phg foundation
making science work for health

Black box medicine and transparency

Annexes: report of roundtables and interviews

A PHG Foundation report for the Wellcome Trust

UNIVERSITY OF CAMBRIDGE

Authors

Johan Ordish, Tanya Brigden, Colin Mitchell, and Alison Hall

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:

Ordish J, Brigden T, Mitchell C, Hall A. *Black Box Medicine and Transparency: Report of Roundtables and Interviews*. PHG Foundation. 2020.

Contents

1. Roundtables and Interviews

Three roundtable workshops were held as part of the *Black Box Medicine and Transparency* project. These workshops were held in June-September 2019 after the ethical and legal/regulatory phases had been completed. The purpose of these workshops were to inform our 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 our analysis; and to develop, test and refine the *Interpretability by Design Framework* described in the previous report.

In advance of the Roundtables, two discussion papers were drafted to provide background briefings to delegates. These papers, *A Right to Explanation?* and *Why explainable learning matters for health* are included in Section 6 of this report. Further, in advance of Roundtables 1 and 2, hypothetical worked examples were generated and they were used during these Roundtables to provoke and support structured discussion. These examples are also included in Section 6 of this report.

The delegates, proceedings, and findings of each of the roundtable workshops are set out below. In addition, key comments from delegates are seeded throughout the series of *Black Box Medicine and Transparency* reports where pertinent to the content. The comments from the Roundtables and from the Interviews are deliberately not intended to be attributable to any specific delegate or interviewee, as this was one of the conditions of their participation in the project.

These summaries form part of a larger project, *Black Box Medicine and Transparency*, consisting of 7 reports available at www.phgfoundation.org

2. Roundtable 1 - Black Box Medicine and Transparency: Developing Transparency

Roundtable 1 - *Black Box Medicine and Transparency: Developing Transparency* took place on the 3rd of June 2019 at the Tamburlaine Hotel, Cambridge.

The group of invitees primarily consisted of developers of machine learning products in the healthcare or research spaces.

The Roundtable was purposely kept small, emphasising interactivity, discussion, and development of ideas through several facilitation methods.

In advance of each Roundtable, delegates were provided with two discussion papers prepared as briefings: *A Right to Explanation?* and *Why explainable learning matters for health.*

a. Proceedings

The primary purposes of the day were:

a) To understand how developers approached explanation and interpretability of machine learning for healthcare and research
b) To test preliminary findings encapsulated in the machine learning 'personality profile' framework
c) To consider emerging points of consensus and recommendations

Roundtable 1 sought to fulfil these purposes through a number of sessions:

Explainable AI outlined the literature on interpretable machine learning, including select methods to render otherwise uninterpretable models interpretable and how the details of interpretability might differ depending on the device in question and the wider context.

Chaired discussion: how do developers currently approach explanation? Participants were asked to describe how interpretability factored into the design of their system, their current practice and plans for implementing interpretability, and their perceived importance or unimportance of interpretability in each sub-sector.

Interactive session: factors influencing explanation This session outlined preliminary ideas for a machine learning 'personality profile' framework that attempted to capture the axes of a machine learning system for healthcare or research that might require the model to be human interpretable. We asked each participant to score and plot on a radar diagram their system and discussed whether the framework was helpful to think through the interpretability of machine learning in the context of healthcare or research.

Key findings included:

- *Broad agreement* that the framework was useful to think through interpretability with respect to each machine learning model
- *Broad agreement* that the framework needed further axes added and the relationship between axes to be explored more thoroughly

Break-out session: explanations in context This session provided four hypothetical examples based on real applications of machine learning in healthcare:

a) A triage tool to assign a risk score to patients, this risk score determining (along with clinical input) whether patients are admitted to the ICU, assign beds, and assist with discharge.
b) A diabetes management application that incorporates glucose level readings, self-reported food logging, and other relevant information to provide personalised insulin recommendations.
c) A neural network to automatically highlight suspected stress fractures in x-rays for closer examination by a radiographer.
d) A text-based chatbot application to deliver personalised cognitive behavioural therapy.

Participants were asked to consider their assigned example with respect to the framework, plot its axes, and consider what interpretability or explanation might be appropriate.

Key findings included:

- *Broad agreement* that the framework was useful for thinking through interpretability of machine learning with respect to a diverse range of tools and framing the kind of explanation that might be of assistance
- *Broad agreement* that social impact should be considered as an element in the framework
- *Broad agreement* that explanation is neither necessary nor sufficient for many contexts in machine learning for healthcare. For instance, explanation is no guarantee of accuracy - a model can be explainable yet wrong
- *Broad agreement* that machine learning tools need to demonstrate promise of tangible benefit, especially if the tools are more opaque by comparison with other tools which do not utilise machine learning
- *Broad agreement* that interpretability of machine learning models for healthcare does not sit in isolation but fits within the context of scientific (and other) evidence bases

Discussion: key conclusions and recommendations In the final session, participants were asked to feedback on *Black Box Medicine and Transparency's* preliminary findings, consider improvements to be made, and identify gaps in the analysis or priorities for interpretability in the context of healthcare or research.

Broadly, participants found:

- *Broad agreement* that the framework was helpful in thinking through interpretability with respect to developers' machine learning systems and assisted in framing the interpretability or explanation needed
- *Broad agreement* on modifications to the framework - the addition of an axis, possible weighting of axes, and clarification over the purpose and set up of the personality profile framework

- *Partial agreement* noting the gulf in methods between commercial development of machine learning versus those systems developed in an academic setting - how the systems are designed, the priorities of the designers, how interpretability might be included
- *Partial agreement* noting the difference in approach in the commercial sector - different organisations had different risk appetites - some organisations saw themselves as developing highly innovative but riskier products, others had a more conservative approach to developing machine learning systems
- *Partial agreement* emphasizing that there are domains where machine learning can outperform humans. However, there are other domains of healthcare and research that will be much harder to implement accurate, usable machine learning

Key findings from Roundtable 1 are seeded throughout the series of *Black Box Medicine and Transparency* reports in the 'A Salient Feature' boxes.

b. List of attendees

We thank the following attendees for their time, engagement, and generous sharing of ideas:

Name	**Job Title**	**Organisation**
Peter Fish	Head of Clinical Strategy	Mendelian
Andrew Fried	Life Sciences, Global Industry Leader	IBM
Sharanya Gajapathy	European Healthcare and Life Sciences Operations Manager	IBM
Hannah Murfet	Senior Compliance Manager	Microsoft Research
Gabriel Recchia	Research Associate	Winton Centre for Risk & Evidence Communication, University of Cambridge
Anthony Rix	CEO/CTO	Granta Innovation and Amiri Health
Hannah Thompson	Chief Product and People Officer	Cambridge Cancer Genomics

3. Roundtable 2 - Black Box Medicine and Transparency: Clinical Focus

Roundtable 2 - *Black Box Medicine and Transparency: Clinical Focus* took place on the 8th of July 2019 at the Tamburlaine Hotel, Cambridge.

The group of invitees primarily consisted of clinicians, clinical communication specialists, and patient representatives.

The Roundtable was purposely kept small, emphasising interactivity, discussion, and development of ideas through several facilitation methods.

a. Proceedings

The primary purposes of the day were:

a) To understand what interpretability (if any) clinicians and patients might require of machine learning systems
b) To test preliminary findings encapsulated in the machine learning 'personality profile' framework
c) To consider emerging points of consensus and recommendations

Roundtable 2 sought to fulfil these purposes through a number of sessions:

Explainable AI outlined the literature on interpretable machine learning, including select methods to render otherwise uninterpretable models interpretable and how the details of interpretability might differ depending on the device in question.

Notably:

- *Dissenting viewpoints* noted that the concept of a 'black box' was inherently flawed - all machine learning is interpretable, depending on the knowledge and capacities of the person examining the model
- *Dissenting viewpoints* were incredulous that interpretability was an important trait for machine learning models. These participants emphasised concern that interpretability would be better served by reporting out of sample errors and reporting of predictive accuracy. We noted that the importance of interpretability was more controversial in Roundtable 2 compared to Roundtable 1

Explanations in healthcare outlined the different key audiences of interpretability in a clinical context: healthcare professionals, patients (via a healthcare professional), and consumers in a direct to consumer context. Broadly, this presentation sought to spark conversation about what each audience might require in regards to interpretability of machine learning models, their specific needs, and the challenges with respect to each audience.

