HHAI2022: Augmenting Human Intellect
S. Schlobach et al. (Eds.)
© 2022 The authors and IOS Press.
This article is published online with Open Access by IOS Press and distributed under the terms of the Creative Commons Attribution Non-Commercial License 4.0 (CC BY-NC 4.0).
doi:10.3233/FAIA220202

Monitoring AI Systems: A Problem Analysis, Framework and Outlook

Annet ONNES [a,1,2]
[a] *Information and Computing Sciences, Utrecht University, The Netherlands*

In order to collaborate or interact with AI systems we have expectations about their behaviour. In this project we present a novel framework to study how to monitor AI systems in such a way that the expectations for the performance are formulated into an interpretable, knowledge-based system for monitoring. Monitoring is comparing the in- and output of the AI system, i.e. observed model behaviour, to intended behaviour. We have to consider how to represent this intended behaviour because modelling the intended behaviour using data would result in another model of a target system similar to the AI system's model. We suggest a knowledge-based model constructed using the prescriptions is more effective.

The problem setting consists of an AI system, which has to be monitored because there are expectations about its functioning. The AI system has to be monitored while it is running, since the expectations can be context specific. An impression of this is given in Figure .

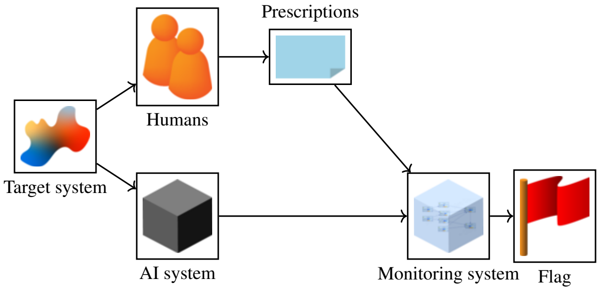

Figure 1. Impression of the setting

The *target system* is the process the *AI system* attempts to model [1]. The focus here is on statistical models, trained using parts of the data from the target system that are observable. *Prescriptions* are expectations about the output, that is the behaviour of the AI system, as defined by humans interacting with the system. This is information that is not included in the AI system because it can be context specific. Output that is accurate according to the AI system and that also adheres to the prescription is referred to as *intended behaviour*. If this is *not* the case, the output is *flagged*. A *monitoring system* would be designed to use the prescriptions in order to monitor the AI systems behaviour.

[1] Corresponding Author: Annet Onnes; E-mail: a.t.onnes@uu.nl.
[2] This research was funded by the NWO Hybrid Intelligence Center.

It is one thing to monitor an AI system to see whether expectations are fulfilled, but another to make decisions about what the expectations are and thus whether the output is ‘good’ or ‘bad’. The monitoring system is not defined as deciding what output is *wrong*; Prescriptions are determined externally, meaning monitoring itself is a technical operation devoid of assigning value. In social situations *norms* are expectations of behaviour. A norm is characterised as a particular rule or normative principle that is accepted in a particular group or community [2] or as “an expected behaviour in a social setting” [3]. *Norm monitoring* is then described as “determining whether an individual is adhering to this expected behaviour”. This fits the characterisation of monitoring. Norms are however only suggested as a potential partial representation of intended behaviour because they are indicators of the expectations of humans. They do not dictate what is ’good’, define a strict form of representation or fully model intended behaviour.

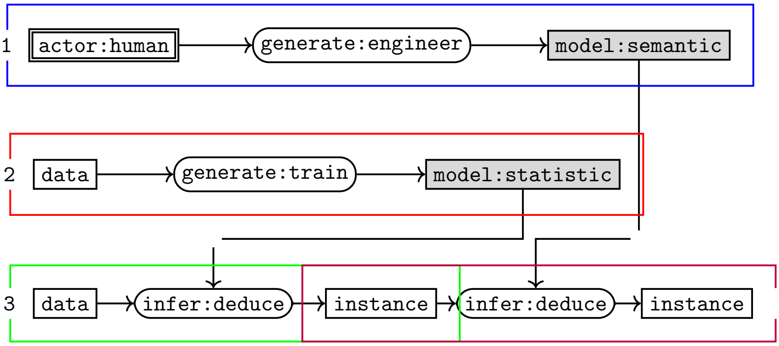

Figure 2. The full monitoring setting with AI system and monitoring system. Marked are four individual sections which are elementary patterns. This framework represents all elements of the setting that are designed systems with in- and outputs, this does not include the target system.

Using the design patterns from Bekkum et al.[4], we can formally depict in Figure 1 the systems and processes of the monitoring setting. Box 1 describes the process through which prescriptions are modelled from human knowledge. The resulting semantic model informs the monitoring system. Box 2 describes training a statistical model from data. `model:statistic` is the model used by the AI system represented by pattern 3. Box 4 is the monitoring process, a deductive process.

Thusfar we only have the stepping stones from which to continue research in this monitoring setting. This analysis helps to identify the questions: *how to do the comparison between the intended behaviour and the observed AI system behaviour?* and *how to represent the intended behaviour using prescriptions?* Like with most research there are a vast amount of possible research directions and factors to take into consideration. Further work will have to look into uncertainty, comparing distributions and learning AI systems. Using the abstract framework of the setting presented here these challenges can be formulated in a more structured manner. The next step in this endeavour will also include finding technical representations and notations. For a further exploration of future research possibilities, as well as a more thorough analysis, see [5].

References

[1] Frigg R, Hartmann S. Models in Science. In: Zalta EN, editor. The Stanford Encyclopedia of Philosophy. Spring 2020 ed. Metaphysics Research Lab, Stanford University; 2020. .

[2] Brennan G, Eriksson L, Goodin RE, Southwood N. Explaining Norms. Oxford University Press; 2013.

[3] Dastani M, Torroni P, Yorke-Smith N. Monitoring Norms: a Multi-Disciplinary Perspective. The Knowledge Engineering Review. 2018;33:1–22.

[4] van Bekkum M, de Boer M, van Harmelen F, Meyer-Vitali A, Teije At. Modular design patterns for hybrid learning and reasoning systems: a taxonomy, patterns and use cases. Applied Intelligence. 2021;51(9):6528–6546.

[5] Onnes A. Monitoring AI systems: A Problem Analysis, Framework and Outlook. arXiv; 2022. Available from: https://arxiv.org/abs/2205.02562.