[28] C. Gruhl, B. Sick, A. Wacker, S. Tomforde, and J. Hähner, "A building block for awareness in technical systems: Online novelty detection and reaction with an application in intrusion detection," in *2015 IEEE 7th International Conference on Awareness Science and Technology (iCAST)*, 2015, pp. 194–200.

[29] C. Gruhl, B. Sick, and S. Tomforde, "Novelty detection in continuously changing environments," *Future Generation Computer Systems*, vol. 114, pp. 138–154, 2021.

[30] M. T. Ribeiro, S. Singh, and C. Guestrin, ""Why Should I Trust You?": Explaining the Predictions of Any Classifier," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, ser. KDD '16. New York, NY, USA: Association for Computing Machinery, Aug. 2016, pp. 1135–1144.

[31] S. Reyd, A. Diaconescu, and J.-L. Dessalles, "Searching for actual causes: Approximate algorithms with adjustable precision," Jul. 2025.

[32] B. Scholkopf, F. Locatello, S. Bauer, N. R. Ke, N. Kalchbrenner, A. Goyal, and Y. Bengio, "Toward Causal Representation Learning," *Proceedings of the IEEE*, vol. 109, no. 5, pp. 612–634, 2021.

APPENDIX

A. Relation to full causal formalism

1) Causal variables: Sec. II-A presents a simplified intervention notation adapted to the setting of this paper. This appendix clarifies how this notation relates to the standard Structural Causal Model (SCM) formalism [1], [12], [2].

An SCM is usually defined as a tuple $\mathcal{M} = (\mathcal{U}, \mathcal{V}, \mathcal{R}, \mathcal{F})$, of exogenous variables $\mathcal{U}$, endogenous variables $\mathcal{V}$, ranges $\mathcal{R}$, and structural equations $\mathcal{F}$. Endogenous variables are those that support interventions; structural equations fix the values of the endogenous variables, and exogenous variables encode all the additional information needed to compute the endogenous values. Given exogenous values $\mathbf{u} \in \mathcal{R}(\mathcal{U})$, the structural equations fix the values of all endogenous variables. An intervention $[\mathbf{Y} \leftarrow \mathbf{y}']$ corresponds to replacing the structural equation for each $X \in \mathbf{Y}$ by $X = \mathbf{y}_{|X}$. When a Boolean formula φ holds in a given context, we write $(\mathcal{M}, \mathbf{u}) \models \varphi$. When an intervention cancels it, we write $(\mathcal{M}, \mathbf{u}) \models [\mathbf{Y} \leftarrow \mathbf{y}']\neg\varphi$.

The endo/exogenous distinction is formal and should not be confused with the distinction between variables that are internal or external to the simulated system. Some internal variables can be exogenous to the SCM and vice versa, depending on the modeled interventions.

The parameter-only setting used in our experiments corresponds to a degenerate SCM with no endogenous interactions. The equations trivially copy the simulation parameters from

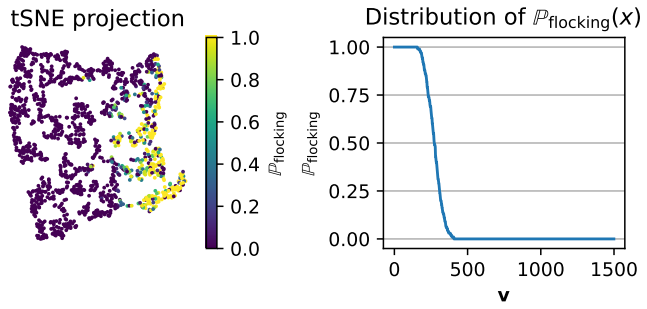


Fig. 2: Distribution of the dataset used in our experiments. The left figure shows the tSNE projections. The right figure shows the sorted probability of flocking for each parameter value.

exogenous to endogenous versions. The target predicate is then evaluated using the simulation and behavior detector. An SCM used to explain the simulation variables would include non-trivial structural equations derived from the simulation. For example, the position of an agent at time $t = 5$ depends on the positions and headings of neighboring agents at time $t = 4$. Additionally, initial states of the simulation would need to be added to exogenous variables, and the target predicate would be evaluated using only the behavior detector.

