phg foundation
making science work for health

Black box medicine and transparency

Ethics of transparency and explanation

A PHG Foundation report for the Wellcome Trust

UNIVERSITY OF CAMBRIDGE

Authors

Johan Ordish, Tanya Brigden, 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

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, Hall A. *Black Box Medicine and Transparency: The Ethics of Transparency and* Explanation. PHG Foundation. 2020.

Contents

1. The Ethics of Transparency and Explanation[i]

Conversations on the ethics of AI/machine learning and digital health often bemoan the lack of trust from publics and users.[1] Tentatively, surveys vindicate such concerns, many finding that consumers tend to be sceptical of digital health. For example, Baker McKenzie's *Outside the Comfort Zone* survey found that only 47% of consumers trust digital health products overall and that algorithmic diagnosis tools are treated with particular suspicion, only 7% of consumers trusting such tools.[2] This general mood of scepticism paired with equally bleak surveys on public understanding of AI and machine learning, suggests that promising applications of machine learning for healthcare and research could be undermined or curtailed by misunderstanding, scepticism, or poor messaging.

In response to this crisis of trust, international initiatives like the Institute of Electrical and Electronics Engineers' (IEEE) *Ethically Aligned Design*,[3] supranational initiatives like the European Commission's High-Level Expert Group on Artificial Intelligence,[4] and national initiatives like NHSX's Principle 7 guidance prescribe more trust sought through increased transparency.[5] In this report we argue that this prescription requires pause, heeding the cautionary words of O'Neill in her Reith Lecture *Trust and Transparency*: "A crisis of trust cannot be overcome by a blind rush to place more trust."[6] Accordingly, we examine the nature of transparency, its limits, and the adjunct value of explainability as well as related terms. This is necessary to consider whether the response to any crisis of trust is being aimed at the proper values. That is, whether the correct regime has been prescribed for a dearth of trust.

We pose two distinct arguments with regards to transparency and explanation. We argue that transparency, while valuable, often results in mass disclosure of information without regards for one's audience in lieu of genuine communicative action. As a result, we note that transparency has limits. Further, we also caution against seeking trust as an end in of itself - to seek trust in the absence of trustworthiness is pernicious. Consequently, we consider the philosophy of explanation and explanatory pragmatism in particular, to consider how disclosure might be best tailored to each audience.

We argue that a proper account of the act of explanation is highly context sensitive, requiring knowledge of at least a) what being explained (the *explanandum*) and b) an understanding of *why* the explanation is sought. We note that in the context of machine learning for healthcare and research, there are i) multiple different stakeholders ii) seeking explanations of different phenomena iii) for different reasons. Accordingly, explanation of machine learning models is met with different standards, depending on the context in which the explanation is sought.

This report has several purposes:

I. To provide a conceptual understanding and normative account of transparency
II. To provide a conceptual understanding and normative account of explanation
III. To outline the variety of reasons why we might request or demand explanations of machine learning models in the context of healthcare or research

Accordingly, this report provides an ethical standard by which we might consider the legal requirements of the GDPR and its supposed 'right to explanation' considered in the Regulating Transparency report.

[i] We are grateful for the assistance of Dr Rune Nyrup (Leverhulme Centre for the Future of Intelligence) who discussed and contributed to ideas in this report.

2. Transparency and related terms

Literature on explainable machine learning (XAI) centres around a number of key terms like 'interpretability,' 'explainability', 'comprehensibility' and 'legibility.' This literature also invokes a number of related concepts like 'predictability,' 'fairness,' 'generalisability,' and 'trust' to bolster the importance of interpretability, explainability, and related terms. In parallel is the proliferation of AI ethics guidelines. In this literature, Jobin et al note that there is emerging consensus around five ethical principles:[7]

- Transparency, including related terms: explainability, explicability, understandability, interpretability, communication, disclosure, and showing
- Justice and fairness, including related terms: consistency, inclusion, equality, equity (non-)bias, (non-)discrimination, diversity, plurality, accessibility, reversability, remedy, redress, challenge, access, and distribution
- Non-maleficence, including related terms: security, safety, harm, protection, precaution, prevention, integrity (bodily or mental), and non-subversion
- Responsibility, including related terms: accountability, liability, acting, and integrity
- Privacy, including related terms: personal or private information

The Jobin et al analysis provides a good basis for understanding consensus in regards to ethics guidelines. Still, the project is empirical, reporting the prevalence and clustering of ethical concepts in the artificial intelligence space. While this work is certainly valuable, often the terms included are given divergent definition and interpretation. This report seeks to clarify and demarcate the limits of transparency and explainability, considering what lessons the literature provides in relation to each.

