Normative Monitoring Using Bayesian Networks: Defining a Threshold for Conflict Detection

Annet Onnes(✉), Mehdi Dastani, and Silja Renooij

Department of Information and Computing Sciences, Utrecht University, Utrecht, The Netherlands
{a.t.onnes,m.m.dastani,s.renooij}@uu.nl

Abstract. Normative monitoring of black-box AI systems entails detecting whether input-output combinations of AI systems are acceptable in specific contexts. To this end, we build on an existing approach that uses Bayesian networks and a tailored conflict measure called IOconfl. In this paper, we argue that the default fixed threshold associated with this measure is not necessarily suitable for the purpose of normative monitoring. We subsequently study the bounds imposed on the measure by the normative setting and, based upon our analyses, propose a dynamic threshold that depends on the context in which the AI system is applied. Finally, we show the measure and threshold are effective by experimentally evaluating them using an existing Bayesian network.

Keywords: Bayesian Networks · Conflict Measures · Responsible AI · Normative Monitoring

1 Introduction

Given the omnipresence of AI systems, it is important to be able to guarantee their safety and reliability within their context of use, especially when the AI system is a black-box that is not easily interpretable or sufficiently transparent. To this end, we previously introduced a novel framework for model-agnostic normative monitoring under uncertainty [8,9]. Since the exact design underlying the system being monitored is irrelevant, we simply refer to it as the 'AI system'; our only assumption is that this system has excellent *general* performance on the task for which it is designed. However, the AI system can be employed in different environments in each of which additional *context-specific* rules, protocols, or other types of values or norms exist or emerge, which need to be adhered to. To determine whether the AI system operates acceptably in the context of a given environment, we need to monitor the system for adhering at run-time, preferably using interpretable and model-agnostic methods.

The framework for normative monitoring under uncertainty includes, in addition to the AI system, a *normative model* and a *monitoring process* [8,9]. The normative model captures the input-output pairs of the AI system, as well as variables that describe information that is relevant to the specific environment

Z. Bouraoui and S. Vesic (Eds.): ECSQARU 2023, LNAI 14294, pp. 149–159, 2024.
https://doi.org/10.1007/978-3-031-45608-4_12

and context in which the AI system operates. In some situations, this context may dictate that the output provided by the AI system is undesirable or unacceptable. In that case, a warning has to be issued by the monitoring process when it compares the input-output pair of the AI system against the normative model.

Previously, we proposed the use of Bayesian Networks (BNs) to implement the normative model and we adjusted a conflict measure for BNs to compare the in-context behaviour of the AI system against the normative model [9]. The adjusted measure, IOconfl, comes with an intrinsic threshold that can be used to determine whether or not to flag an input-output pair of the AI system as possibly unacceptable in the current context. However, the suitability of this threshold for the purpose of normative monitoring was not investigated.

In the current paper, we therefore further study and evaluate the adjusted measure and the intrinsic threshold for the normative monitoring setting. We analyse the bounds imposed on the IOconfl measure by our normative setting and propose a new dynamic, context-specific, threshold. In addition, we compare different measures and thresholds in a controlled experimental setting and demonstrate that our proposed measure and threshold serve to take context into account and result in flagging behaviour that is different from the considered alternatives.

Our paper is organised as follows. After providing preliminaries in Sect. 2, we further review different measures in Sect. 3. In this section, we also discuss the suitability of the related thresholds for the purpose of monitoring. In Sects. 4 and 5, we analyse the bounds on the IOconfl measure and define the dynamic threshold. We experimentally evaluate the use of the measure and impact of the chosen threshold in Sect. 6 and conclude the paper with Sect. 7.

2 Preliminaries

In this section we introduce our notations and provide formal definitions of the different components in our framework. A schematic overview of the framework is given in Fig. 1 (see [9] for further details). The framework assumes that both the AI system and the normative model can be interpreted as probabilistic models that somehow represent or reflect a probability distribution Pr over a set of random variables, that is, both models capture the dependencies along with their uncertainties as present in a part of the real world (the *target system*).

