phg foundation
making science work for health

Black box medicine and transparency

Executive summary

A PHG Foundation report for the Wellcome Trust

UNIVERSITY OF CAMBRIDGE

Executive summary

The series of reports *Black Box Medicine and Transparency* examines the human interpretability of machine learning in healthcare and research:

1. *Machine learning landscape* considers the broad question of where machine learning is being (and will be) used in healthcare and research for health
2. *Interpretable machine learning* outlines how machine learning can be or may be rendered human interpretable
3. *Ethics of transparency and explanation* asks why machine learning should be made transparent or be explained, drawing upon the many lessons that the philosophical literature provides
4. *Regulating transparency* considers if (and to what extent) does the General Data Protection Regulation (GDPR) require machine learning in the context of healthcare and research to be transparent, human interpretable, or explainable
5. *Interpretability by design framework* distils the findings of the previous reports, providing a framework to think through human interpretability of machine learning in the context of healthcare and health research
6. *Roundtables and interviews* summarises the three roundtables and eleven interviews that provided the qualitative underpinning of preceding reports

Each report interlocks, building on the conclusions of preceding reports. The following provides a high-level summary of each report and its conclusions.

The machine learning landscape

Where is machine learning being (and will be) used in medical research and healthcare?

Machine learning is a subset of artificial intelligence. Machine learning models are not explicitly programmed but are trained with many examples (data) relevant to a task, allowing them to develop rules to automate the process.

Machine learning has a range of applications across both medical research and healthcare. Some applications are already in use or near implementation, others are more speculative.

In medical research, it is being used for tasks such as systematic literature reviews or selecting potential participants for clinical trials; in healthcare, applications range right across the patient pathway from early detection, to stratification, diagnosis, treatment, and beyond, as well as administrative tasks. Many machine learning applications for medical research will blur the lines between research and clinical care.

The machine learning field is focusing increasingly on the replicability and generalisability of models, especially in the context of healthcare.

There has been a proliferation of related policy guidance and regulation that applies to the use of machine learning in medical research and healthcare, varying in specificity and statutory weight. Most highlight transparency as being a key principle underpinning responsible development of AI.

Interpretable machine learning

What is human interpretability of machine learning? How may machine learning models be rendered human interpretable?

The term 'black box' can refer to opacity that is due to the model being inherently uninterpretable to humans, uninterpretability due to the user lacking technical expertise, or because of information being restricted due to commercial sensitivity.

'Human interpretability' of machine learning (opacity due to the model being inherently uninterpretable) is a complex concept with three main dimensions:

Algorithmic transparency relates to how an algorithm learns relationships between data and produces the eventual trained machine learning model.

Global interpretability concerns the ability to 'understand the whole logic of a model and follow the entire reasoning leading to all different possible outcomes.'[i]

Local interpretability refers to the ability to understand 'only the reasons for a specific decision' made by a machine learning model.[ii]

There are many methods to make otherwise uninterpretable machine learning models interpretable, different methods suiting different dimensions and better serving different purposes. For instance, ensuring a model is interpretable may facilitate successful interaction between the model and user, underpin confidence in the model's outputs, or ensure accountability.

There are different methods to make medical machine learning models interpretable. For instance, interpretable models can be trained side by side with opaque models to infer how the opaque model operates, visualisations can demonstrate model function, and a variety of other 'post hoc explainers' such as LIME can approximate model function. Each method has its own strengths and weaknesses, for example, post hoc explainers may not faithfully represent the model they represent but they allow the developer to pick a machine learning model irrespective of how interpretable that model might be.

There are three broad methods to evaluate the interpretability of machine learning models: functionally-grounded evaluation, human-grounded evaluation, and application-grounded evaluation – each more rigorous, although more demanding than the last. Evaluating the interpretability of models is likely to differ depending on how developed the application is, the task the model is intended to assist with, and the risk profile of this application.

Ethics of transparency and explanation

Why should machine learning be transparent or be explained? What lessons can be drawn from the philosophical literature on transparency and explanation?

