phg foundation
making science work for health

Black box medicine and transparency

Technical summary

A PHG Foundation report for the Wellcome Trust

UNIVERSITY OF CAMBRIDGE

Authors

Alison Hall and Johan Ordish

Acknowledgements

The *Black Box Medicine and Transparency* project was funded by the Wellcome Trust as a part of their 2018 Seed Awards in Humanities and Social Sciences [Grant Number: 213623/Z/18/Z]. We thank the Wellcome Trust for their support.

The series of reports is informed and underpinned by a series of roundtables and interviews. These roundtables and interviews are detailed in the Report of Roundtables and Interviews. Further, highlights from both are seeded throughout all reports, being found in 'A Salient Feature' boxes.

Disclaimer

The following report is intended to provide general information and understanding of the law. The report should not be considered legal advice, nor used as a substitute for seeking qualified legal advice.

Hannah Murfet (PHG Foundation Fellow) contributed to this report by way of in-kind support from Microsoft Research Ltd. Any opinions expressed are the author's own, and may not represent the view of Microsoft Research.

URLs in this report were correct as of February 2020

This report is available from www.phgfoundation.org

Published by PHG Foundation 2 Worts Causeway, Cambridge, CB1 8RN, UK

+44 (0)1223 761900

February 2020

Correspondence to: intelligence@phgfoundation.org

How to reference this report:

Hall A, Ordish J. *Black Box Medicine and Transparency: Summary*. PHG Foundation. 2020.

PHG Foundation is an exempt charity under the Charities Act 2011 and is regulated by HEFCE as a connected institution of the University of Cambridge. We are also a registered company No. 5823194, working to achieve better health through the responsible and evidence based application of biomedical science

Contents

Technical Summary

Black Box Medicine and Transparency examines artificial intelligence (AI) (machine learning in particular) in the context of healthcare and health research. The series of reports considers the problem of interpretability (or the 'black box' problem) of machine learning in these sectors. Notably, there are sharply divergent views about the extent to which it is necessary to understand the inner working of machine learning models. Indeed, this uncertainty is made more pressing in the context of health with ethical principles often emphasising transparency and trust. If the opacity of machine learning models undermines trust in these models, their uptake and implementation may be slowed. Moreover, with the introduction of the General Protection Regulation (GDPR), lack of clarity over any legal requirement to render models transparent or explainable also threatens to slow or stifle innovation.

The *Black Box Medicine and Transparency* project addresses the nature of transparency and interpretability in black box medicine. Through iterative analysis of the philosophical and legal/regulatory principles underpinning the requirement for transparency, the project provides an in depth assessment of how these principles apply in the context of health care. The findings from these analyses have been used to construct a novel approach to systematically capture key indices of proposed machine learning models in the form of the *Interpretability by Design Framework*. This framework is intended as an aid for developers to think through interpretability of their machine learning models for healthcare.

The *Black Box Medicine and Transparency* Project consists of six discrete reports: these are complementary and can be read either as separate parts of a whole or as free-standing elements. This approach has been adopted because some areas of the analysis are necessarily complex and technical, and are likely to attract different audiences.

The six discrete reports include:

- *Machine Learning Landscape* considers the broad question of where machine learning is being used in healthcare and research for health
- *Interpretable Machine Learning* considers how machine learning is or may be made human interpretable
- *Ethics of Transparency and Explanation* considers why machine learning should be made transparent or be explained, drawing on many of the lessons the philosophical literature provides
- *Regulating Transparency* considers if and to what extent the GDPR requires machine learning in the context of healthcare and research to be interpretable
- The *Interpretability by Design Framework* distils the findings of the previous reports, providing a framework to think through interpretability of machine learning in the context of healthcare and research
- The *Report of Roundtables and Interviews* summarises the findings from the three roundtables and eleven interviews which provide the qualitative input seeded throughout the other reports.

This project has been made possible through a Wellcome Trust Seed Award Grant Number [213623/Z/18/Z] and the support of the many delegates and interviewees acknowledged in the *Report of Roundtables and Interviews* who generously gave their time and enthusiasm to the project.