We use capital letters to denote variables, bold-faced in case of sets, and consider distributions $\Pr(\mathbf{V})$ over a set of random variables $\mathbf{V}$. Each variable $V \in \mathbf{V}$ can be assigned a value v from its domain $\Omega(V)$; a joint value assignment (or configuration) $v_1 \wedge \ldots \wedge v_n$ to a set of n variables $\mathbf{V} = \{V_1, \ldots, V_n\}$ is denoted by $\mathbf{v} \in \Omega(\mathbf{V}) = \bigtimes_{i=1}^{n} \Omega(V_i)$. The normative model and AI system can now be defined as follows (generalised from [9]):

- the *AI system* represents a joint distribution $\Pr^S(\mathbf{V}^S)$ over a set of variables $\mathbf{V}^S = \mathbf{I}^S \cup \mathbf{O}^S$, where $\mathbf{I}^S$ and $\mathbf{O}^S$ are non-empty sets of input variables and output variables, respectively;

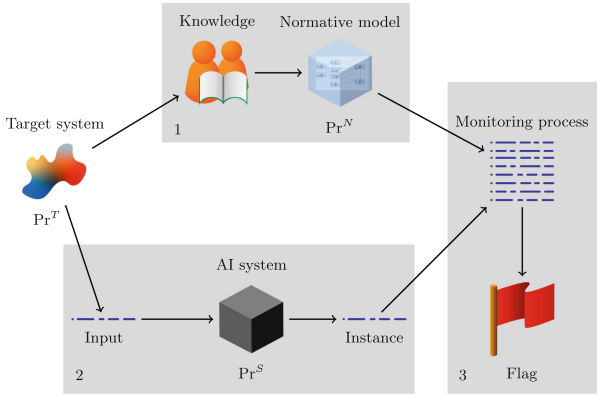

Fig. 1. Overview of the framework for normative monitoring under uncertainty, including (1) a normative model, (2) an AI system, and (3) a monitoring process.

– the *normative model* represents a joint distribution $\Pr^N(\mathbf{V}^N)$ over a set of variables $\mathbf{V}^N = \mathbf{I}^N \cup \mathbf{O}^N \cup \mathbf{A}$, where $\mathbf{I}^N$ and $\mathbf{O}^N$ result from (easy) mappings $\mathbf{I}^S \rightarrow \mathbf{I}^N$ and $\mathbf{O}^S \rightarrow \mathbf{O}^N$, and $\mathbf{A} = \mathbf{C} \cup \mathbf{H}$ is a set of additional variables, including a non-empty set of context variables $\mathbf{C}$ and possibly other hypothesis or hidden variables $\mathbf{H}$.

Superscripts indicate the type of model that the variables and distributions belong to, where S refers to the AI system and N to the normative model. Without loss of generality and for ease of exposition, we take $\mathbf{I}^S = \mathbf{I}^N$ and $\mathbf{O}^S = \mathbf{O}^N = \{O\}$ in this paper.

In the current paper we assume that the normative model is implemented by a Bayesian network. Bayesian networks (BNs) are probabilistic graphical models that are interpretable and can be handcrafted [6]. Interpretability and transparency of (part of) the normative model is important since it includes variables specific to the context in which the AI system operates, and we assume that expert knowledge is needed to design and interact with it. To be precise, a BN $\mathfrak{B} = (G, \Pr)$ is a compact representation of a joint probability distribution $\Pr(\mathbf{V})$ that combines an acyclic directed graph G, with nodes $\mathbf{V}$ and directed edges that describe the (in)dependences among $\mathbf{V}$, with local distributions specified for each variable, conditional on its parents in the graph G [3]. As such, BNs allow for computing any probability of interest from their distribution, which facilitates the computation of various measures that could be employed in the monitoring process to flag for unacceptable input-output pairs.

3 Measures and Thresholds

In our normative monitoring setting we want to *flag* an input-output pair of the AI system when the input-output combination is considered to be undesirable or unacceptable in the current context, according to the normative model. To

determine the extent to which the AI system's input-output pair is acceptable, a measure and a corresponding threshold are required to determine when to raise a flag. In this section we review two measures used for the purpose of Anomaly Detection (AD), a setting related to ours. In addition, we consider a measure proposed explicitly for the normative monitoring setting. Subsequently we will discuss the associated thresholds and argue that the choice of such a threshold, even for measures that have a seemingly intrinsic one, is not trivial.