The concept of 'transparency' is best analysed as not necessarily including ideas of accessibility, communication, or interpretability. These ideas are highlighted in the related philosophy on 'explanation' and pragmatist accounts of this concept. In the health context, trust has also been highlighted as an important concept. Transparency should be viewed as a means to secure trustworthiness, not trust.

There is a significant literature on the concept of 'explanation.' An explanation consists of the thing to be explained (the *explanandum*) and the thing that does the explaining (the *explanans*). Clarifying the precise question being asked, and ensuring that there is a match between these elements will be key in how satisfactory the resulting explanation will be. In the context of scientific and medical research, understanding is typically the end goal of explanation for machine learning. In healthcare, the most common form of explanation sought will not be scientific, but an everyday causal explanation of what led to a particular decision or outcome.

We can distinguish between correct explanations that are true and good explanations, where the explanation is appropriate to the interests and requirements of the audience being addressed and context. The philosophical literature on explanatory pragmatism suggests that there is probably no single archetypically good explanation; how satisfactory an act of explanation is, can be measured against the following criteria:

- Answering the audience's specific question
- Selecting the closest comparator (fact-foil combination) to approximate the specific explanatory information the audience seeks
- Ensuring sufficient explanatory information is provided to satisfy the questioner
- Using concepts and phenomena that the audience is already familiar with to explain unfamiliar concepts or phenomena

Context will be indispensable when developing satisfactory explanations of machine learning models; there will often be distinct but overlapping purposes behind explanations of machine learning for health or research:

A. Interpretability to evidence the safety and effectiveness of a system
B. Interpretability to facilitate human-computer interaction
C. Interpretability to assist in scientific or causal understanding
D. Interpretability to underpin control by the data subject or controller accountability

The prominence of each purpose will vary according to context. The *Interpretability by Design Framework* highlights the relative importance of each purpose for different audiences and contexts.

Regulating transparency

> *Does (and if so, to what extent) the GDPR require machine learning in the context of healthcare and research to be transparent, interpretable, or explainable?*

The GDPR is only one source of regulation that might generate a duty of transparency or explanation, forming part of a complex web of regulation that protects personal data. Accordingly, the GDPR is no panacea for transparency but offers only data protection solutions to protect data protection interests and values.

The GDPR is limited by its material scope ('what' it applies to), namely 'personal data' , which may be used by machine learning in different ways, for instance: as a part of the training/test datasets, as an input into a model, and as the output of a model.

The GDPR is also limited by its territorial scope ('where' the GDPR applies). This scope is broad – potentially applying to establishments both within, and outside, the European Economic Area (EEA), where the data processing is 'within their scope of activities' but also where they 'offer goods and services' or 'monitor' individuals within the EEA. Consequently, machine learning models that are trained on non-EEA datasets or use 'personal predictors to provide person recommendations' for EEA individuals, may have to comply with the GDPR's requirements.

The GDPR includes three interrelated claims that could together generate a duty of transparency, interpretability, or explainability:

A. The general principle of transparent processing
B. The interaction of this principle with specific data subject rights (for instance, the rights to information, of access, rectification, and to object)
C. The specific automated individual decision-making requirements

The general principle of transparent processing is context-specific and user-centric. It requires data controllers to consider the form in which they communicate (accessibility, simplicity, and intelligibility) as well as the content. For example, Recital 60 clarifies that controllers should: 'provide the data subject with any further information necessary to ensure fair and transparent processing taking into account the specific circumstances and context in which the personal data are processed.' This places a triple obligation on controllers, requiring that they comply with the principle when communicating with data subjects, disclose information required under the rights to information, and facilitate other data subject rights found in Articles 15-22.

Depending on the context, data subject rights may be qualified, restricted, or excluded (derogated from). In the context of healthcare and research, four general restrictions to data subject rights and data protection principles are particularly relevant:

1. Where the controller is no longer in a position to identify the data subject (Articles 11 and 12(2))
2. If Member States apply further restrictions, in this context with respect to health data (Article 23(1))
3. For scientific research (Article 89)
4. To prevent the disclosure of trade secrets and intellectual property (Recital 63 and Article 23(1)(i)).