In a non-trivial SCM, an intervention is not a simple value update but replaces a structural equation and implies consequences for the other endogenous variables. This notion is notably missing in counterfactual explanations in XAI and motivates our use of actual causation.

2) Actual causes: The full HP definition accounts for endogenous dependencies and allows for variables to be held fixed as "witnesses". For instance, to explain the absence of flocking with an agent's behavior, we check whether an alternate behavior restores flocking. In this counterfactual world, other agents, called "witnesses", may keep their actual behavior despite the intervention's indirect influence.

Formally, $\mathbf{X} = \mathbf{x}$ is an actual cause of φ in $(\mathcal{M}, \mathbf{u})$ if three conditions hold [2]. First, the cause and the outcome are true in the actual situation: $(M, \mathbf{u}) \models (\mathbf{X} = x)$ and $(\mathcal{M}, \mathbf{u}) \models \varphi$. Second, there exist $\mathbf{x}' \in \mathcal{R}(\mathbf{X})$, and "witness" variables $\mathbf{W} = \mathbf{w}^*$, that prevent the outcome: $(\mathcal{M}, \mathbf{u}) \models [\mathbf{X} \leftarrow \mathbf{x}', \mathbf{W} \leftarrow \mathbf{w}^*]\neg\varphi$. Third, $\mathbf{X}$ is minimal: no strict subset of $\mathbf{X}$ satisfies the previous conditions.

The simplified definition presented in the main text, adapted to our degenerate SCM, corresponds to the special case in which no "witness" is needed. The proposed normality model is independent of this simplification and can be used when "witnesses" are needed.

B. Dataset distribution

Fig. 2 presents the distribution of the dataset used in our experiments to model normality and for evaluation. The clear bimodality of the label distribution shows our naive range selection includes numerous points whose flocking or non-flocking is unambiguous. The projection shows that these points are distributed in robust neighborhoods.

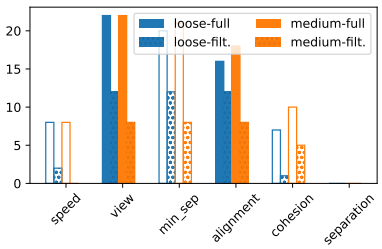


Fig. 3: Parameters used for the global high-normality regions. Filled portions correspond to lower thresholds and vice versa.

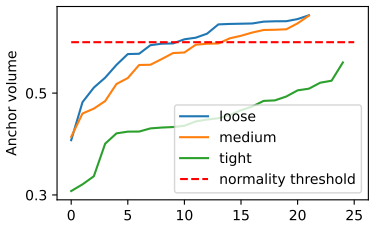


Fig. 4: Volumes of the global high-normality-zones and the normality threshold of 0.6.

C. Global explanations

While this is not a contribution highlighted in this paper, the global high-normality zones produced by our method also produce global explanations of the system.

Fig. 3 shows which parameters are used to define the global normality regions. Notably, most models can be interpreted as: flocking occurs for low speeds, minimum separations, cohesion factors, and high perception radii, and alignments. The separation factor is more nuanced.

Fig. 4 shows that filtering removed around half of both the loose and medium anchors, while it removed all the tight ones. Additionally, around three times more tight anchors were initially identified.

Tab. IV shows that tighter anchors are more numerous and defined on more dimensions, which is aligned with contrast results from Sec. VI-C. Additionally, filtering the anchors barely degrades their average precisions. Finally, the volumes and coverages of the anchors are well aligned and confirm that the tight anchors, while carrying precise and rich information, only support small and low-coverage zones.

Regime	Filt.?	Nb.	# Dims	Precision	Size	Coverage
Loose	Full	48	3.29±0.5	96%±3	51%±6	2%±1
	Filt.	22	3.32±0.5	92%±2	59%±6	5%±2
Medium	Full	48	3.56±0.5	98%±2	50%±6	2%±1
	Filt.	22	3.59±0.5	97%±1	56%±6	4%±2
Tight	Full	48	3.5±0.5	97%±4	48%±8	2%±1

TABLE IV: Statistics about the identified anchors for the global high-normality zones. "Nb." refers to the number of anchors produced by our method. "Filt." is short for filtered.