3.1 Measures

The purpose of AD is to determine whether a set of observations in the real world should be classified as anomalous [1]. To this end, the observed behaviour is typically compared against a model of normal behaviour using one of many anomaly detection techniques. Among such techniques are Bayesian networks, used in combination with a likelihood measure [4] or a measure of conflict [7].

Johansson and Falkman [4], for example, train a BN to represent normal maritime vessel behaviour and use the *likelihood* $\Pr(\mathbf{v})$ of an instance of vessel behaviour $\mathbf{v}$ to detect anomalous behaviour. An instance is flagged when its probability of occurrence is low. However, rare behaviour is not necessarily anomalous [5]. To overcome this issue, likelihood of an instance $\mathbf{v} = v_1 \wedge \ldots \wedge v_n$, $n \geq 2$, can be compared to the probability of the observations occurring independently: $(\Pr(v_1) \cdot \ldots \cdot \Pr(v_n))/\Pr(\mathbf{v})$. This *conflict measure*, introduced by Jensen et al. [2], was used by Nielsen and Jensen [7] to detect anomalies in production plants based upon sensor readings. In case of normal behaviour, again captured by a BN, the sensor readings should be positively correlated, regardless of whether their combination is rare. An instance is flagged when the combination of observations seems internally incoherent.

In the normative monitoring setting, we want to detect input-output pairs for which the output seems inconsistent with the input in the context prescribed by the normative model. For this it does not matter whether or not the input-output pair is rare. To this end we proposed an adapted version of the conflict measure, $\text{IOconfl}(o, \mathbf{i} \mid \mathbf{c})$ [9]:

$$\text{IOconfl}(o, \mathbf{i} \mid \mathbf{c}) = \log \frac{\Pr^N_{\mathbf{c}}(o) \cdot \Pr^N_{\mathbf{c}}(i_1 \wedge \ldots \wedge i_n)}{\Pr^N_{\mathbf{c}}(o \wedge i_1 \wedge \ldots \wedge i_n)} \tag{1}$$

where $\mathbf{i} = i_1 \wedge \ldots \wedge i_n$, $n \geq 1$, is input for the AI system, o is the associated output returned by the AI system, and $\Pr^N_{\mathbf{c}}(\cdot)$ is a short hand for $\Pr^N(\cdot \mid \mathbf{c})$ with $\mathbf{c}$ a configuration for one or more of the context variables $\mathbf{C}$ from the normative model. Note that the IOconfl measure differs from the original conflict measure by separating only the marginal over the output o from the joint over the inputs $i_1, \ldots, i_n$, which effectively eliminates the effect of conflict within the input of the AI system [9]. Moreover, the IOconfl measure takes into account the specific context prescribed by the normative model. The probabilities in Eq. 1 are therefore computed from the normative model, and conditioned on a specific context $\mathbf{c}$.

3.2 Thresholds

In the monitoring process, any measure needs a threshold to decide between flagging or not flagging. The likelihood measure, aside from only detecting rare cases, requires a threshold δ to be set to capture when a case is rare enough to flag: $\Pr(\mathbf{v}) \leq \delta$. The choice of threshold must be based on expert knowledge, taking into account the cost of false positives and false negatives in the domain of application [4]. A benefit of the Jensen conflict measure is that the choice of threshold appears easy, since it has an intrinsic threshold of 0: if the measure exceeds $\log 1 = 0$ then it is more likely to find the combination of observations assuming they are independent (the product of marginals) rather than by assuming their dependencies as captured in the BN's joint distribution. According to the BN, therefore, the instance is incoherent if its conflict value exceeds 0.

The same intrinsic threshold of $\log 1 = 0$ seems an intuitively appealing default threshold for the adjusted conflict measure IOconfl. Using this threshold would entail that an input-output pair of the AI system is flagged when the input and its associated output are not correlated positively according to the normative model. This situation is, however, not necessarily what we want to flag. Instead, we want to find a threshold that enables flagging for a situation where, according to the context prescribed by the normative model, the output is not acceptable given the input. To reconsider the choice for this intrinsic threshold, we study how the constraints imposed by the normative monitoring setting affect the values of the IOconfl measure.