Structured discussion of AI, explanations, and healthcare sought to collect from participants the following information: how does AI fit into your practice, now or in the near-future? What are the key challenges associated with explaining AI technologies to healthcare professionals? Are there different challenges explaining AI technologies to patients/consumers, and if so, what are these challenges?

Key conclusions include:

- *Broad agreement* on importance of having a concrete idea of what the explanation or interpretability is for, who it is for, and what dimension of interpretability the explanation seeks to illuminate
- *Broad agreement* that patients more often than not trust their clinician to make the best decision based on the tools they have available to them and have little desire to engage with explanations of such tools
- *Broad agreement* that the pathology community is worried that their workforce will be replaced by automation and that they are training their replacements when labelling datasets for supervised machine learning models
- *Broad agreement* that often precise numbers are perceived as being authoritative and accurate by virtue of their specificity

- *Partial agreement* that there is often bias in clinical judgment - machine learning will be acceptable to clinicians to the extent that it confirms findings but met with scepticism where it disagrees with human clinical judgment
- *Partial agreement* that machine learning is not special or exceptional - it is one instrument for use and contextualisation by healthcare professionals

- *Dissenting viewpoints* outlined that perhaps we should not use the terms 'machine learning' or 'artificial intelligence' when communicating with patients or consumers but instead use terms like 'complex modelling' to remove the stigma and mystique associated with the technology

A tool for developing explanations This session outlined the next iteration of the machine learning 'personality profile' framework presented at Roundtable 1. We asked participants for feedback on the framework, its usefulness for their practice, as well as possible uses for patients, consumers, and regulators.

Key findings include:

- *Broad agreement* that the existing framework was too complex and not fit the needs for communicating the attributes of a machine learning system to patients and consumers
- *Broad agreement* that further clarification is necessary with respect to: the intended audience of the framework, how the axes fit together, and reworking of the positioning of the framework

- *Partial agreement* that the framework was helpful to think through interpretability of machine learning models and of assistance when considering how to frame explanations of machine learning for different audiences

Break-out session: explanations in context This session presented revised hypothetical examples from Roundtable 1:

a) A triage tool to assign a risk score to patients, this risk score determining (along with clinical input) whether patients are admitted to the ICU, assign beds, and assist with discharge.
b) A diabetes management application that incorporates glucose level readings, self-reported food logging, and other relevant information to provide personalised insulin recommendations.
c) A neural network to automatically highlight suspected stress fractures in x-rays for closer examination by a radiographer.
d) A text-based chatbot application to deliver personalised cognitive behavioural therapy.

Participants were asked to consider their assigned example with respect to the framework, plot its axes, and consider what interpretability or explanation might be required by healthcare professionals, patients and consumers.

Key points included:

- *Broad agreement* that healthcare professionals typically look to restricted forms of evidence when considering what tools to use: CE marking, NICE evaluation, and, in limited circumstances, enquire about the training and test sets of data
- *Broad agreement* that it is important to consider the specific intended use of devices and frame any explanation or requirement of interpretability with respect to that intended use
- *Broad agreement* that there are other tools apart from interpretability to assist in ensuring systems are safe and meet their intended use - for instance, out of sample error reporting, clear labelling, and the inclusion of monitors/alarms

- *Partial agreement* that there needs to be persuasive reason to use a complex, opaque model over a simple, interpretable model, for instance, gains in terms of predictive accuracy

Discussion: key conclusions and recommendations In the final session, participants were asked to feedback on *Black Box Medicine and Transparency's* preliminary findings, consider improvements to be made, and gaps in the analysis or priorities for interpretability in the context of healthcare or research.

Broadly, key findings included:

- *Broad agreement* that the machine learning 'personality profile' framework was a useful exercise but in need of iteration and improvement. It was especially clear that the purpose the framework serves and the audience to whom it is directed needs to be clearly articulated
- *Broad agreement* that healthcare professionals typically rely on heuristics to decide whether to use a system or not – for instance, CE marking, NICE recommendations and so on. It is unclear how explanation of machine learning models might fit in with these forms of evidence
- *Broad agreement* that healthcare professionals, if they are interested in the underpinning of the machine learning model in question, are interested in the inputs (the training and test datasets) and the outputs, but not necessarily the weightings or significance of the features of the model
- *Broad agreement* that we need to guard against AI exceptionalism - the idea that AI is something new, categorically different from other technology, and inherently dangerous

- *Partial agreement* that machine learning models are no more than and, in some instances, more interpretable than human healthcare professionals
- *Partial agreement* that communication strategies are needed to inform how to communicate machine learning models to patients and consumers in the direct to consumer context

Key findings from Roundtable 2 are seeded throughout the series of *Black Box Medicine and Transparency* reports in the 'A Salient Feature' boxes.

b. List of attendees

We thank the following attendees for their time, engagement, and generous sharing of ideas:

Name	**Job Title**	**Organisation**
Stephanie Archer	Research Associate	University of Cambridge, School of Clinical Medicine
Areeq Chowdhury	Head of Think Tank	Future Advocacy
Shah Islam	Academic Neuroradiologist	Imperial College London
Parashkev Nachev	Senior Clinical Research Associate	University College London (UCL) Institute of Neurology
Gabriel Recchia	Research Associate	University of Cambridge, Centre for Research in the Arts, Social Sciences and Humanities (CRASSH)
Saskia Sanderson	Research Psychologist & Senior Research Fellow	UCL Institute of Health Informatics and PHG Foundation Associate
Bethany Williams	Digital Pathology Fellow	Leeds Teaching Hospital NHS Trust
Evan Wroe	Communications Officer	Genetic Alliance

4. Roundtable 3 - Black Box Medicine and Transparency: Policy and Regulatory Focus

Roundtable 3 - *Black Box Medicine and Transparency: Policy and Regulatory Focus* took place on the 9th of September at the Wellcome Collection, London.

The group of invitees primarily consisted of policymakers and representatives from regulatory bodies in the healthcare and research spaces.

The Roundtable emphasised interactivity, discussion, and development of ideas through several facilitation methods.

a. Proceedings

The primary purposes of the day were:

a) To understand how interpretability fits into policy and regulation in the healthcare and research sectors
b) To test preliminary findings on the ethical and legal requirements of interpretability in healthcare and research
c) To test preliminary findings encapsulated in the 'model for transparency by design'
d) To consider emerging points of consensus and recommendations

Roundtable 3 sought to fulfil these purposes through a number of sessions:

Welcome and policy context Outlined the complex policy landscape that surrounds machine learning in healthcare and research. This outline situated the work of *Black Box Medicine and Transparency*, noting possible synergies and points for collaboration.

Machine learning in healthcare and research Highlighted the breadth of near-use applications for machine learning in healthcare and research: tools to assist with meta-analyses, machine learning for drug discovery, machine learning for image analysis, machine learning for diagnosis and symptom checking, monitoring and management of conditions, public health surveillance, and so on.

Knowledge gathering: current work on transparency and machine learning Asked participants to outline their current and future work on transparency and machine learning - the ethical, legal, and policy initiatives that might be of relevance.

Requirements for transparency Presented interim findings on: the literature on interpretable machine learning, a philosophical analysis of transparency and explanation, and a legal analysis of the GDPR's requirements for transparency and explanation in the context of machine learning for healthcare and research.

Key findings included:

- *Broad agreement* that the problem of interpretability should not be reduced to just a problem of complexity - simple models can be opaque too
- *Broad agreement* that understanding of how healthcare professionals engage, interpret, and interact with these tools is important
- *Broad agreement* that what counts as 'the decision' at stake when considering Article 22(1) GDPR needs clarification and is perhaps a problematic concept for the healthcare

space. That is, healthcare usually consists of a number of decisions being made, leading to diagnosis or treatment - not just one
- *Broad agreement* that it is important to consider what the patient could reasonably expect of their healthcare professional to explain in regards to their model
- *Broad agreement* that machine learning for this sector will likely progress in stages: scepticism, acceptance as an equal, and then the healthcare professional becoming subservient to the model
- *Broad agreement* that public expectations will impact the regulatory thresholds for machine learning in the context of healthcare. Currently, there is often a large gap between what the public expects machine learning to do and what it can currently do

- *Partial agreement*, in the context of the GDPR and DPA 2018, to more clearly distinguish between information that is only relevant to general duties of transparency and information that might constitute interpretability of explainability - many of the duties of transparency do not require any kind of interpretability or explainability
- *Partial agreement* that there will be some applications for which only intrinsically interpretable machine learning will be acceptable

Developing machine learning devices The first half of this session provided a developer's perspective of developing machine learning in this space - the various ethical, commercial, and regulatory reasons why machine learning might be rendered interpretable. The second half of this session considered to what extent medical device regulation also addresses interpretability and how the requirements of the GDPR might intersect with the Medical Device Regulation and In Vitro Diagnostic Medical Devices Regulation.

