2.8 Interpretability requirements

Practical guidance – cross-domain

Authors: Rhys Ward

The need for interpretability in safety critical domains

'Interpretability is used to confirm other important desiderata of ML systems...' [1]. Ultimately Interpretability is one way in which we can provide assurance that systems will be sufficiently safe. Different projects may need their systems to be interpretable for different reasons; one project may need to be able to explain events (e.g. accidents) after they occur, for legal reasons; for another project an accident may be impermissible and so transparency which allows future model behaviour to be predicted may be prioritised.

Interpretability may:

- Increase insight into model behaviour (and also into the operational domain).
- Identify weaknesses of the model, known cases where the model under-performs.
- Enable the increase of robustness - i.e. assurance that the system will behave safely in new environments/situations.
- Inform contestability and allow effective improvements/corrections.
- Allow prediction of future model behaviour and avoid future harmful decisions.
- Aid in understanding mistakes/accidents/anomalies
- Protect against unfair models helping to avoid discrimination.
- Improve trust in the model and allow informed consent [3]
- Help us to explain individual (local) decisions and answer the question 'What were the important factors in this decision?'

We will now discuss **explanations** of system behaviour. An explanation may be the model itself, if the model is transparent and easily understood; or an explanation may be an approximate model. However, these explanations would not be suitable for a non-technical audience, when using the term 'explanation' here we refer to any format which an interpretation of the system behaviour may take.

WHAT to explain

The different types of interpretability result in the interpretability or explanation of a set of distinct things. Transparency may refer to: the transparency of the whole model, wherein the entire global logic of the model can be explained and understood by a human; the transparency of the learning algorithm, we may understand that some algorithms converge to a solution in reasonable time (e.g. linear models), whereas we may not know whether another algorithm does find a solution or not (e.g. neural networks) [2]; transparency of parameters and model structures, do we understand what these are referring to and do they even map to human-understandable concepts? Similarly post-hoc explainablity methods may try to explain and interpret these things, e.g. through approximating the global logic of a model, or they may explain local decisions.

So, the following needs to be explained:

- The **learning algorithm** which produces the decision-making model
- The global **model logic** (either through transparency or approximation methods)
- The **system logic** as a whole
- **Local decisions** i.e. how (and which) specific inputs relate to an output
- How **parameters and model structures** relate to human-understandable concepts

WHEN explanations are needed

Explanations will be needed for different reasons during development and operation. ML developers may seek global explanations to better understand the logic of the model to help in design; stakeholders will need different types of explanations during operation (perhaps local explanations will be more important during operation to explain individual cases - e.g. when explaining why an accident occurred).

- **Development**
    - **Data management** - interpreting the model may identify weaknesses/gaps in the data.
    - **Model selection** - the interpretability of a model should influence this stage.
    - **Model learning** - being able to interpret the model will inform the model learning stage, e.g. in aiding hyper-parameter selection. Interpretability will also aid understanding of the operation domain.
    - **Model verification** - being able to explain model decisions will aid verification and help to identify the cause of model flaws/weaknesses.
- **Operation**
    - **Normal operation** - e.g. for advisory systems such as diagnostic tools explanations may be compulsory.
    - **In cases where the model is known to under perform** - which will enable contestability.
    - **Accident or incident investigation** - local explainability to discover why decisions were made.
    - **Model run-time improvement/learning** - to improve models as new data and situations are encountered.

WHO explanations are for

Different stakeholders need different types of explanations, lay users, expert users, designers. Developers need explanations and transparency to understand how the model works in order to predict when undesirable model behaviour will occur and make corrections and improvements. Whilst developers may need some local explainability to understand and account for edge cases, in general they will need global interpretability to aid design. End-users will need local explanations to satisfy understanding of individual decisions. Figure 1 shows potential stakeholders and the explanation needs for each.