4 Bounding IOconfl

To better understand the IOconfl measure from Eq. 1, we will study its boundaries under various conditions specific to the normative monitoring setting. Firstly, we assume that the AI system returns the output that is most likely, according to $\Pr^S$, given the input, i.e. the AI system returns (ties disregarded):

$$o^* = \arg\max_{o_k \in \Omega(O)} \Pr^S(o_k \mid \mathbf{i})$$

Thus, if outcome variable O has r values, then $\Pr^S(o^* \mid \mathbf{i}) \in [\frac{1}{r}, 1]$.

To facilitate our analysis of $\text{IOconfl}(o, \mathbf{i} \mid \mathbf{c})$ with $o = o^*$, we disregard the log term and rewrite the remaining expression using the definition of conditional probability. Recall that $\{O\} = \mathbf{O}^S = \mathbf{O}^N$; we thus consider boundaries on α as defined by:

$$\alpha \stackrel{\text{def}}{=} \frac{\Pr^N_{\mathbf{c}}(o^*) \cdot \Pr^N_{\mathbf{c}}(\mathbf{i})}{\Pr^N_{\mathbf{c}}(o^* \wedge \mathbf{i})} = \frac{\Pr^N_{\mathbf{c}}(o^*)}{\Pr^N_{\mathbf{c}}(o^* \mid \mathbf{i})} \tag{2}$$

In general, the IOconfl measure can take on any value in the interval $(-\infty, \infty)$, and therefore $\alpha \in (0, \infty)$. Here we exclude the possibility of a degenerate 'prior' where $\Pr^N_{\mathbf{c}}(o^*) = 0$ or $\Pr^N_{\mathbf{c}}(o^*) = 1$, since in that case $\Pr^N_{\mathbf{c}}(o^* \mid \mathbf{i}) = \Pr^N_{\mathbf{c}}(o^*)$ for all $\mathbf{i}$. Given that $\Pr^N_{\mathbf{c}}(o^* \mid \mathbf{i}) \leq 1$, we now in fact find a lower bound: $\Pr^N_{\mathbf{c}}(o^*) \leq \alpha$.

To find an upper bound, we first consider the special case where the AI system and the normative model have the same distribution over the shared variables, and no context variables are observed. That is, $\Pr^N_{\mathbf{c}} = \Pr^N$ and $\Pr^N(o^* \mid \mathbf{i}) = \Pr^S(o^* \mid \mathbf{i}) \in [\frac{1}{r}, 1]$ too. We now find the following boundaries on α:

$$\alpha = \frac{\Pr^N(o^*)}{\Pr^N(o^* \mid \mathbf{i})} = \frac{\Pr^N(o^*)}{\Pr^S(o^* \mid \mathbf{i})} \in [\Pr^N(o^*), r \cdot \Pr^N(o^*)] \tag{3}$$

In this case, the conflict between o and $\mathbf{i}$ as computed from the normative model is equivalent to the conflict we would compute for the AI system, had we known the distribution $\Pr^S$. This is however not the aim of normative monitoring.

Upon including context we must generally assume that $\Pr^N_{\mathbf{c}} \neq \Pr^S$ and $\Pr^N_{\mathbf{c}}(o^* \mid \mathbf{i}) \neq \Pr^S(o^* \mid \mathbf{i})$. In fact, o^* need not be the most likely value of O given $\mathbf{i}$ according to $\Pr^N_{\mathbf{c}}$.[1] We will distinguish three cases where according to $\Pr^N_{\mathbf{c}}$ o^* is (1) *definitely not* the most likely value, (2) *not guaranteed* to be the most likely value, and (3) *definitely* the most likely value. These three cases, together with the associated range of posterior probabilities, are illustrated in Fig. 2.

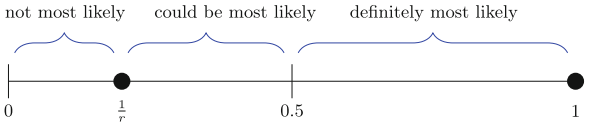

Fig. 2. The range of values for $\Pr^N_{\mathbf{c}}(o \mid \mathbf{i})$ for which o is or is not guaranteed to be the most likely value of O given $\mathbf{i}$ in context $\mathbf{c}$.

