Normative Monitoring of Black-Box AI systems using Bayesian Networks

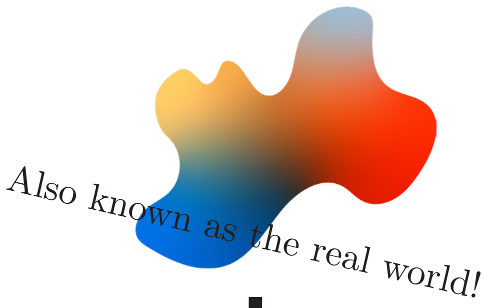
Utrecht University

#13

Annet Onnes

Supervised by: Silja Renooij and Roel Dobbe

Target System

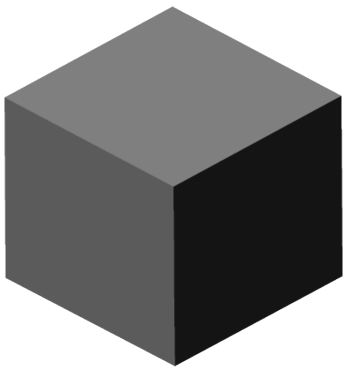


Knowledge

Human experts can observe the target system and establish (uncertain) knowledge about what normative behaviour of the real world system looks like.

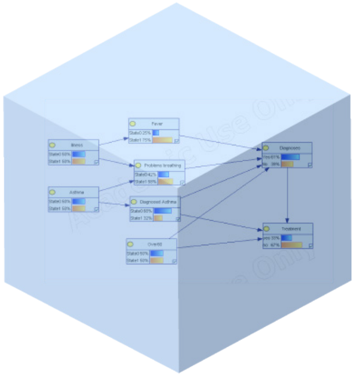
Bayesian networks (BNs) can be handcrafted by eliciting knowledge from experts (users and stakeholders).

AI system



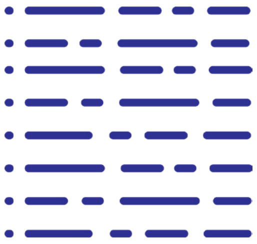
Only in- and output are available of the one instance we have to decide about.

Normative model

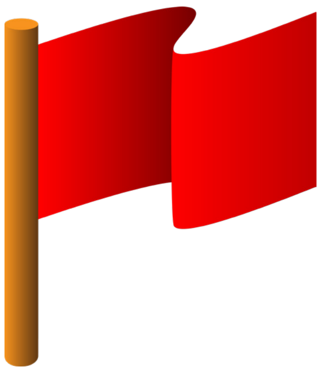


Normative behaviour is represented in a normative model using a BN.
This has to capture requirements for a specific context, and should be interpretable and transparent, like a Glass Box.

Monitoring Process



To monitor, a measure is needed which indicates to what extend the AI systems in- and output is adhering to the norms in a particular context.



In normative monitoring an input-output pair of the AI system is flagged when it is detected the behaviour is not as it ought to be according to the norms in the normative model for that context.

Modelling user requirements in BNs

BNs can be constructed using human expert knowledge. The aim is to develop a method to translate this knowledge about the requirements, expectations and protocols for the AI systems *in a specific context.*

The BN - that is the normative model - has to model required behaviour using context additional to the in- and output of the AI system.

Defining measures and thresholds

Inspired by Anomaly Detection literature, we adjusted a conflict measure to work for this novel monitoring setting:

$$\mathrm{IOconfl}(o, \mathbf{i}) = \log \frac{\Pr(o) \cdot \Pr(\mathbf{i})}{\Pr(o \wedge \mathbf{i})}$$

We also reconsidered what threshold to use and developed a dynamic threshold:

$$\tau = r \cdot \Pr(o^*)$$