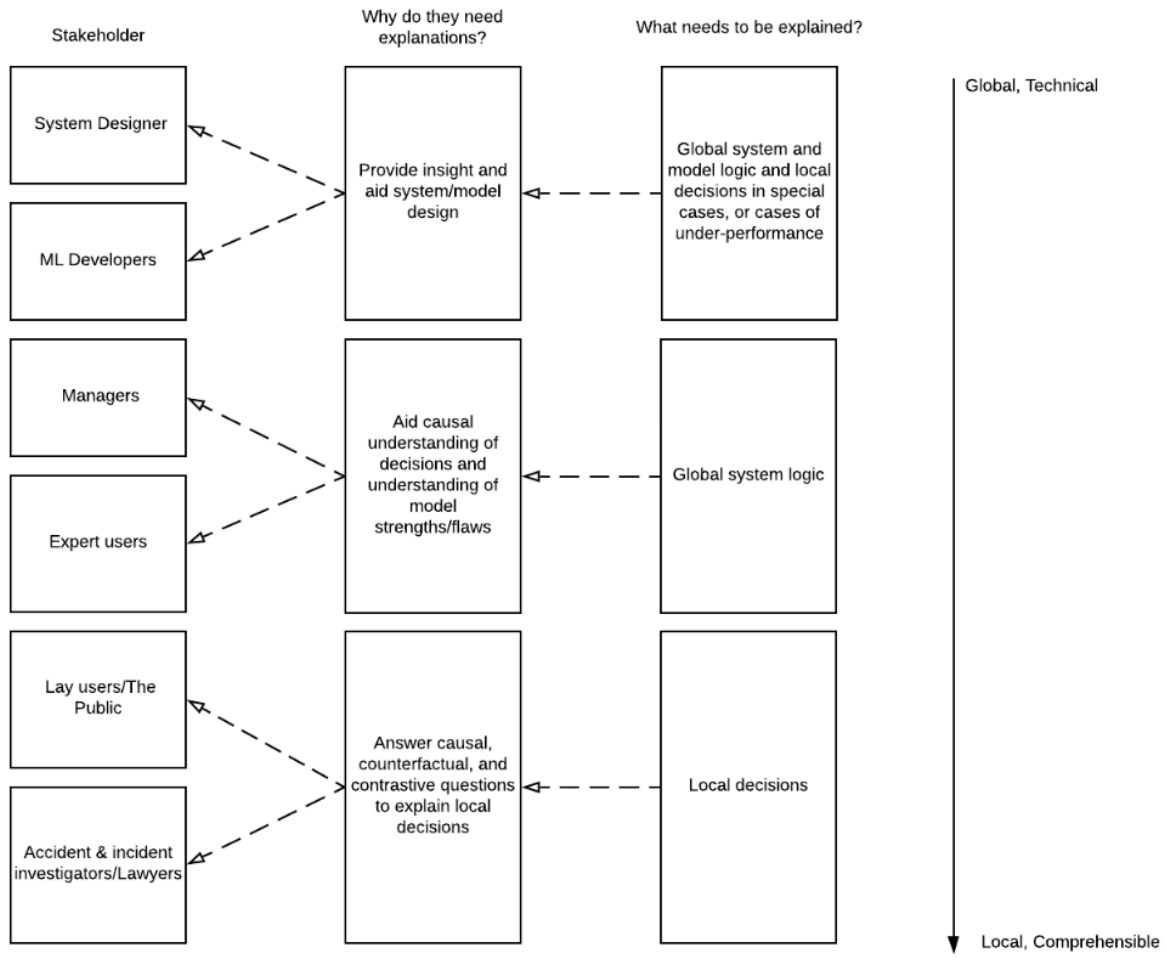

Figure 1 - Explanation needs for different stakeholders

References

- [1] Finale Doshi-Velez and Been Kim. “Towards A Rigorous Science of Interpretable Machine Learning”. In: arXiv:1702.08608v2 [stat.ML] 2Mar 2017 (2017).
- [2] Zachary C. Lipton. “The Mythos of Model Interpretability”. In: arXiv:1606.03490v3 [cs.LG] (2017).
- [3] David Watson et al. “Clinical applications of machine learning algorithms: beyond the black box”. In: BMJ. 2019.