Utrecht University

Bayesian Network Conflict Detection for Normative Monitoring of Black-Box Systems

Annet Onnes, Mehdi Dastani, Silja Renooij

Overview

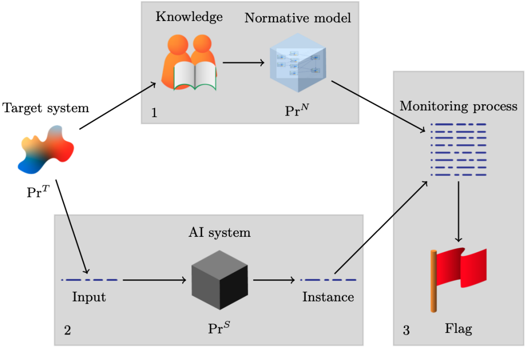
1. Overview of the setting

2. Bayesian Network Conflict Detection for Normative Monitoring
    - 2.1 Background
    - 2.2 Monitoring an Black-Box AI System
    - 2.3 Using conflict measures

3. Constructing Bayesian Networks Normative Models

Monitoring a Black-Box AI System

Overview of Normative Monitoring setting



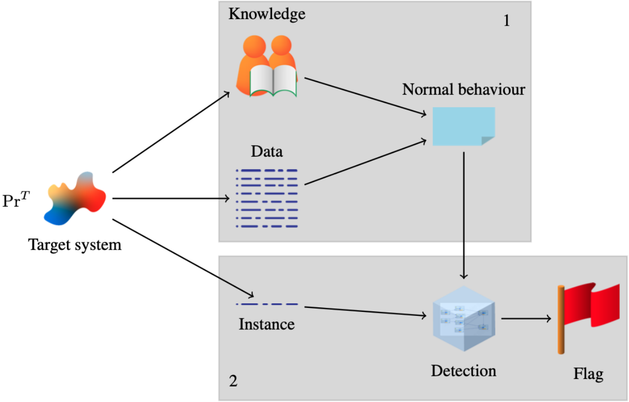
Bayesian Network Conflict Detection for Normative Monitoring

Background

- Motivation: We need to monitor operations to ensure AI technology is safe and reliable.
- Techniques: Anomaly Detection using Bayesian Networks [1, 3]

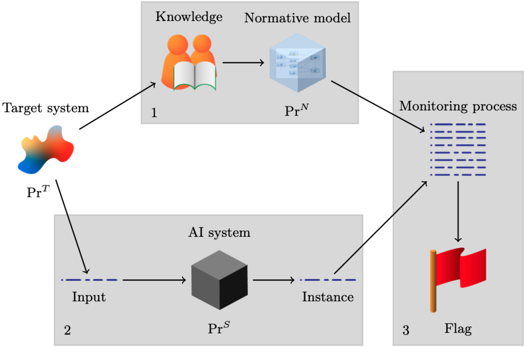
Anomaly Detection

Quick Impression



Monitoring a Black-Box AI System

Overview of Normative Monitoring setting



Detecting unacceptable input-output pairs

Conflict Measure

$$\mathrm{confl}(e_1, \ldots, e_t) = \log \frac{\Pr(e_1) \cdot \ldots \cdot \Pr(e_t)}{\Pr(\mathbf{e})} \tag{1}$$

Introduced by Jensen et al. [2].

Detecting unacceptable input-output pairs

Adjusting the conflict measure

Using the distribution from the normative model and given the context, Pr is $\mathrm{Pr}^N(\cdot \mid \mathbf{a}')$, $\mathrm{Pr}^N_{\mathbf{a}'}(\cdot)$ abbreviated.

$$\begin{aligned} \mathrm{IOconfl}(o, \mathbf{i}) &= \mathrm{confl}(o, i_1, \ldots, i_n) - \mathrm{confl}(i_1, \ldots, i_n) \\ &= \log \frac{\mathrm{Pr}(o) \cdot \mathrm{Pr}(\mathbf{i})}{\mathrm{Pr}(o \wedge \mathbf{i})} \end{aligned} \tag{2}$$

Defining a Threshold for Conflict Detection

- To flag using any measure a threshold is needed.
- Both the original and adjusted conflict measure have a intrinsic threshold at 0.

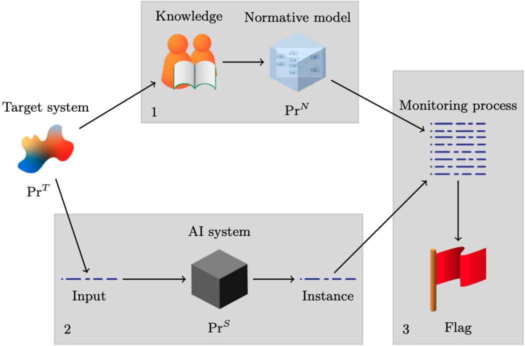
Defining a Threshold for Conflict Detection

A Dynamic Threshold

- After analysing the bounds on the measure we determined limitations on the intrinsic threshold.
- $$\mathrm{IOconfl}(o^*, \mathbf{i}) > \tau, \quad \text{where } \tau \overset{\text{def}}{=} \log(r \cdot \Pr(o^*)) \tag{3}$$

Thanks for your attention

Any questions so far?



Current Research

Constructing BNs for Normative Monitoring

- Taking inspiration from knowledge elicitation for Bayesian networks.
- Translating the expectations of acceptable behaviour into the Bayesian network.

Responsible Hybrid Intelligence

Discussion

- Ensuring Responsible HI: What do we want to monitor for?
- How do these expectations arise in context?
- Aim of using BNs is increasing transparency and interpretability

Bibliography

[1] Varun Chandola, Arindam Banerjee, and Vipin Kumar. "Anomaly Detection: A Survey". In: *ACM Computing Surveys* 41.3 (2009), pp. 1–58.

[2] Finn Verner Jensen et al. "Analysis in HUGIN of Data Conflict". In: *Proceedings of the Sixth Conference on Uncertainty in Artificial Intelligence*. 1990, pp. 546–554.

[3] Andrew Kirk, Jonathan Legg, and Edwin El-Mahassni. *Anomaly Detection and Attribution Using Bayesian Networks*. Tech. rep. Defence Science and Technology Organisation Canberra, 2014.