The following summary outlines the broad conclusions of each report below. The paragraph numbers indicate the section numbers in the full report to which they refer.

1. Machine Learning Landscape

Where is machine learning being used in healthcare and research for health? What policy landscape surrounds AI for healthcare and research?

1.1 Machine learning is a subset of artificial intelligence.

1.2 Machine learning models are not explicitly programmed but are trained with many examples (data) relevant to the task, the algorithm finding structure in these examples to provide rules to automate the task. Machine learning often counts as an 'algorithmic modelling approach.' This approach assumes a black box that is complex and unknown, predicting input variables to output variables often without explaining what happens in between.

1.3 Machine learning has wide application across medical research and healthcare. Whilst some applications are already in use or close to implementation, others are more speculative. Machine learning for medical research has a number of applications: increasingly these may blur into delivery of care or applications with therapeutic intent. Machine learning for healthcare has broad potential application, involving much of the patient pathway and direct to consumer market. Machine learning as applied to health is undergoing systemisation with reporting standards being improved.

1.4 Although there has been a proliferation of policy guidance, it varies in specificity and statutory weight. In the UK, key sources of advice and guidance include the Information Commissioner's Office, which is the UK statutory authority for upholding information rights, and, within the health sector, NHSX, which has responsibility for establishing a framework for developing AI in the health and care sector.

There is increasing recognition that determining appropriate ethical and regulatory oversight of AI is a universal challenge which is best met by consistent and harmonised approaches. Important sources of guidance include the AI Ethics Guidelines developed by the EU High-level Expert Group on Artificial Intelligence, and the UK NHSX code of conduct for data-driven health and care technology. Both sources of guidance highlight transparency as a principle underpinning the development process, but do not provide more granular information about the form, content and timing of an explanation.

1.5 The Landscape Report concludes by highlighting two instances of machine learning applications where systems had wrongly identified confounding factors as being relevant: Caruana (2015) notes an example where a confounding factor was found to underpin a model predicting risk of death of pneumonia; Zech (2015) outlines an example where a confounding factor was found to dictate the outcome of a convolutional neural network to screen for pneumonia in x-rays. These examples demonstrate the potential (although not universally required) importance of interpretability in ensuring models are safe and meet their intended purpose - highlighting the central role of interpretability in the robust implementation of machine learning for health.

2. Interpretable Machine Learning

How is machine learning human interpretable? What methods are there to render machine learning human interpretable?

2.1 The term 'black box' encompasses opacity which is due to the uninterpretability of a model, and opacity that results from the intentional restriction of information.

2.2 A model might be interpretable to one person and not to another given the same information, depending on the technical literacy and understanding of the user.

2.3 'Black box medicine' captures wider concerns about how opaque forms of computational modelling might change the practice of medicine. Commentators such as Price distinguish between black boxes arising from the inherent complexity of biological relationships and those arising due to the machine learning process itself being opaque. It is unclear how this distinction fits with the burgeoning literature on explainable machine learning.

2.4 Interpretability of machine learning is complex and has multiple dimensions. Many terms encompass explainable machine learning but may emphasise different elements. Algorithmic transparency focuses on how an algorithm learns relationships and a model from the data it is provided with. Global interpretability of machine learning models concern the ability to 'understand the whole logic of a model and follow the entire reasoning leading to all different possible outcomes': this is a very demanding standard. Local interpretability of machine learning models illuminate 'only the reasons for a specific decision' which in healthcare, will typically be the output for a particular patient.

There are many methods to render machine learning interpretable or somewhat interpretable, each having their own strengths and weaknesses. Sometimes interpretable models (e.g. global surrogate models) are trained alongside uninterpretable models to infer an explanation. Other approaches involve visualisation which utilise visual representations of machine learning models to illuminate their function. 'Post hoc explainers' can render otherwise uninterpretable models somewhat interpretable by approximating model function. However, post hoc explainers do have weaknesses, the explainers being approximations of underlying models, sometimes only producing a partial explanation, and, in any case, machine learning models sometimes are limited in the extent to which they may be manually calibrated in light of any explanation.