3. Transparency

It is common in the literature on machine learning in healthcare to see calls for transparency and trust. One notable example is the EU High-Level Expert Group on Artificial Intelligence (AI HLEG) includes transparency as an important element in delivering trustworthy AI.[8] This principle is further broken down into three parts:

> "Traceability" includes, amongst other elements, 'data sets and the processes that yield the AI system's decision, including those of data gathering and data labelling as well as the algorithms used, should be documented to the best possible standard to allow for traceability and an increase in transparency.'[9]
>
> "Explainability" concerns the 'ability to explain both the technical processes of an AI system and the related human decisions (e.g. application areas of a system).'[10]
>
> "Communication" notes that 'AI systems should not represent themselves as humans to users; humans have the right to be informed that they are interacting with an AI system.'[11]

The AI HLEG is by no means alone in stressing the connection between transparency and trust. Indeed, as mentioned, Jobin et al identify 'transparency' as one of the most common principles appearing in AI ethics statements.[12] Further, the Joint Research Centre of the European Commission, in their report *Artificial Intelligence: A European Perspective,* outline transparency of AI as a major societal challenge for algorithms and automated decision making.[13] From an international perspective, the IEEE in their *Ethically Aligned Design* consider transparency to be a major challenge for autonomous and intelligent systems, linking transparency with other concepts like accountability.[14] This begs the question: what is 'transparency' and what is its relationship to other concepts such as explanation, trust, and accountability?

a. What is transparency?

The meaning of transparency is contested. One of the most common methods to define transparency is to counterpose transparency with similar terms to test where the boundaries of transparency lie.[15] In this regard, there is little agreement in the literature. For instance, some in the literature on governance and transparency, counterpose transparency with openness, others, transparency with trust. In regards to transparency and openness, Larsson distinguishes the two terms, noting that transparency goes beyond openness to also include comprehensibility.[16] With respect to transparency and trust, some note that transparency contributes positively to trust by building credibility.[17] However, others argue to the contrary, perhaps the most prominent being O'Neill's caution that a myopic focus on transparency can actually increase deception and erode trust.[18] As we outline below, transparency on O'Neill's account does not necessarily require that disclosure be interpretable, accessible, or take into account those to whom it is addressed.

We make a tentative observation that the drive to interpretability and explainability may be driven by a crisis of trust in machine learning systems. However, we echo O'Neill's warning: 'A crisis of trust cannot be overcome by a blind rush to place more trust.'[19] We must direct our efforts toward trustworthiness, providing observers and the public with the ability to trust *intelligently*.[20] We analyse this message below.

b. Transparency's limits

Transparency works by making the 'very evidence needed to place or refuse trust intelligently more available.'[21] Where appropriately sought, transparency is a means to support *trustworthiness* rather than merely as a means to secure *trust*.[22] However, even transparency appropriately sought has its limits. In O'Neill's words, transparency 'has few enemies but offers fewer and more limited benefits than is widely assumed.'[23] As conceived by O'Neill, transparency only requires that specified 'types of informational content be disclosed' - nothing in the concept requires effective communication.[24] In this way, transparency requirements by themselves are too little for others to *trust intelligently*.[25]

On O'Neill's account, transparency and its requirements only require disclosure but do not in and of themselves require that this disclosure be relevant or accessible to their audiences.[26] According to O'Neill, transparency alone can - and often does - fail as a *communicative act*. In the context of governance and legislation like the Freedom of Information Act 2000, O'Neill notes that transparency emphasises mass dissemination of information often with little regard for their audience.[27] Accordingly, we note that within the philosophy of explanation literature, explanatory pragmatism in particular may prove instructive when considering how best to take regard of one's audience.