Discussion: the wider legal and policy landscape This session asked participants to share where they think the policy/regulatory gaps, tensions, or uncertainties are.

Key conclusions include:

- *Broad agreement* that there may be sector-specific regulation for AI emerging in the near future
- *Broad agreement* that the lines between classes of data can be uncertain and difficult to sensibly draw. For instance, under the GDPR, where does data become 'data concerning health' given that data like supermarket shopping habits could be linked to draw health-related conclusions
- *Broad agreement* that the material scope of the GDPR is uncertain - it is unclear, for example, whether pseudonymised data is always 'personal data.' Given this, it is also unclear how many machine learning systems will be caught by the GDPR's requirements
- *Broad agreement* that often patients and data subjects do not want a technical analysis of the features of the model and their weighting. Tentative analysis shows that patients rate predictive accuracy over interpretability
- *Broad agreement* that interpretability is no replacement for good design but should be included in robust design and development processes. Ideally, these processes should include input from a host of parties, for example, patients and healthcare professionals

- *Partial agreement* that addressing the interests of data subjects might be insufficient to address the interests of other 'output recipients' (that might not count as data subjects) who are those potentially impacted by findings

A proposed framework for transparency by design This session presented the latest iteration of the machine learning 'personality profile,' now called the 'transparency by design framework.' We asked participants for feedback on the framework, whether it might assist

developers and regulators to think through interpretability, and how the framework might fit with current regulatory and policy frameworks.

Key conclusions include:

- *Broad agreement* that scoring of the axes needed clarification and that the relationship between axes requires reorienting
- *Broad agreement* that the framework would complement recently released guidance from NHSX
- *Broad agreement* that we should be careful to avoid giving the impression that completion of the framework's checklist allows the applicant to abdicate responsibility and consider the job of responsible design to be completed

Discussion: priorities for regulators and policymakers In the final session, participants were asked to draw on conversations throughout the day to propose issues missed and consider policy priorities as well as ways forward.

The main points included:

- *Broad agreement* that decision supports tools in general are not new - healthcare professionals are accustomed to contextualising these tools for patients, sometimes differently for different patients. It is important to consider how healthcare professionals will manage conversations with their patients about how machine learning has influenced their decision
- *Broad agreement* that it is important to consider what is reasonable for healthcare professionals to understand and communicate - are we asking them to assess the performance of devices even though much of the information required to do this is not in the public domain?
- *Broad agreement* that high accuracy does not necessarily mean better outcomes for patients. Better diagnostic yield does not necessarily result in better therapeutic yield, especially if considerations like overdiagnosis are taken into account
- *Broad agreement* that a part of the conversation about machine learning in healthcare is about trust, ensuring that we have trustworthy experts to test, interpret, and contextualise machine learning outputs so the patient or frontline clinician does not have to
- *Broad agreement* that machine learning likely requires the tweaking of existing regulatory and policy regimes rather than wholesale upheaval or replacement
- *Broad agreement* that the NHS has a large repository of trust and goodwill from the public but that this trust can be fragile. Given this, engagement with the commercial sector in the form of co-development of models or sharing of data should be carefully considered

- *Partial agreement* to separate out what GDPR requirements relate to transparency and which relate to explainability or interpretability
- *Partial agreement* that the term 'black box' can be misleading, indicating that we know nothing about the model, where, in reality, we can still subject the model to testing, measure its predictive accuracy, and so on
- *Partial agreement* that the direct to consumer context may be key - the promise of machine learning is to automate some processes, freeing up time for healthcare professionals to redistribute their time. However, not all tools will be mediated by a healthcare system and healthcare professional
- *Partial agreement* that we should re-examine what term we use to describe 'machine learning' - should we be picking something less threatening and scary?

Key findings from Roundtable 3 are seeded throughout the series of *Black Box Medicine and Transparency* reports in the 'A Salient Feature' boxes.

b. List of attendees

We thank the following attendees for their time, engagement, and generous sharing of ideas:

Name	**Job Title**	**Organisation**
Mark Birse	Group Manager, Device Safety and Surveillance and Device Software and Apps	Medicine and Healthcare products Regulatory Agency
Vicky Chico	Lecturer in Law / Data Policy Adviser	Sheffield University / Health Research Authority
Alice Clay	Assistant Programme Manager	NESTA
Rachel Coyle	Public Health Registrar	Public Health England
Alastair Denniston	Consultant Ophthalmologist	University of Birmingham
Heather Draper	Professor of Bioethics	University of Warwick
Nick Fuggle	Dunhill Clinical Research Fellow	University of Southampton
Jonathan Hope	Principal Data Manager - Data Science	NHS Digital
Xiao Liu	Clinical Research Fellow	University of Birmingham
Hannah Murfet	Senior Compliance Manager	Microsoft Research
Will Navaie	Engagement Manager	Health Research Authority
Rune Nyrup	Postdoctoral Research Associate	Leverhulme Centre for the Future of Intelligence
Florian Ostmann	Policy Fellow	Alan Turing Institute
Sara Payne	Associate	PHG Foundation
Aidan Peppin	Researcher	Ada Lovelace Institute
Helena Quinn	Senior Policy Officer	Information Commissioner's Office
Vibha Sharma	Regulation Policy Manager,	General Medical Council

	Strategy and Policy Directorate	
Adam Steventon	Director of Data Analytics	Health Foundation
David Watson	DPhil Candidate	University of Oxford
Zoe Webster	Director of AI and Digital Economy	Innovate UK

5. Interviews

As a part of the *Black Box Medicine and Transparency* project 11 separate interviews were conducted with 13 interviewees.

We thank all interviewees for their time and input into *Black Box Medicine and Transparency*.

a. Purpose of the interviews

Interviews were conducted for the following purposes:

a) To add detail in relation to specific technical areas or questions
b) To sense check preliminary findings
c) To understand where *Black Box Medicine and Transparency's* findings might fit in the wider policy landscape

Key findings from the interviews conducted are seeded throughout the series of *Black Box Medicine and Transparency* reports in 'A Salient Feature' boxes

b. List of interviewees

We thank the following interviewees for their time, engagement, and generous sharing of ideas:

Name	**Job Title**	**Organisation**	**General area of expertise**
Reuben Binns	Postdoctoral Research Fellow in AI	The Information Commissioner's Office	Philosophy
Loubna Bouarfa	CEO	OKRA Technologies	Machine learning
Alastair Denniston	Consultant Ophthalmologist	University of Birmingham	Clinical
David Erdos	University Senior Lecturer in Law and the Open Society	University of Cambridge	Law
Eddie Korot	Clinical Research Fellow	Moorfields Eye Hospital	Clinical
Xiaoxuan Liu	Clinical Research Fellow	University of Birmingham	Clinical
Brent Mittelstadt	Research Fellow	Oxford Internet	Philosophy

	and British Academy Postdoctoral Fellow in data ethics	Institute	
Christoph Molnar	PhD Candidate	Department of Statistics, LMU Munich	Interpretable Machine Learning
Rune Nyrup	Postdoctoral Research Associate	Leverhulme Centre for the Future of Intelligence	Philosophy
Adrian Price	Policy Lead – Innovation and Horizon Scanning	NHSX	Policy
Helena Quinn	Senior Policy Officer	Information Commissioner's Office	Policy
Marco Riberio	Researcher	Adaptive Systems and Interaction Group, Microsoft	Interpretable Machine Learning
Carl Wiper	Group Manager – Strategic Policy Projects	Information Commissioner's Office	Policy

6. Roundtable materials

The following materials were produced to facilitate discussion in Roundtables 1, 2, and 3 described above.

a. Worked hypothetical examples

GROUP A

- *Please discuss the following two (hypothetical) worked examples of machine learning for health, considering:*

    A. What kind of explanation of the tool would a clinician desire from developers?

    B. How should this explanation be given?

    C. When should this explanation be given?

    D. What kind explanation do you think patients or consumers want or need? When/how? From clinicians and developers?

    E. What are some of the challenges explanation of this tool might face?

- Is this a tool which is too uninterpretable to feel safe about using or communicating to patients?