In addition, multiple practical reasons to render a machine learning model interpretable emerge from the explainable AI (and related) literature. These utilise different methods and emphasise different dimensions of interpretability. Interpretability may be necessary to facilitate successful interaction between the human and the machine learning system. Moreover, because most models are necessarily 'incomplete', interpretability may be necessary to contextualise and rely upon the model's outputs.

The standard for interpretability for machine learning models is sometimes compared with human reasoning: healthcare professionals could be regarded as human black boxes, with human reasoning being subject to bias and impulse. Nevertheless, we might forgive some of this opacity because we can interrogate and question human professionals. We ought to ensure that we have similar tools to examine machine learning, especially if the system is safety critical.

There are three broad methods to evaluate the interpretability of machine learning models: functionally-grounded evaluation, human-grounded evaluation, and

application-grounded evaluation. Each method has its appropriate place: application-grounded evaluation being the most thorough but the most difficult to perform, functionally-grounded evaluation being the easier of the three to perform but also being the least rigorous.

3. 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?

3.1 Many groups call for transparency as a key ethical principle or note the concept as a key challenge for artificial intelligence. It is important that we understand how transparency and its limitations might impact on machine learning for healthcare and research.

3.2. Transparency is best analysed as a distinct concept that does not necessarily incorporate ideas of accessibility, communication, and interpretability. It should be viewed as a means to secure trustworthiness, not trust. To seek trust but not trustworthiness is pernicious. Often transparency facilitates users and publics to trust in machine learning systems. However, there are limits to transparency. Transparency emphasises disclosure but often underestimates the importance of communication and its associated virtues including accessibility, interpretability, and interest-sensitivity. These qualities are highlighted in the philosophy of explanation literature, especially in pragmatist accounts.

3.3. Where explanation is sought, the thing which is to be explained (the *explanandum*) should be defined with careful thought and, where appropriate, precision. If there is a specific purpose in mind, that which does the explaining (the *explanans)* should be crafted to serve this specific purpose. We should be clear what kind of explanation we seek. Not all explanations of machine learning will be scientific but some scientific explanations of machine learning may, through referring to general laws, be appropriate for healthcare or research. Different accounts of scientific explanation may be relevant to the interpretability of machine learning.

In the context of healthcare or medical research, machine learning will be used, ultimately, to guide diagnosis, treatment and management. Thus explanations will be sought of what led to a particular output. Patients may seek to understand how a decision or outcome might have been altered if different factors applied but finding an appropriate comparator might not always be straightforward. As a consequence, in the healthcare context, 'every day', casual, or counterfactual explanations may be the most common forms of explanation sought.

3.4. Explanatory pragmatism emphasises context, suggesting that it is misguided to find an explanation appropriate to all machine learning applications but that the appropriate explanation will depend on the context and the machine learning application at stake.

3.5. Explanation can be intrinsically valuable and/or instrumentally valuable. Explanation offers us the intrinsic good of understanding through providing a sound reason to believe certain propositions, link the unfamiliar to the familiar, tell us how a phenomena fits with others, highlight how a phenomena necessarily had to occur, and give us information about the cause of the phenomena in question. Understanding may not be the focus of an explanation in some contexts, but the promotion of understanding may be especially important for machine learning in health research. Explanation may also be instrumentally valuable, if it can be used to select, manipulate and control different features within a model, to obtain a different outcome.

3.6. The act of explaining aims to make something explainable to someone: it is an illocutionary act, in that something can be explained even if the act did not in fact render something explainable to another. We can distinguish between 'correct explanations' (where the 'propositional member of the ordered pair is true') and 'good explanations' (that take broader ideas of being appropriate to the addressee into account). There is likely no one

archetypically 'good' explanation since good explanations are sensitive to the interests and requirements of their audience.

3.7 Acts of explaining can be more or less satisfactory. Factors that might assist include:

- Meeting the specific questions the audience had in mind
- Selecting an appropriate comparator when approximating the specific explanatory information the addressee requires
- Selecting sufficient information to satisfy the questioner
- Using concepts and phenomena that the audience is already familiar with to explain concepts or phenomena the audience is unfamiliar with

3.8 The philosophical literature suggests that context is vital in understanding the key audiences and their interests in explaining or using machine learning to explain. One single explanation is unlikely to meet the expectations of each of these audiences and fulfil the purposes that they have in mind. There will often be distinct but overlapping purposes behind interpretability:

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 providing data subjects control and a means to secure controllers' accountability

These purposes will vary in importance depending on context. The weighting that might be applied to each of these purposes are explored in more detail in the Interpretability by Design Framework.