Section 3 key messages:

- **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 the concept of transparency and its limitations if we are to consider what the concept might require of machine learning for healthcare and research.**
- **Transparency is best analysed as a distinct concept that does not necessarily incorporate ideas of accessibility, communication, and interpretability.**
- **Transparency should be viewed as a means to secure trustworthiness, not trust. To seek trust but not trustworthiness is pernicious. In many cases, transparency should result in disclosure of information that assists users and the public to intelligently place trust in the machine learning system.**
- **Transparency has limits. Transparency emphasises disclosure but often misses the importance of communication and the virtues associated with communicative acts, that is, accessibility, interpretability, and interest-sensitivity.**
- **The importance of communication in the context of understanding machine learning models may be found in the philosophy of explanation literature, especially pragmatist accounts.**

4. What is explanation?

The question of what counts as an explanation is often interpreted as asking: 'what information has to be conveyed in order to have explained something?'[28] That is, there is an attempt to formally describe a general concept of explanation. This was the dominant approach in the philosophy of science during the 20th Century. Undoubtedly, these methods offer important contributions to the literature and so are rightfully emphasised by meta-studies such as Miller's 2019 paper.[29] However, we think that there are valuable lessons for explanation of machine learning in healthcare and research to be found in the literature on explanatory pragmatism and explanation as an act. In this regard, we seek to address what the components of a successful explanatory act are and what goods explanation provides in Section 6.

It is relatively uncontroversial to say that an explanation has at least two elements:[30]

I. The *explanandum*, or that which is explained.[31] This fact may be a *particular* fact, for example, the explosion of the Challenger space shuttle.[32] Alternatively, the fact may be *general*, for example, the law of conservation of linear momentum.[33]
II. The *explanans*, or that which does the explaining.[34] To quote Salmon, it 'consists of whatever facts, particular or general, are summoned to explain the *explanandum*.' In the case of the Challenger disaster event, this may be the failure of the O-ring seals.

Notably, explanations focus on limited aspects of the event or phenomena in question. When explaining the Challenger disaster, we are not interested in the fact that the astronaut was a woman, her previous occupation, and so on - the full 'richness and complexity' of the event is compressed to just what is needed to explain.[35] Having concrete ideas of what the precise *explanandum* is - what is to be explained - will likely be helpful when crafting a successful explanation.

In the context of machine learning for healthcare and research, given the diversity of both fields, there are a great many *explanandums* to be explained. For instance, perhaps the machine learning model itself is *the* fact to be explained. Alternatively, perhaps our inquiry is more limited, perhaps we want to explain only a particular output of a model. Moreover, especially in the arena of research, the machine learning model might itself be the *explanans* - the fact which does the explaining. As far as possible, we should be clear what we seek to explain, this will assist when figuring out the precise *explanandum* sought and the appropriate *explanans*.

a. General epistemology

There are many accounts of explanation. The philosophy of explanation is inherently tied to (and often dominated by) the specific question of what constitutes *scientific explanation*.[36] However, scientific explanation is not synonymous with *explanation simpliciter*. There are many different kinds of explanation in addition to those classified as 'scientific.'[37] Ordinary explanations need not be cast in 'scientific' terms. We can explain the snow's melting not by reference to a natural law but by the oncoming of warmer weather.[38] We can ask a poet to explain the meaning of their poem.[39] In these cases - and in many more - the explanation proffered has no obvious link to science, nor is it scientific by nature. Accordingly, the analysis of explanation belongs to general epistemology, not merely to the philosophy of science.[40] Locating explanation in the domain of general epistemology is important because the kind of explanation we seek of machine learning models is not necessarily a scientific explanation. It depends on the kind of explanation sought - the kind of questions we wish to ask, and why we

ask them. Depending on the explanation sought and the tools we seek to explain, perhaps an ordinary, everyday explanation is sought, in which case the explanation sought should be analysed and held to the standard of general epistemology.