- Would you make recommendations to change the device or its intended use, i.e. provide more information, a visualisation etc?

Hypothetical Example A1

Hello! My name is...

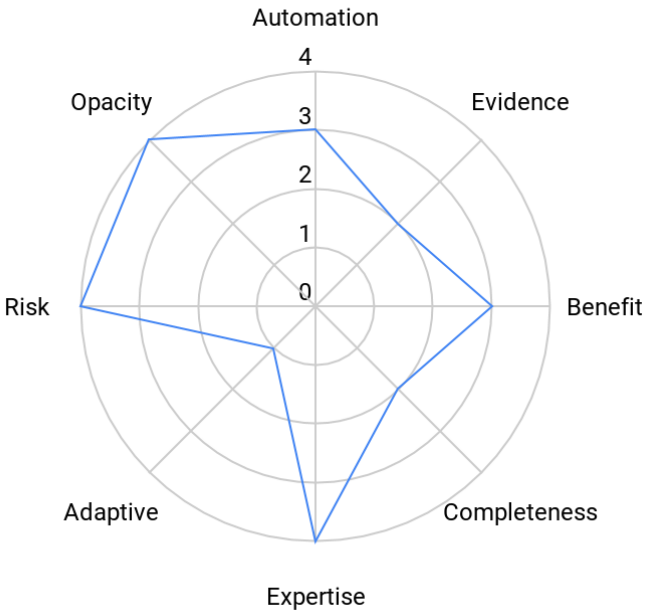
TRIAGETOOL

Description of tool: TRIAGETOOL is trained using deep learning and assigns a risk score to patients. This risk score is used to triage patients and determine whether patients are admitted to the ICU or not, assign beds, and discharge. The model appears to be highly accurate, initial investigations demonstrate that the model generally assigns an accurate risk score and triages patients according to their medical priority. However, at present, the generalisability of these findings is in doubt, as the model has been trained and tested in only a small set of academic teaching hospitals.

Characteristics of the tool:

a) **Automation (3/4):** TRIAGETOOL is highly automated, collecting data from patient EHRs, generating risk scores largely without clinical input and oversight. However, the decision to admit patients is still ultimately up to clinicians and they may overrule the risk score assigned.
b) **Evidence (2/4)**: TRIAGETOOL uses a number of indicators established in the general scientific literature to be linked to patient outcome. There is no general literature establishing the clinical utility of models like TRIAGETOOL.
c) **Benefit (3/4)**: ICUs are already overburdened, so gains in efficiency will pay dividends. However, QALY/DALY/HYE analysis notes that gains are restricted to a small subset of the population.
d) **Completeness (2/4)**: triaging decisions involve clinical judgment - the ultimate triaging decision often involves a number of variables that are either unquantifiable or simply not included in the TRIAGETOOL model.
e) **Expertise (4/4):** TRIAGETOOL is designed and marketed only to hospitals and their intensive care clinicians. The tool also comes with a short training programme and support service.
f) **Adaptive (1/4):** The machine learning model is largely static, the model does not incorporate streaming data nor does it learn incrementally. The developers may update the training data and model but only with pushed updates and through a rigorous change management process.
g) **Risk (4/4):** TRIAGETOOL can determine admission into the ICU and other resource allocation largely without human scrutiny. A failure to triage properly can cause death or serious injury.
h) **Opacity (4/4):** TRIAGETOOL incorporates multiple layers in its deep learning network. While ostensibly highly accurate, the model's decisions remain human uninterpretable if no model-agnostic or example-based explainer is used.

TRIAGETOOL

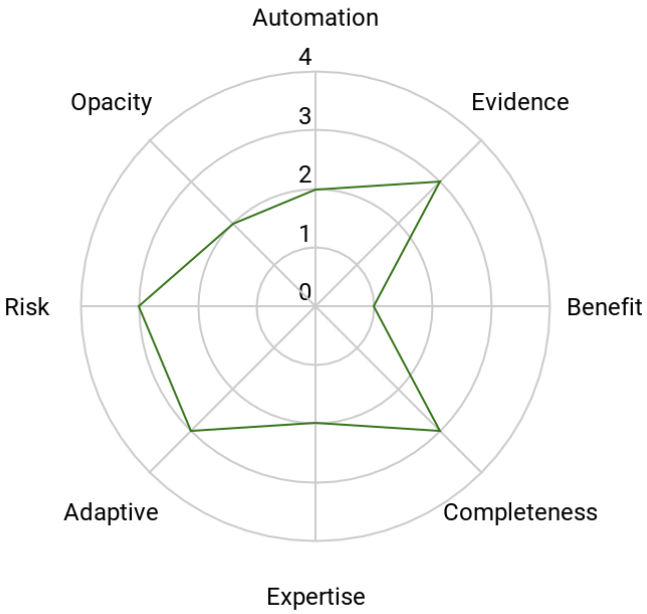

Automation
4
3
2
1
0
Evidence
Benefit
Completeness
Expertise
Adaptive
Risk
Opacity

Hypothetical Example A2

Description of tool: TRACKDIA is a machine learning enabled mobile application that assists patients to better manage their diabetes. TRACKDIA incorporates glucose level readings, self-reported food logging, and other relevant information. Using decision trees, the machine learning model provides personalised management recommendations to the patient, suggesting, for example, when to take insulin. The application requires that finger stick data be returned on a regular basis to function. TRACKDIA has been praised by a number of diabetes patient groups and has received funding from a small number of NHS Trusts.

a) **Automation (2/4):** TRACKDIA relies on user input to generate its recommendations. The labelling makes clear that TRACKDIA's recommendations are advisory only and to be used in combination with qualified medical advice and proper diabetes management. The app also includes functionality for a patient to give their GP access to their data and TRACKDIA recommendations. Patients may correct the recommendations provided and control some parameters.
b) **Evidence (3/4)**: there is extremely strong general evidence establishing a link between insulin and glucose levels. However, there is only a developing evidence base establishing the effectiveness of personalised management recommendations that include other variables.
c) **Benefit (1/4)**: utility of these personalised recommendations over and above simple glucose level to insulin dose recommendations represents a modest improvement in patient outcome.
d) **Completeness (3/4)**: The problem of recommending an insulin dose based on glucose level readings is well-understood and quantifiable. However, the personalised elements of the recommendations introduce a level of incompleteness and unquantified variability.
e) **Expertise (2/4):** TRACKDIA is marketed as a prescription only tool and requires GP referral to use. Instructions for use are provided within the application. When prescribed, the patient is given general information about the device and how to use it.
f) **Adaptive (3/4):** The machine learning model that underpins TRACKDIA, while relatively simple, retrains in batches, incorporating new data every once in a while to produce its personalised recommendations.
g) **Risk (3/4):** TRACKDIA maintains that its tool is assistive and ought to be used in conjunction with qualified medical advice and regular diabetes management and care. However, improper dosage of insulin can cause serious health issues.
h) **Opacity (2/4):** TRACKDIA, while incorporating multiple inputs and a variety of data, decision trees can still be human interpretable. In addition to this, the application also produces a graph to show trajectory of glucose level, suggesting to the user how the recommendation was arrived at. Nevertheless, the decision tree remains publically unavailable due to commercial sensitivity.

TRACKDIA


Automation
4
3
2
1
0
Evidence
Benefit
Completeness
Expertise
Adaptive
Risk
Opacity

Group B

- *Please discuss the following two (hypothetical) worked examples of machine learning for health, considering:*

    A. What kind of explanation of the tool would a clinician desire from developers?

    B. How should this explanation be given?

    C. When should this explanation be given?

    D. What kind explanation do you think patients or consumers want or need? When/how? From clinicians and developers?

    E. What are some of the challenges explanation of this tool might face?

    - Is this a tool which is too uninterpretable to feel safe about using or communicating to patients?