In the first case, we have that $\Pr^N_{\mathbf{c}}(o^* \mid \mathbf{i}) < \frac{1}{r}$. As a result, we may find values of $\alpha > r \cdot \Pr^N_{\mathbf{c}}(o^*)$, which means that the upper-bound in Eq. 3 may no longer hold and all we know is that $\alpha \in [\Pr^N_{\mathbf{c}}(o^*), \infty)$. In the second case, we have that $\Pr^N_{\mathbf{c}}(o^* \mid \mathbf{i}) \in [\frac{1}{r}, \frac{1}{2})$ and, as a result, $\alpha \in (2\Pr^N_{\mathbf{c}}(o^*), r \cdot \Pr^N_{\mathbf{c}}(o^*)]$. In this case we either have that o^* is the most likely value given $\mathbf{i}$ in $\Pr^N_{\mathbf{c}}$ too, or there exists an $o \in \Omega(O)$, $o \neq o^*$, with $\Pr^N_{\mathbf{c}}(o \mid \mathbf{i}) > \Pr^N_{\mathbf{c}}(o^* \mid \mathbf{i})$. Note that for binary-valued output variables ($r = 2$), this case does not exist. Finally, in the third case, $\Pr^N_{\mathbf{c}}(o^* \mid \mathbf{i}) \geq \frac{1}{2}$, resulting in $\alpha \in [\Pr^N_{\mathbf{c}}(o^*), 2\Pr^N_{\mathbf{c}}(o^*)]$. Here o^* is the most likely value (disregarding ties) given $\mathbf{i}$ in both $\Pr^N_{\mathbf{c}}$ and $\Pr^S$. Figure 3 summarises the intervals found for α in the different cases.

5 Choosing a Threshold

Recall that the IOconfl measure has an intrinsic threshold of $0 = \log 1$ which corresponds to $\alpha = 1$. We will now use our above analyses to propose an alternative threshold on α, and hence on IOconfl.

[1] Even without including context, there can be various reasons why o^* need not be the most likely value of O given $\mathbf{i}$ in $\Pr^N$, for one thing because the normative model is not designed to make predictions regarding the value of O.

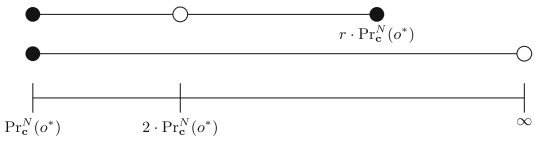

Fig. 3. Bounds on α, depending on $\Pr_{\mathbf{c}}^{N}(o^*)$ and $r = |\Omega(O)|$. The top line corresponds to cases (2) and (3); the second line to case (1). $r \cdot \Pr_{\mathbf{c}}^{N}(o^*)$ and $2\Pr_{\mathbf{c}}^{N}(o^*)$ coincide for binary-valued O.

From our analyses above we have that in the absence of context-specific information, $\alpha \in [\Pr^N(o^*), r \cdot \Pr^N(o^*)]$ for an r-ary output variable, under the assumption that $\Pr^N = \Pr^S$ (Eq. 3). Whether or not α can exceed the default flagging threshold of 1, therefore depends on the number of possible values of output variable O and the prior $\Pr^N(o^*)$. More specifically, α can only exceed 1 if $\Pr^N(o^*) > \frac{1}{r} > \Pr^N(o^* \mid \mathbf{i})$. That is, it can only flag cases for which the output from the AI system is a priori the most likely, or possibly most likely, according to both the AI system and the normative model (see Fig. 2), and becomes less likely upon observing input $\mathbf{i}$.

The above case captures a situation in which the normative model is not truly exploiting any context-specific information and hence does not add anything on top of what the AI system is doing. Assuming that the AI system is in essence an accurate model for the task it is designed to perform, we should therefore refrain from flagging in cases where $\Pr_{\mathbf{c}}^{N}$ and $\Pr^S$ agree. This suggests that an appropriate threshold on α for this case is $r \cdot \Pr_{\mathbf{c}}^{N}(o^*)$.