4. Regulating transparency

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

The Regulating Transparency Report considers the requirements of the General Data Protection Regulation (GDPR) on machine learning used for healthcare and medical research. It analyses the legal requirements for transparency and interpretability, and explores how these impact on the nature, timing and content of explanations.

4.1 The GDPR is one source of regulation that might generate a duty of transparency, interpretability, or explainability but is no panacea. Any duty of transparency, interpretability, or explainability that the GDPR offers will be a data protection solution that seeks to protect data protection interests and values. Data is protected by a complex web of regulation in England and Wales. Notably, ICO has statutory authority over other law that governs the same space. In parallel with data protection is a system of common law which includes duties of confidentiality as well as the tort of misuse of private information, and there may be other potential sector-specific regulation for AI in the future.

4.2 The GDPR is limited by its material scope: 'personal data.' It is important to consider how machine learning might be caught as personal data, distinguishing between training/test data as personal data, and data used as an input to a model as personal data. The GDPR is also limited by its territorial scope. First, what counts as 'processing of personal data in the context of the activities of an establishment' is likely to be broad, potentially including training/test datasets outside the Union if the models are eventually sold in the European Economic Area. Second, where not established in the Union, provisions relating to 'offering of goods or services' and 'monitoring' are expansive. In this way, no money needs be exchanged to count as 'offering of goods or services' and European Data Protection Board examples of 'monitoring' include applications that use personal predictors to provide personal recommendations.

The GDPR furnishes data subjects with data subject rights and assigns correlated duties to controllers and processors. It is the controller's responsibility to ensure these rights are complied with and data protection principles are upheld. In the context of healthcare and research, 'biometric data', 'genetic data', and 'data concerning health' all count as special category data and are subject to special restrictions and safeguards.

4.3 Three interrelated claims could be marshalled to generate a duty of transparency, interpretability, or explainability:

1. The general principle of transparency; and
2. How this principle interacts with specific data subject rights; and
3. Automated individual decision-making requirements.

4.4 The general principle of transparent processing is context-specific and user-centric. It requires controllers to consider the form in which they communicate (accessibility, simplicity, and intelligibility) as well as the content. In regards to the content, 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 derogated from. In the context of healthcare and research, four restrictions to data subject rights and data protection principles are particularly relevant. First, where the controller is no longer in a position to identify the data subject (Articles 11 and 12(2)). Second, the Article 23(1) restrictions that apply to health data according to the DPA 2018's Schedule 3, Part 2. Third, the flexibility for research purposes found in Article 89 and in the DPA 2018's Section 19 and Schedule 3, Part 6. Fourth, the restrictions relating to disclosure of trade secrets and intellectual property in 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 some rights may be recurrent: the right of access, is available at 'reasonable intervals' throughout the lifecycle of processing upon request from the data subject. However the rights to rectification and to object arguably require some interpretability or explainability to be vindicated. This suggests that the general principle of transparency combined with the data subject rights outlined is more than the sum of its parts and the wise data subject will use all the rights available to them to leverage interpretability or an explanation.

4.5 Automated individual decision-making provisions are the most prominent tools used to construct a 'right to explanation.' Two broad questions arise: first, when the right is triggered and second, what the right requires 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 lays down the broad conditions for automated individual decision-making but Article 22(1) captures only a narrow range of processing. Namely, '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. 'Based solely on automated processing', is interpreted by Working Party 29 to mean that any human in the loop must have 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 fuzzy but in the context of machine learning for healthcare and research, those systems that approve or deny a social benefit (including healthcare) may count as having 'legal effect.' 'Similarly significant effect,' is difficult to interpret 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 to interpret: 'meaningful information about the logic involved', 'significance and the envisaged consequences', the 'right to contest under Article 22(3).' 'Meaningful information about the logic involved' as applied to machine learning may require the disclosure of a variety of information and likely 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 the 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. Emphasising 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 both explanation before processing and explanation after processing. This may require both global interpretability of the model overall but also local interpretability of particular instances of processing. If a controller is caught by Article 22(1), the requirements of Article 22(2) also narrow and complicate the legal position of the controller, especially if the controller processes special category data.