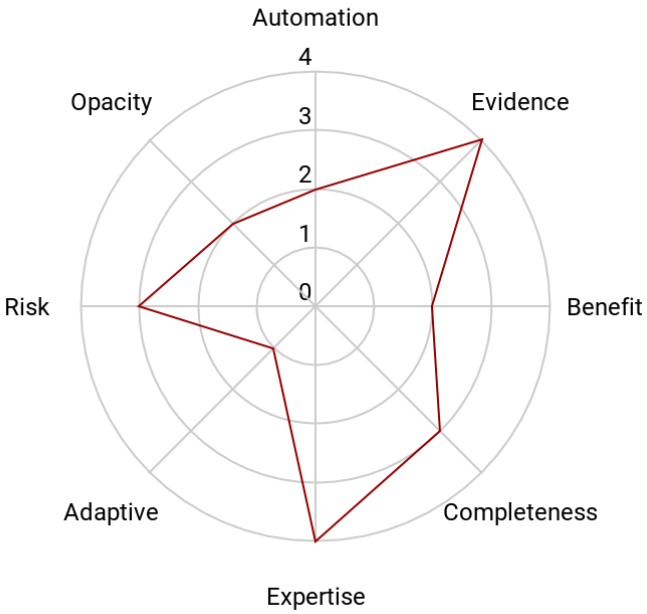
    - Would you make recommendations to change the device or its intended use, i.e. provide more information, a visualisation etc?

Hypothetical Example B1

Description of tool: FOOTFRACFIND uses deep neural networks to assist in the diagnosis of stress fractures in the foot and ankle. FOOTFRACFIND highlights to radiographers suspected stress fractures using a heat map overlay to indicate probability of fracture. Trained on a dataset of confirmed stress fractures, FOOTFRACFIND has performed well in clinical investigations and has recently been given FDA clearance.

a) **Automation (2/4):** FOOTFRACFIND generates its heat map overlay automatically without further human input. However, FOOTFRACFIND only suggests regions of suspected fracture, its labelling making clear that it is assistive and not a replacement for trained radiologists. Radiologists may correct or add further annotations based on their clinical judgment.
b) **Evidence (4/4)**: radiological analysis of stress fractures has a rich literature and clinical investigations conducted for this particular tool appear to be generalisable across populations.
c) **Benefit (2/4)**: FOOTFRACFIND is assistive only, it seeks to augment current workflow and will prove especially valuable as a second assessor. Nevertheless, there is a chronic shortage of qualified radiologists, so any efficiency improvement in workflow would be welcome.
d) **Completeness (3/4)**: FOOTFRACFIND uses radiological data and the clinical judgment inherent in its labelled dataset to produce its heat map. Given this, the tool includes many of the quantifiable variables relevant to diagnosing a stress fracture. However, diagnosis from radiological analysis appears to almost always include unquantifiable variables that require clinical judgment.
e) **Expertise (4/4):** FOOTFRACFIND in its labelling and marketing materials makes clear that the tool is for clinical use only and must be used in conjunction with a qualified radiologist.
f) **Adaptive (1/4):** The machine learning model is static. New datasets may be added and the model retrained but only via a rigorous change management process, the developer having no plans to retrain its model at this stage.
g) **Risk (3/4):** FOOTFRACFIND does not diagnose stress fractures but only assists in their diagnosis. However, if fractures remain undetected this can present with long term, serious complications.
h) **Opacity (2/4):** FOOTFRACFIND uses deep neural networks to produce its heat map. Deep neural networks are not readily human interpretable. Currently, radiologists are only given general information about how the model functions. However, the heat map provides some idea of what the model found significant and the ability to correct recommendations based on clinical judgment renders the model somewhat interpretable.

FOOTFRACFIND



Hypothetical Example B2

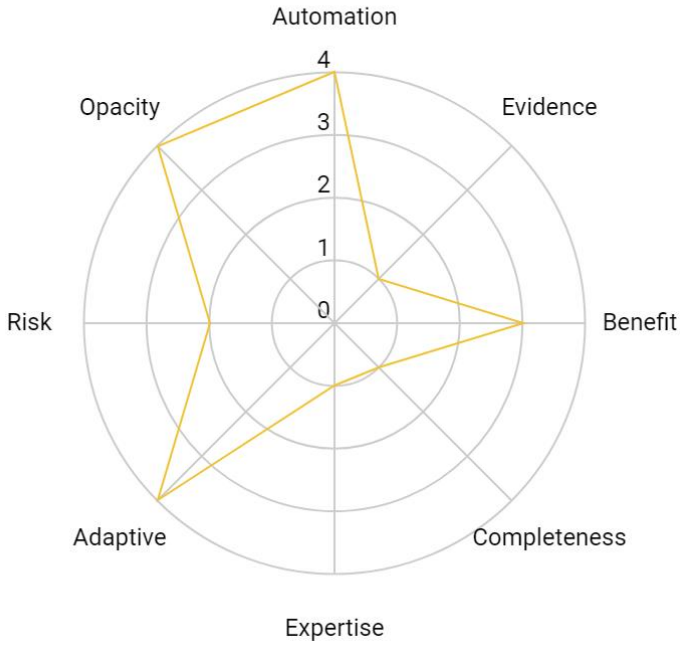
Hello my name is...

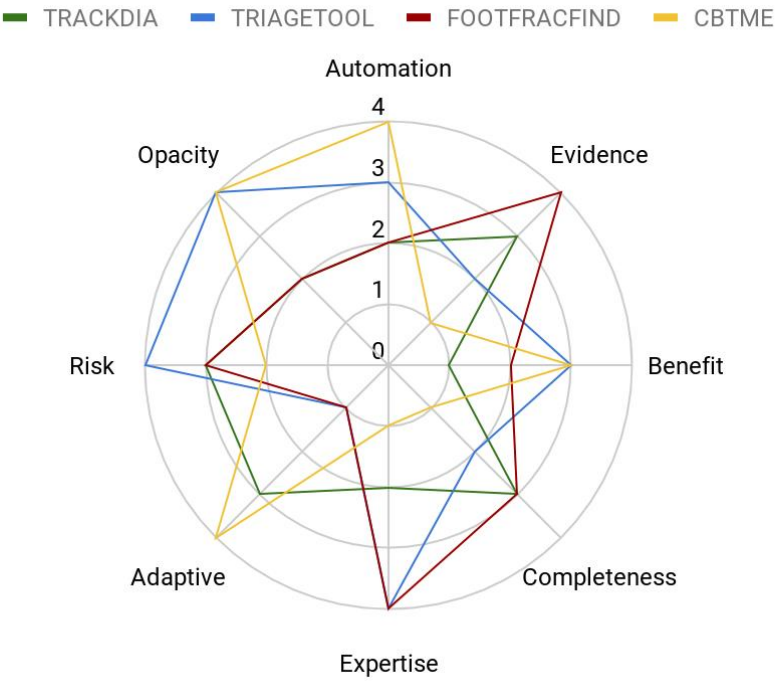
CBTME

Description of tool: CBTME is a text-based chatbot designed to deliver cognitive behaviour therapy to patients. The tool uses multiple machine learning methods, including natural language processing, neural networks, and decision trees to directly interact with patients and administer therapy. Explicitly, CBTME is only for low-risk patients and not for the treatment of serious psychiatric conditions nor those in crisis. Studies examining the clinical effectiveness of the chatbot have produced mixed results.

a) **Automation (4/4):** CBTME is highly automated, the system being marketed as a cost effective tool to assist with the overburdened mental health services. Patients may self-refer or be referred by their GP. Signing up is easy, with only an email and minimal personal details being required. There is no clinician in the loop, and no system apart from general information to refer patients to further mental health services.
b) **Evidence (1/4)**: the tool relies on four tiers of evidence, all of which may be challenged. First, the effectiveness of CBT in general. Second, the effectiveness of digital delivery of CBT. Third, the effectiveness of CBTME and its chatbot system in particular. Fourth, the generalisability across different populations.
c) **Benefit (3/4)**: mental health services are severely stretched. Further, the tool, if effective, can scale to treat many patients. Perhaps modest gains for each patient but across a large patient population.
d) **Completeness (1/4)**: there is a high degree of clinical judgment involved in administering CBT. While the tool includes many quantified variables, there are many variables not represented in the model - the model is highly incomplete.
e) **Expertise (1/4):** Much of the appeal of CBTME is that it requires no clinical input, the tool being marketed directly to consumers and GPs being encouraged to refer patients to the chatbot. The chatbot was initially rolled out on a limited basis in London but is now available to anyone in England and Wales.
f) **Adaptive (4/4):** The machine learning model is highly adaptive, the model utilising streaming data and incremental learning to constantly retrain. Trials of the chatbot were conducted on a locked, static model. Since these trials, the chatbot appears to be relatively stable, the chatbot working largely as intended in the limited London roll out.
g) **Risk (2/4):** CBTME in its labelling and marketing emphasises that the tool is only for low-risk patients. If the chatbot detects 'worrying behaviour' the bot is programmed to display information for further relevant services like suicide hotlines, abuse hotlines, and so on. However, a recent exposé revealed that the tool often fails to recognise those in crisis, delivering patently inappropriate responses.
h) **Opacity (4/4):** CBTME combines multiple machine learning methods to analyse and respond accordingly. The combination of these methods means that it is very difficult to explain exactly how in each instance the tool arrived at the response it issued.