For the cases in which the provided context actually makes a difference in the normative model, we expect to find that $\Pr_{\mathbf{c}}^{N} \neq \Pr^S$. As a result, α can become larger than $r \cdot \Pr_{\mathbf{c}}^{N}(o^*)$, which happens when $\Pr_{\mathbf{c}}^{N}(o^* \mid \mathbf{i}) < \frac{1}{r}$, i.e. the normative model considers the combination $o^* \wedge \mathbf{i}$ less likely than the combination $o' \wedge \mathbf{i}$ for some $o' \in \Omega(O), o' \neq o^*$. In the given context, therefore, the output returned by the AI system may not be acceptable, which should be a reason for the monitoring system to flag. We note that differences between $\Pr_{\mathbf{c}}^{N}$ and $\Pr^S$ can of course also be due to the AI system and the normative model representing different joint distributions over their shared variables; however, this is not easily verified since $\Pr^S$ is in fact unknown to us.

Given the above, we propose to flag an input-output instance $o^* \wedge \mathbf{i}$ whenever $\alpha > r \cdot \Pr_{\mathbf{c}}^{N}(o^*)$, that is, for

$$\text{IOconfl}(o^*, \mathbf{i}) > \tau, \quad \text{where } \tau \stackrel{\text{def}}{=} \log(r \cdot \Pr_{\mathbf{c}}^{N}(o^*)) \tag{4}$$

Note that τ is in fact below the default threshold of 0 whenever $\Pr_{\mathbf{c}}^{N}(o^*) < \frac{1}{r}$.

The proposed threshold τ is a dynamic threshold, which depends on the output predicted by the AI system and additional context taken into account by the normative model. Since the normative model is a transparent BN, both the number of values r for O and $\Pr_{\mathbf{c}}^{N}(o^*)$ are known, so we can easily determine this context-specific threshold.

6 Experimental Evaluation

The adjusted conflict measure, IOconfl, and the corresponding dynamic threshold are specifically designed for monitoring input-output pairs in a context. In this experiment, we evaluate the flagging behaviour of different monitoring processes, that is, combinations of measures and thresholds, to qualitatively establish the impact of varying contexts.

6.1 Experimental Set-Up

For our experiment we need a normative model, an AI system, a monitoring process and test cases. As normative model we use an existing Bayesian network, the CHILD network[2] [10], which was manually elicited from medical experts at the Great Ormond Street Hospital for Sick Children in London, and developed for preliminary diagnosis of congenital heart diseases using information reported over the phone. In this network, we let $O = \{\texttt{Disease}\}$, $\mathbf{I} = \{\texttt{GruntingReport}, \texttt{CO2Report}, \texttt{XrayReport}, \texttt{LVHreport}\}$, and the context variables be $\mathbf{C} = \{\texttt{BirthAsphyxia}, \texttt{Age}\}$; $\mathbf{H}$ consists of the remaining 13 variables.

To simulate an AI system with $\mathbf{I}^S = \mathbf{I}$ and $O^S = O$, we construct a BN with $\Pr^S(O^S, \mathbf{I}^S) = \Pr^N(O, \mathbf{I})$ by using GeNIe[3] to marginalise out the variables $\mathbf{C} \cup \mathbf{H}$ from the original CHILD network. Although in practice the distributions $\Pr^S$ and $\Pr^N$ over the shared variables might not be exactly the same, they both approximate part of the target system and should therefore be rather similar. Assuming $\Pr^S$ and $\Pr^N$ to be equivalent in the experiment, allows us to evaluate the impact of the context on flagging behaviour in isolation.

The monitoring process computes the measure and decides for a given input-output instance and context whether or not to flag, based upon the thresholds; we implemented a script for these computations using SMILE. In this experiment, we compare three monitoring processes: $\mathcal{J}_0$, the original Jensen conflict measure with its intrinsic threshold of 0; $\mathcal{I}_0$, the IOconfl measure with the intrinsic threshold 0; and $\mathcal{I}_\tau$, the IOconfl measure with dynamic threshold τ.