The rights to information, of access and portability (provisions relating to automated individual decision-making aside) generally require little interpretability or explanation in order to be vindicated, although the right to access is recurrent, being available at 'reasonable intervals' throughout the lifecycle of data processing, potentially requiring some explanatory material. Arguably, the rights to rectification and to object, although highly limited, may require interpretability if they are to be fully upheld. Successful attempts to leverage interpretability or explanation will likely invoke multiple rights in combination with the general principle of transparency,

The GDPR's special provisions on automated individual decision-making are the most prominent tools used to construct a right to explanation. Two broad questions arise: first, when is the right to explanation triggered and second, what does the right require once triggered?

The right may be triggered by a variety of provisions which support a right to explanation (spread across Article 22, Recital 71, Articles 13(2)(f), 14(2)(g), and 15(1)(h)). However, there is a lack of consensus about how these should be interpreted.

Article 22(1) captures only a narrow range of data processing, providing a right to explanation only where 'a decision based solely on automated processing, including profiling, which produces legal effects concerning him or her or similarly significantly affects him or her.' It is manifestly unclear what counts as 'a decision' and how to frame such 'decisions' in the context of healthcare and research where it is common to have strings of decisions rather than just one. The wording 'based solely on automated processing' excludes processes where a human is involved with meaningful, authoritative input. Most health professionals will meet this threshold in the near-term, since most machine learning for healthcare and research is assistive, requiring healthcare professionals to contextualise and interpret its results.

'Legal effect' relates to a change in legal rights, status, or rights under contract for a data subject. The term is inherently unclear but in the context of machine learning for medical applications, systems that approve or deny a social benefit (including healthcare) may count as having 'legal effect.' Interpreting the phrase 'similarly significant effect' is also challenging with the addition of 'similarly.' Nevertheless, the more core the interest at stake, the more likely the decision will have 'similarly significant effect.'

If triggered, the right requires three elements:

'Meaningful information about the logic involved' as applied to machine learning may require the disclosure of a variety of information and probably requires a user-centric, layered approach. Accordingly, there may not be a one-size-fits-all approach to render machine learning interpretable.

'Significance and the envisaged consequences' appears to require some idea of how inputs into the model influence its outputs and any eventual decision. In the context of machine learning, this may be difficult as there is often not a linear relationship between an input and a particular output.

The right to contest (under Article 22(3)) may add extra interpretative depth to any right to explanation, perhaps requiring disclosure of information to allow data subjects to interrogate the model for fairness.

Any right to explanation may require explanation prior to and after processing, which could in turn necessitate both global interpretability of the overall machine learning model and also local interpretability of particular instances of processing. If a controller is caught by these provisions on automated individual decision-making, the added requirements of Article 22(3)-(4) also narrow and complicate the legal position of the controller, especially if the controller processes special category data (and so is subject to even more restrictions).

Interpretability by Design framework

Building on the ethical and legal analysis, a new *Interpretability by Design framework* (ID framework) assists developers to think through interpretability of machine learning models intended for medical applications. As an aid to good practice, this free framework enables the systematic review of various dimensions of the proposed tool and its application via a set of seven principles and seven steps to assist developers to consider the interpretability of their machine learning model.

References

[i] Guidotti R, Monreale A, Ruggieri S, et al. A survey of methods for explaining black box models. *ACM Computing Surveys*. 2018; 51(5): 5-6

[ii] Ibid, 5-6.

The Black box medicine and transparency report was funded by the Wellcome Trust as part of the 2018 Seed Awards in Humanities and Social Sciences [Grant Number: 213623/Z/18/Z].

We thank the Wellcome Trust for their support.

The PHG Foundation is a non-profit think tank with a special focus on how genomics and other emerging health technologies can provide more effective, personalised healthcare and deliver improvements in health for patients and citizens.

For more information contact:

intelligence@phgfoundation.org

UNIVERSITY OF CAMBRIDGE

phg foundation

making science work for health