CBTME


Automation
4
3
2
1
0
Evidence
Benefit
Completeness
Expertise
Adaptive
Risk
Opacity



b. Roundtable discussion papers

phg foundation
making science work for health

A right to explanation?

Black box medicine

Discussion paper

UNIVERSITY OF CAMBRIDGE

Authors

Johan Ordish and Alison Hall

Acknowledgements

This work was supported by the Wellcome Trust, grant number: 213623/Z/18/Z

URLs in this paper were correct as of November 2019

This discussion paper can be downloaded from:
www.phgfoundation.org

Published by PHG Foundation
2 Worts Causeway
Cambridge
CB1 8RN
UK
+44 (0)1223 761900

May 2019

Correspondence to:
intelligence@phgfoundation.org

How to reference this publication
A right to explanation?
PHG Foundation (2019)

A right to explanation?

From research that underpins scientific discovery to how we diagnose and ultimately treat patients, machine learning is set to transform healthcare[1]. Machine learning's implementation into practice will, in part, depend upon how this technology is perceived by potential users and patients.

Machine learning models are built upon training and test data. The data processing which underpins the development of machine learning applications and their continued use is therefore key. This data may count as personal data (and sensitive personal data) and be regulated by the General Data Protection Regulation (GDPR). One of the most contentious elements of the GDPR is the right to explanation. The very existence of this right, its interpretation, and how it might be satisfied is contested. This paper outlines the right to explanation and other mechanisms the GDPR provides that might require explanation of machine learning models and their outputs.

Summary

- Machine learning for healthcare is a promising technology but some models may be black boxes - their workings may be opaque
- The GDPR contains a specific right to explanation under Article 22. However, only a subset of machine learning for healthcare will trigger this narrow right
- It is unclear how the right to explanation and transparency requirements will apply to machine learning, key questions include: when does the right apply, what has to be explained, and what kind of explanation would suffice?

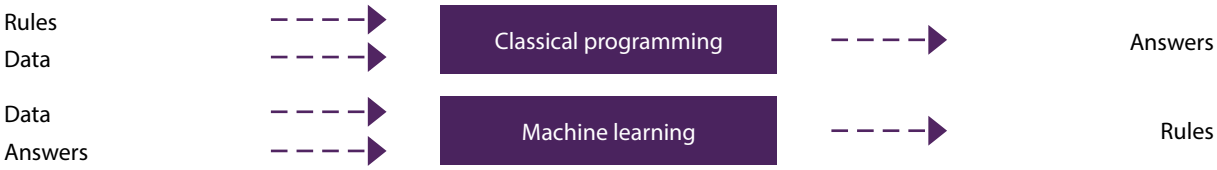
Machine learning for healthcare

Machine learning has many potential applications in healthcare, the table below details three near-implementation applications .

Challenge the tool addresses	Example of a tool	Solution the tool provides
Manual interpretation of radiological images is time consuming	Microsoft Research's Inner Eye	Machine learning for automatic delineation of healthy anatomy from tumours
Diagnosing wrist fractures in a timely manner is difficult	OsteoDetect	AI analysis of wrist radiographs to highlight regions of distal radius fractures
An estimated 1.5 to 3 million people in the UK who attended emergency departments 'could have had their needs addressed in other parts of the urgent care system[2]	Babylon Health's Babylon Check	Automated triage system to route patients to the appropriate service

Machine learning and black boxes

Classical programming combines rules and data to provide answers. Machine learning combines data and answers to provide the rules (see diagram below). Machine learning systems are trained with many examples (data) relevant to the task, the system finding structure in these examples to provide rules to automate the task[3].



A potential disadvantage of using these tools is that many machine learning models may be black boxes, that is, models 'whose internal workings are either unknown to the observer or known but uninterpretable to humans'[4]. In short, it may be difficult to explain why a machine learning model generated a certain output. However, not all machine learning models are human uninterpretable - some techniques are visualizable and so susceptible to human interpretation. Moreover, there are methods to make an otherwise opaque machine learning model somewhat transparent by creating a model-agnostic explanation that approximates the relationship between inputs and outputs, illustrating the model's internal workings. Further, rather than explaining the model as a whole, example-based explanations may be used to explain particular decisions of the model. However, this raises the question: why explain?

Why explain?

A legal obligation to provide an explanation is only one reason to ensure a machine learning model is human interpretable. In the context of healthcare, it might be necessary to explain the workings and contextualise the outputs of a machine learning model for it to be regarded as a viable product and be trusted by clinicians and patients. There may also be an ethical imperative to explain models, especially if models are used for an individual's diagnosis or treatment or for maintaining accountability.

These reasons aside, various sources of law may generate an obligation to explain otherwise human uninterpretable models. Chiefly, medical negligence, medical device law, administrative law, and human rights instruments may individually or collectively generate a duty to explain. In this paper we focus on obligations found in the GDPR.

Duties to explain under the GDPR

The GDPR provides data subjects with at least two potential routes to open black boxes, namely:

I. the right to explanation under Article 22(1); and

II. the general principle of transparency spread across the Regulation but rooted in Article 5(1)(a).

We examine both the right to explanation and the general principle of transparent processing in turn.

Structure of the right to explanation

> Article 22(1) The data subject shall have the right not to be subject to a decision based solely on automated processing, including profiling, which produces legal effects concerning him or her or similarly significantly affects him or her.

This right to explanation is a composite right found across the GDPR. Article 22(1) contains a general prohibition against automated processing. However, most elements referencing explanation are found elsewhere in the rights to information and access, specifically Articles 13(2)(f), 14(2)(g), 15(1)(h) and supporting interpretative aids (recitals).

It is these provisions triggered by Article 22(1) that contain reference to giving 'meaningful information about the logic involved' and the consequences of data processing.

Will my device trigger Article 22(1)?

Not all machine learning for healthcare will be caught by the right to explanation (narrowly interpreted) in Article 22(1). To trigger Article 22(1), the processing of data in question must be:

I. based solely on automated processing; and

II. produce legal effects concerning or similarly significantly affects the data subject.

Working Party 29 Guidelines on Automated individual decision-making elaborate on each of these elements:

'Based solely on automated processing' means there is no human involvement in the decision process. However, this human involvement cannot be 'fabricated' and must be more than a token gesture - the human must have actual authority and influence over the decision.

'Legal effects' means the decision affects the data subject's legal status, legal rights, or rights under contract.

'Similarly significant affects' means that the decision must have similar significance to legal effects, being sufficiently important to be 'worthy of attention.' Recital 71 gives some examples: 'e-recruiting practices without human involvement' and 'automatic refusal of online credit applications.'

While the above guidance is vague, it is clear that only a subset of near-use machine learning for healthcare will either be solely automated and also produce legal/similarly significant effect (see table below). Only machine learning devices in Category A will engage the narrow right to explanation in Article 22(1).

Triggering Article 22(1)	Based solely on automated processing	NOT based solely on automated processing
Produces legal effects or similarly significant affects	(Article 22(1) triggered) A	B
Does NOT produce legal effects or similarly significant affects	C	D

Transparency apart from Article 22

Article 22 is not the only mechanism under the GDPR that might generate a duty to explain machine learning models and their outputs. The general principle that personal data be 'processed lawfully, fairly and in a transparent manner' underpins the rights to information, access, and other GDPR rights[5]. In short, general principles of transparency and the need to meet other GDPR rights may necessitate explanation, even if this requirement is less onerous than that found in Article 22.

What does transparency in general require? The rights to information and access, accompanied by their recitals will likely require some explanation of machine models as a whole. More controversially, these general transparency requirements may require explanation of a specific decisions and processing of machine learning models. While it is unclear what a duty to explain under the general principle of transparency might require, it is clear that explanation of specific decisions would be a more demanding requirement.

What does a duty to explain require?

The proper interpretation of the right to explanation and its relation to the broader principle of transparency is highly contentious. These interpretative debates have real consequences for what the GDPR will require in terms of explanation of machine learning. Broadly, there are those that emphasise the human rights pedigree of the GDPR, noting that the purpose of the right to explanation is to vindicate more general rights to transparency[6]. These commentators typically think that the right to explanation can require explanation of systems as a whole but as well as individual decisions.