As test cases we use 240 configurations from $\Omega(\mathbf{I}) \times \Omega(O) \times \Omega(\mathbf{C})$: each of the 40 possible value assignments $\mathbf{i} \in \Omega(\mathbf{I})$ is paired with its most likely value $o^* \in \Omega(O)$ according to the AI system ($\Pr^S$), and every resulting input-output instance is subsequently considered in each of 6 possible contexts $\mathbf{c}_k \in \Omega(\mathbf{C})$. For each of the 240 configurations, we compute both the IOconfl and the original conflict measure, as well as the dynamic threshold, using $\Pr^N_{\mathbf{c}_k}$ from the original CHILD network. Note that context is also included for the original conflict measure to enable a fair comparison. For both measures, we determined how many and which test cases were flagged using the intrinsic threshold and, for IOconfl, our dynamic threshold.

[2] Available from https://www.bnlearn.com/bnrepository/discrete-medium.html.

[3] The experiment was executed using the GeNIe Modeler and the SMILE Engine by BayesFusion, LLC (http://www.bayesfusion.com/).

Table 1. Number of contexts in which a specific input-output instance is flagged by a process; the remaining $40 - 13 = 27$ instances are never flagged.

Instance ID	1	2	3	4	5	6	7	8	9	10	11	12	13	# cases
Process $\mathcal{J}_0$	2	4	1	0	2	0	0	1	6	6	4	1	1	28
$\mathcal{I}_0$	0	0	3	1	0	3	1	0	0	0	0	3	1	12
$\mathcal{I}_\tau$	0	0	0	1	1	0	1	0	0	1	0	0	1	5

6.2 Results and Discussion

For the three monitoring processes $\mathcal{J}_0$, $\mathcal{I}_0$ and $\mathcal{I}_\tau$, we find the following flagging results for the 240 test cases. There are 38 cases in which at least one monitoring process flags: process $\mathcal{J}_0$ flags 28 times; process $\mathcal{I}_0$ flags 12 times; and process $\mathcal{I}_\tau$ flags five times (see Table 1). Six cases are flagged by two processes and only a single case is flagged by all three monitoring processes. We conclude that *what* these monitoring processes measure and flag differs notably. In addition, we note that the frequency with which they flag differs: the original conflict measure flags far more often than the IOconfl measure, even when using the same intrinsic threshold, and the fewest cases are flagged with the dynamic threshold.

To consider the effect of context, we look at which instances were flagged and in which contexts. A total of 13 input-output instances are flagged by the monitoring processes, in at least one context (see Table 1). For each 0 and 6 in Table 1 the context did not matter, since a process either flags in none of the contexts or in all. In all other cases, we find that context affects the flagging behaviour. Let $F(\mathcal{I}_\tau) = \{4, 5, 7, 10, 13\}$ denote the set of all instances (IDs) flagged by process $\mathcal{I}_\tau$, and let $F(\mathcal{I}_0)$ and $F(\mathcal{J}_0)$ be likewise defined. We then find that none of these three sets is a subset of either of the other two sets, and that each set partly overlaps with both other sets. This shows that the three processes truly differ in the way they take context into account for a given instance. Consider e.g. the instance `GruntingReport` = `no` ∧ `CO2Report` = `x7_5` ∧ `XrayReport` = `Asy_Patchy` ∧ `LVHreport` = `no` ∧ `Disease` = `Fallot` (ID 10 in Table 1), this instance is flagged in all six contexts by process $\mathcal{J}_0$, and in none of the contexts by process $\mathcal{I}_0$. However, IOconfl in combination with the dynamic threshold (process $\mathcal{I}_\tau$) flags in one specific context only (see Table 2). We conclude that for this input-output instance only process $\mathcal{I}_\tau$ flags context-specifically. It indicates that for babies younger than 3 days with birth asphyxia, diagnosing fallot should be questioned, despite the input indicating this output.

Table 2. Example of flagging behaviour of the monitoring processes for one input-output instance (ID 10) in six different contexts.

Context	Age	x_3_days	x_3_days	x_10_days	x_10_days	x1_30_days	x1_30_days
Variables	BirthAsphyxia	yes	no	yes	no	yes	no
Flagged by	$\mathcal{J}_0$	yes	yes	yes	yes	yes	yes
	$\mathcal{I}_0$	no	no	no	no	no	no
	$\mathcal{I}_\tau$	yes	no	no	no	no	no

Note that we cannot determine whether any combination of measure and threshold is better than another from this experiment, since we have no ground truth available. Such an assessment would require insight into the quality of the CHILD network as well as the expertise of a paediatric cardiologist.