4.6 Many different spheres of regulation potentially require transparency, interpretability or explainability of machine learning used for healthcare and research. The most prominent of these is the GDPR. The general principle of transparency underpins and informs any duties of transparency, interpretability, or explanation, which are context-sensitive. This principle is instantiated by associated data subject rights. Although the rights to information, access and data portability do not directly require interpretability or explanation, they may assist in leveraging these. Other rights to rectification, object, erasure and restriction of processing may, depending on context, require some interpretability or explanation for their vindication. These rights may often be blocked, restricted or excluded.

5. Interpretability by design framework

Building on the ethical and legal analysis, the *Interpretability by Design Framework* (ID Framework) assists developers to think through interpretability of their machine learning models for healthcare. Intended as an aid to good practice, the ID Framework provides a structure and process to systematically review various dimensions of the proposed tool and its application. A number of principles underpin and inform the ID Framework. Interpretability is a design choice, which like other attributes, accuracy and usability, should be incorporated into the development process. However the weight placed on interpretability over other criteria depends strongly on context. Reaching a judgment about what is required will often require input from a multidisciplinary team.

The ID Framework provides seven steps to assist developers in thinking through the interpretability of their system. These steps provide an iterative and comprehensive process for considering the key audiences and benefits of interpretability, and then weighting and integrating key attributes of machine learning systems, including automation, adaptivity, risk, (lack of) ground truth, and (in) completeness. The final step involves comparing this composite score with weighted assessments of risk and opacity. The ID Framework uses processes and concepts already used in applicable law and ethical principles.

6. Roundtables and Interviews

Three roundtable workshops involving a total of 35 external delegates were held as part of the *Black Box Medicine and Transparency* project. The purpose of these workshops were to provide an understanding of the perspectives, approaches, and challenges faced by stakeholders in explanation and interpretability of machine learning for healthcare and research, to address gaps and queries arising from the ethical and legal analysis, and to develop, test and refine the Interpretability by Design Framework. Direct quotations from these roundtables are seeded throughout the separate reports.

A first roundtable on *Developing Transparency,* held on 3rd June 2019, involved developers of machine learning products for healthcare or research. Sessions covered explainable AI, and explored how developers currently approach explanation and the factors influencing explanation. There was broad agreement that the prototype Interpretability by Design Framework was useful, and various suggestions were made as to how it could be improved.

A second roundtable *Clinical Focus*, held on 8th July 2019, involved clinicians, clinical communication specialists and patient representatives. After introducing explainable AI and explanations in healthcare, the perspectives of healthcare professionals and patients were explored: in practice patients often delegate decision making to their health professional. Some felt that there needed to be a persuasive reason to use a complex opaque model over a simpler interpretable model. Potential communication challenges were highlighted. Professionals tailor their evidence requirements to the specific intended use of the device and also utilise other tools apart from interpretability to ensure that systems are reliable and safe. The Interpretability by Design Framework was further iterated but felt to be less useful for the clinical community than for developers.

A third roundtable *Policy and Regulatory Focus* took place on 9th September 2019 with the objective of testing preliminary findings from the ethical and legal phases, further iteration of the Interpretability by Design Framework and to consider emerging points of consensus and recommendations. Key conclusions included that 'the decision' at stake pursuant to GDPR Article 22(1) needs clarification particularly in the healthcare context. Discussions also highlighted interpretative difficulties around 'data concerning health' and 'personal data', and the fact that the views of patients and their healthcare professionals might be divergent, with patients' likely rating predictive accuracy over interpretation. Both groups agreed that interpretability should be included in robust design and development processes, and highlighted the central role of trust and trustworthiness.

These reports are intended as a resource for all those interested in optimising the utility of machine learning and AI for health through responsible and proportionate policy development.

For more information see www.phgfoundation.org.

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