On the other hand, there are those that base their interpretation on the gradual evolution of the GDPR from the Data Protection Directive, drawing a sharp distinction between the interpretative recitals and the legally effective articles of the GDPR[7]. These commentators typically think that the right to explanation does not require explanation of individual decisions and prefer to call the right a 'right to be informed' instead.

These interpretative disputes over the right to explanation and transparency have deep implications for determining what the duty to explain requires. Major uncertainties include:

When is explanation required? Is explanation required before the data is processed and/or after processing?

What is to be explained? Must data controllers explain the model and how it functions as a whole and/or must they provide an explanation of individual decisions post-processing?

What kind of explanation is required? Might counterfactual explanations (that describe the nearest possible world where the result sought was obtained) suffice[8]?

The GDPR's right to explanation and transparency requirements were implemented to protect data subjects and foster good data protection practice. However, the interpretation of these requirements and how they apply to machine learning is in a state of chronic uncertainty. This uncertainty threatens to undermine the goals of the Regulation and acts as a barrier to the development and implementation of machine learning for healthcare. Further guidance clarifying the above questions is urgently needed.

References

1. Ordish J, Hall A. Algorithms as Medical Devices. PHG Foundation. 2019.

2. NHS England: Next Steps on the NHS Five Year Forward View. 2017.

3. Chollet F. Deep Learning with Python. Manning Publications; 2017: 2-3

4. Guidotti R, Monreale A, Ruggieri S, et al. A Survey of Methods for Explaining Black Box Models. ACM Computer Surveys. 2019; 51(5): 1-42

5. Article 5(1)(a), Articles 13-15. Regulation (EU) 2016/679 GDPR

6. Selbst AD, Powles J. Meaningful Information and the Right to Explanation. International Data Privacy Law. 2017; 7(4): 233-242

7. Wachter S, Mittelstadt B, Floridi L. Why a Right to Explanation of Automated Decision-Making Does Not Exist in the General Data Protection Regulation. International Data Privacy Law. 2017; 7(2): 76-99

8. Wachter S, Mittelstadt B, Russell C. Counterfactual Explanations Without Opening the Black Box: Automated Decisions and the GDPR. Harvard Journal of Law & Technology. 2017; 31(2): 1-44

The PHG Foundation examine the right to explanation and related questions in our Wellcome Trust funded project Black Box Medicine and Transparency.

PHG Foundation
2 Worts Causeway
Cambridge
CB1 8RN
+44 (0) 1223 761900

@phgfoundation
www.phgfoundation.org

phg foundation
making science work for health

Why explainable machine learning matters for health

Discussion paper

UNIVERSITY OF CAMBRIDGE

Authors

Johan Ordish and Alison Hall

Acknowledgements

This work was supported by the Wellcome Trust, grant number: 213623/Z/18/Z

URLs in this paper were correct as of November 2019

Published by PHG Foundation
2 Worts Causeway
Cambridge
CB1 8RN
UK
+44 (0)1223 761900

May 2019

Correspondence to:
intelligence@phgfoundation.org

How to reference this publication
Why explainable machine learning matters for health
PHG Foundation (2019)

Why explainable machine learning matters for health

Machine learning promises to change the way we diagnose and ultimately treat patients. However, some argue that machine learning for health also threatens to usher in an age of black box medicine, where 'opaque computational models make decisions related to healthcare.'[1] This paper explores the use of machine learning for health, stating to what extent and why machine learning models might be opaque, noting why human interpretability of machine learning matters, and outlining the different ways in which machine learning models can be interpretable to humans. We conclude that while not all machine learning models are black boxes, interpretability of machine learning models will often be important when providing proper assurances that a model is safe and effective.

Summary

- Machine learning models vary in the extent to which they are interpretable - ranging from those that are intrinsically human interpretable to black boxes that are not intrinsically interpretable to humans
- Black box models may be made somewhat human interpretable through the use of post hoc explainers that explain a particular decision of the model (local interpretability) and/or how the model functions generally (global interpretability)
- Post hoc explainers have weaknesses and are ultimately only an estimation of an underlying black box model. Given this, it is unclear where and when we should demand the use of intrinsically interpretable machine learning

A call for interpretability - two examples to demonstrate why interpretability is important:

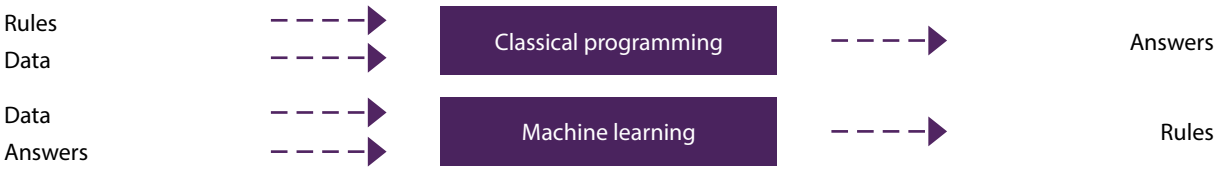
Caruana *et al* (2015) describe a series of models to predict the probability of death for patients with pneumonia.[2] The group found that neural networks produced the most accurate models. However, when the group trained in parallel a less accurate but interpretable rule-based model, the group found that this model learned the following rule: 'HasAsthma(x) ⇒LowerRisk(x).' Consequently, it was shown that a confounding variable influenced the neural networks, the models correctly identifying that those with asthma were less likely to die but only because as a group they were more likely to receive treatment.

Zech *et al* (2015) trained a convolutional neural network to screen for pneumonia using x-rays.[3] Subsequent manual image review noticed that the model was able to differentiate between those x-rays taken by portable scanner (identified by the word 'portable' and inversion of colour in the x-ray) and those by static scanner, the model finding this distinction significant, portable scanners being used in the emergency department but not for inpatient units. Consequently, when the model found the word 'portable' significant it introduced a potentially confounding factor into the screening process.

These two examples illustrate that the mere fact of accuracy may be insufficient; that it is important to know why machine learning models are accurate.

What is machine learning?

Machine learning describes an approach to programming that typically produces algorithms with bounded, task-specific intelligence. In a phrase, machine learning algorithms are narrowly intelligent (they do one thing well), but broadly unintelligent (lacking broad capacity to reason). How does machine learning differ from classical programming? Classical programming combines rules and data to provide answers. Machine learning combines data and answers to provide the rules (see diagram below). Machine learning models are trained with many examples (data) relevant to the task, the system finding structure in these examples to provide rules to automate the task.[4]



A potential disadvantage of using these tools is that some machine learning models may be black boxes.

Black boxes and human interpretability

Black box models are models 'whose internal workings are either unknown to the observer or known but uninterpretable to humans.'[5] In short, because the model has been trained rather than explicitly programmed it may be difficult to explain why a machine learning model generated a certain output or to understand what the model finds significant.

Core to the definition of 'black box' is interpretability. Interpretability has different definitions across different domains. In the machine learning context, a useful definition comes from Miller (2017):

> '*Interpretability* is the degree to which a human can understand the cause of a decision.'[6]

Why is interpretability important? How does the problem of interpretability arise?

Interpretability and incompleteness

Arguably, the problem of interpretability arises because many machine learning problems are (often necessarily) incomplete.[7] In this sense, an incomplete problem leaves a gap which creates the need for a machine learning model to be interpretable.

We should distinguish between the following:

- *Uncertainty*, meaning quantified variance that can be formalised

  For example, false positive rates, confidence intervals

- *Incompleteness*, meaning the problem includes elements that have not been formalised and quantified

  For example, scientific discovery and diagnosis make causative inferences, but causation cannot be formalised or quantified, so scientific discovery and diagnosis remain an incomplete problem for machine learning to address

Many useful machine learning models in health tackle problems that are incomplete and will remain incomplete for the foreseeable future. As a consequence, machine learning for health will have a gap which interpretability may have to bridge. Furthermore, even where a machine learning problem is virtually complete, its implementation as a device or service may still require interpretability. For instance, concepts like 'safety' cannot fully be quantified and so if a model requires some form of human input or the results need to be put into further context by a human, the interpretability of the model may remain important.