Overall, we conclude that the choice between measures and thresholds matters and is not trivial, and that the IOconfl measure in combination with the dynamic threshold seems to be a conservative combination that evidently succeeds in flagging context-specifically.

7 Conclusion and Future Research

In monitoring processes, any measure must be accompanied by a threshold in order to determine whether to flag an observed instance. We considered several measures for flagging input-output instances from an AI system in a normative monitoring setting. In particular, we reconsidered the default threshold for the BN-specific IOconfl measure and studied the measure's boundary conditions to arrive at a new dynamic threshold. This dynamic threshold depends on the context in which the input-output pair of the AI system is observed and the distribution over the output variable, both according to the normative model. As such it is capable of taking context of use into account, as intended. We compared the use of the IOconfl measure with both default and dynamic thresholds in a small controlled experiment; in addition, we compared the IOconfl measure against the original conflict measure from which it was derived. We found that each combination of measure and threshold results in different flagging behaviour, confirming that decisions about which to use are indeed not trivial.

The actual choice for a suitable measure and threshold will depend on the domain and the costs of false positive and false negative warnings. We can therefore not conclude that one is necessarily better than the other. In future research, we would like to further study theoretical differences between the measures and evaluate their use with domain experts for realistic tasks. Moreover, we can study to what extent the attribution method by Kirk et al. [5] can be employed to explain the reason for flagging in terms of violation of the rule, protocol or other type of norm captured by the normative modelled. Finally, to fulfill all steps in our framework for monitoring under uncertainty, future research is necessary into methods for eliciting norms from domain experts and for capturing these in models such as Bayesian networks.

Acknowledgements. This research was supported by the Hybrid Intelligence Centre, a 10-year programme funded by the Dutch Ministry of Education, Culture and Science through the Netherlands Organisation for Scientific Research, https://hybrid-intelligence-centre.nl.

References

1. Chandola, V., Banerjee, A., Kumar, V.: Anomaly detection: a survey. ACM Comput. Surv. **41**(3), 1–58 (2009)

2. Jensen, F.V., Chamberlain, B., Nordahl, T., Jensen, F.: Analysis in HUGIN of data conflict. In: Proceedings of the Sixth Conference on Uncertainty in Artificial Intelligence, pp. 546–554 (1990)
3. Jensen, F.V., Nielsen, T.D.: Bayesian Networks and Decision Graphs, 2nd edn. Springer, New York (2007). https://doi.org/10.1007/978-0-387-68282-2
4. Johansson, F., Falkman, G.: Detection of vessel anomalies - a Bayesian network approach. In: Proceedings of the Third International Conference on Intelligent Sensors, Sensor Networks and Information, pp. 395–400. IEEE (2007)
5. Kirk, A., Legg, J., El-Mahassni, E.: Anomaly detection and attribution using Bayesian networks. Technical report, Defence Science and Technology Organisation Canberra (2014)
6. Kjaerulff, U.B., Madsen, A.L.: Bayesian Networks and Influence Diagrams: A Guide to Construction and Analysis, 2nd edn. Springer, New York (2013). https://doi.org/10.1007/978-1-4614-5104-4
7. Nielsen, T.D., Jensen, F.V.: On-line alert systems for production plants: a conflict based approach. Int. J. Approximate Reasoning **45**, 255–270 (2007)
8. Onnes, A.: Monitoring AI systems: a problem analysis, framework and outlook. In: Proceedings of the First International Conference on Hybrid-Artificial Intelligence. Frontiers in Artificial Intelligence and Applications, vol. 354, pp. 238–240 (2022)
9. Onnes, A., Dastani, M., Renooij, S.: Bayesian network conflict detection for normative monitoring of black-box systems. In: Proceedings of the Thirty-Sixth FLAIRS Conference, vol. 36. Florida Online Journals (2023)
10. Spiegelhalter, D.J., Dawid, A.P., Lauritzen, S.L., Cowell, R.G.: Bayesian analysis in expert systems. Statist. Sci. **8**, 219–247 (1993)