When to demand interpretable machine learning

When should we require that a machine learning model be human interpretable? Doshi-Velez *et al* (2017) suggest that explanation of an incomplete model may be unnecessary where either a) 'there is no significant consequences for unacceptable results' or b) 'the problem is sufficiently well-studied and validated in real applications that we trust the system's decision, even if the system is not perfect.'[8] In the health sector, this might mean that interpretability may be less important for machine learning models that are 'lifestyle/wellbeing devices' (devices that do not have a medical purpose and pose little to no risk to the user).[9] Further, the better studied a problem is, the more confidence we might have in the model and our ability to check for any artefacts or confounding factors, and so the less we might lean on its interpretability. In this way, the closer machine learning models for health get to models that are well-understood and tested (e.g. aircraft avoidance systems), the more content we might be with the accuracy and general evidence base making up for their lack of interpretability.

One concern with Doshi-Velez's account of incompleteness and interpretability is that the account may not fully capture the need to make machine learning models interpretable for other reasons such as fairness, privacy, and other related rights. In this way, we can imagine a fully complete model that produces perfect results but is totally opaque. Arguably, in this situation, an ethical (and possibly legal) obligation to render the model interpretable may still exist. After all, the model may still process sensitive data that may significantly impact the data subject in question.

Dimensions of interpretability

There are different ways in which a machine learning model can be human interpretable or made interpretable. Methods used to make machine learning interpretable can be categorised using a number of different criteria[10]:

- *Intrinsic interpretability or post hoc interpretability* - Is the model intrinsically interpretable due to its simple structure or is a post hoc ('after the fact') method to render the model interpretable necessary?
- *Model-specific or model-agnostic* - Is the explanation specific to the model in question (as with intrinsic interpretability) or is the explanation tool model-agnostic, meaning it can be applied

(theoretically) to explain any machine learning model?

- *Global or local* - Does the interpretability method explain how the model functions in general (global), just an individual decision (local), or a mixture of both?

These dimensions of interpretability as well as their various weaknesses and strengths are outlined below.

Intrinsically interpretable models

Not all machine learning models are black boxes - some techniques are relatively simple and susceptible to human interpretation. The challenge with these models often concerns the communication and visualisation of their decision processes. For instance, decision trees, rules, and linear models are generally recognised as being easily understandable and interpretable for humans[11]. So long as the decision process is accessible to the user, the way the model functions and how a particular decision was arrived at will be human interpretable.

Why not always use intrinsically interpretable models?

The most common argument against insisting upon only intrinsically interpretable models is the supposed trade-off between accuracy and interpretability[12]. Some argue that there is an inverse relationship between the accuracy of a machine learning model and the interpretability of that model. However, this relationship is contentious, some noting that this general proposition remains unevidenced.[13] Nevertheless, it is true that the computational goal of building the most accurate model is not exactly the same as building the most interpretable model.[14] While there might not be an inverse relationship between accuracy and interpretability – in fact, the two concepts can often operate in tandem – there may be a tradeoff to be made between the two at some point.

Post hoc interpretability

If a model is not intrinsically interpretable it is a black box model. These models can be rendered somewhat human interpretable by using methods such as post hoc explainers. These post hoc explainers can explain the overall model (global) or the specific decisions of that model (local) or a combination of both. They may be specific to a particular machine learning model or be model-agnostic, being able to be bolted onto many different models (see table below).

	Global	**Local**
Model-agnostic	Post hoc explanations explain the general function of any given machine learning model	Post hoc explanations explain the specific decisions of any given machine learning model
Model-specific	Post hoc explanations explain the function of the model but are specific to only this machine learning model	Post hoc explanations explain the specific decision of the model but are specific only to this machine learning model

Generally, model-agnostic explainers work by treating the underlying machine learning model as a black box, testing the relationship between inputs and outputs to approximate a view of what the model finds significant generally or in relation to a particular decision. Model-agnostic methods are particularly powerful as they can often be bolted on to elucidate the inner workings of what would otherwise remain an opaque machine learning model.

The weaknesses of black box models and black box explainers

Post hoc explainers to interpret black box models are promising but have limitations. There are three main issues with using post hoc explainers to explain black box models:

1. **Fidelity**. post hoc explainers often approximate the underlying machine learning model to explain its contents. Since these explainers estimate the underlying model they may provide inaccurate answers, especially if these explainers are highly localised and taken outside their local context[15]
2. **Partial explanations.** Even if the post hoc explanation generated is correct, it may be incomplete and (potentially) instil a false sense of confidence.[16] For example, saliency maps provide a heatmap overlay of an image, demonstrating what part of the image the model found relevant. However, knowing where the model is looking does not tell us what the model is doing with that part of the image
3. **Calibration of machine learning models.** If the underlying machine learning model is a black box, it is difficult to calibrate the model in light of external information not input into the model.[17] If contextual information informs the data underpinning the model, it is often not possible to manually calibrate models that use convolutional neural networks to take account of this discrepancy. For instance, suppose we know that our dataset has a racial bias. If the model trained using this data is a black box it is difficult to manually adjust for this discrepancy without removing data points

Following these three weaknesses, authors like Rudin (2019) emphasise that the gains in interpretability by using intrinsically interpretable machine learning often exceeds the cost of reduced accuracy.[18] That is, while the accuracy loss in choosing an intrinsically interpretable model is low, the gain that interpretability brings usually outweighs this loss. This underlines the point that post hoc explanations for black boxes are not a shortcut to interpretability - they are imperfect and inappropriate in some circumstances.

Why interpretability matters for health

There are special reasons to make machine learning for health human interpretable as a) many of the machine learning problems in the sector will be incomplete, and b) many machine learning applications risk serious consequences if unacceptable results are returned. There are strong practical reasons to provide interpretable models to assure users, regulators, and commissioners that the model is safe and effective. Apart from this, many machine learning models will also process health (or health-related, biometric, or genetic) data, meaning that they will process sensitive personal data to draw their conclusions. Given this, and the importance of the decision at stake, there may be a strong ethical (and possibly legal) imperative to provide an explanation to the user, whether that be a clinician or a patient. In summary, there are often strong practical, ethical, and legal reasons to explain machine learning models.

Further questions

- When, if at all, should we demand the use of intrinsically interpretable machine learning models in health?
- Are post hoc methods to interpret a black box model appropriate for machine learning models that might have serious implications?
- Which of these explanations might satisfy the GDPR's right to explanation? See *A right to explanation* for more information
- Which types of explanation might be most appropriate for patients, clinicians, or consumers?
- Which types of explanation might be most appropriate to generate the trust and confidence of health care professionals who might rely on machine learning applications?

References

1. Price W.N. Black-Box Medicine. Harvard Journal of Law & Technology. 2015; 28(2): 420-467.
2. Caruana R. *et al.* Intelligible Models for Healthcare: Predicting Pneumonia Risk and Hospital 30-day Readmission. KDD 2015. 2015; 1721-1730.
3. Zech JR. *et al.* Confounding variables can degrade generalization performance of radiological deep learning models. PLoS Medicine. 2015; 1-15.
4. Chollet F. Deep Learning with Python. Manning Publications. 2017; 2-3.
5. Guidotti R. *et al.* A Survey of Methods for Explaining Black Box Models. ACM Computer Survey. 2019; 51(5): 1-42.
6. Miller T. Explanation in Artificial Intelligence: Insights from the Social Sciences. Artificial Intelligence. 2017; 14.
7. Doshi-Velez F. *et al.* Towards A Rigorous Science of Interpretable Machine Learning. 2017; 1-13.
8. Ibid.
9. Regulation (EU) 2017/745 on medical devices. Recital 19.
10. Molnar C. Interpretable Machine Learning: A Guide for Making Black Box Models Interpretable. 2019; 26-27.
11. Guidotti R. A Survey of Methods for Explaining Black Box Models.
12. DARPA. Broad Agency Announcement: Explainable Artificial Intelligence. DARPA Information Innovation Office. 2016; p14.
13. Rudin C. Please Stop Explaining Black Box Models for High Stakes Decisions. NIPS 2018. arXIV:1811. 10154: 14.
14. Dwork C, *et al.* The Algorithmic Foundations of Differential Privacy. Foundations and Trends in Theoretical Computer Science. 2014; 9(3): 211.
15. Guidotti R. A Survey of Methods for Explaining Black Box Models.
16. Rudin. Please Stop Explaining Black Box Models for High Stakes Decisions. NIPS 2018. arXIV:1811. 10154: 4.
17. Ibid. p5.
18. Ibid. p3.

PHG Foundation
2 Worts Causeway
Cambridge
CB1 8RN
+44 (0) 1223 761900

@phgfoundation
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