Miller in his seminal metanalysis of explanation in social science, notes we should in fact be seeking an 'everyday explanation' from machine learning rather than 'scientific explanation.'[41] Following Miller, everyday explanations may be distinguished from scientific explanations on the basis that *everyday explanations* do not address general, fundamental laws. In this way, everyday explanations seek answers to why particular facts (events, properties, decisions, etc.) occur.[42] Mittelstadt sums up this thinking:

> 'The explanations requested [in the context of machine learning] are thus not full scientific explanations, as they need not appeal to general relationships or scientific laws, but rather at most to causal relationships between the set of variables in a given model.'[43]

The following analysis focuses on forms of explanation often overlooked or distinguished by Miller's analysis.

i. General features of explanation

Lipton outlines three features of explanation that might assist in getting us to the question of what the goods of explanation are.[44] These features are:

I. The gap between knowledge and explanations
II. The why regress: answering 'why' questions with 'why' questions
III. Self-evidencing explanations

These features illuminate key attributes of explanation before we consider particular accounts of explanation.

With respect to the gap between knowledge and explanation, regardless of whether the explanation is ordinary, scientific, or something in between, explanation is often sharply distinguished from knowledge.[45] To take Lipton's example, we all know that the sky is blue, but few of us understand *why*.[46] Indeed, explanations often assume the truth of a statement or the existence of a phenomenon - "Why does a phenomenon occur?" implies that the phenomenon does indeed occur. We also recognise that there is some separation between an explanation and the truth of the premises upon which it relies. Indeed, the quality of explanation may be perfectly good even if its premises are false. Consequently, there is a gap between what we know and what we can explain.

Considering the why regress, Lipton observes that children have a habit of responding to why-questions with more 'whys.'[47] This observation is instructive, it highlights that explanations may answer one why-question, but not further questions 'up the ladder' for which no answer is provided.[48] The nub of the lesson is best stated by Lipton:

> 'This shows that understanding is not like some substance that gets transmitted from explanation to what is explained, since the explanation can bring us to understand why what is explained is so even though we do not understand why the explanation itself is so.'[49]

We consider the related question of partial explanations below in Section 8(b).

In regards to self-evidencing explanations, Lipton notes that some explanations are circular in that they take the form: 'H explains E while E justifies H.'[50] To take Lipton's example:

> 'Seeing the disemboweled teddy bear on the floor, with its stuffing strewn throughout the living room, I infer that Rex has misbehaved again. Rex's actions provide an excellent if discouraging explanation of the scene before me, and this is so even though that scene is my only direct evidence that the misbehaviour took place.'

In this way, we infer a hypothesis is correct because it best explains the facts. Lipton thinks that any account of explanation and indeed of understanding must fit with these key features of explanation.

b. Scientific explanation

Not all explanations are scientific explanations. Not all explanations we seek of machine learning models are scientific in nature either. However, some explanations of machine learning may indeed be scientific or at least directed toward scientific ends. In this regard, especially in the context of research, it may be important that a machine learning model be interpretable - either to draw explicit links to phenomena and general laws or have its conclusions susceptible to human inference. Indeed, some more speculative areas of machine learning research consider machine learning's relationship to such laws, for example, Pearl's work on machine learning and causality.[51] In short, we should not disregard scientific explanation's relevance to machine learning.

We briefly outline and consider four relevant models of 'scientific explanation': the deductive-nomological model, deductive-statistical explanation, inductive statistical explanation, and the statistical relevance model.

The deductive-nomological model (D-N model) of scientific explanation argues that a successful explanation is one where the *explanandum* is a logical consequence of the *explanans* and the sentences constituting the *explanans* are true.[52] In this way, the explanation is a form of sound deductive argument, the *explanandum* necessarily following from the *explanans*.[5354] Further, following Hempel, the *explanans* must contain at least one necessary 'law of nature.'[55] Accordingly, the D-N model not only stipulates the logical relationship between *explanans* and *explanandum* but also notes that the explanation should reference general laws.

Following the same general pattern as the D-N model is deductive-statistical explanation (D-S explanation). D-S explanation 'involves the deduction of a narrower statistical uniformity from a more general set of premises, at least one of which involves a more general statistical law.'[56] In this way, statistical uniformity is deduced from a more general statistical law. This kind of statistical explanation is very different from inductive statistical explanation.

Where D-S explanation involves deduction, inductive statistical explanation (IS explanation) attempts to subsume individual events under statistical laws.[57] To take Woodward's example, suppose we seek to explain the individual event of a patient recovering from a streptococcus infection by invoking a statistical law, in this case, the probability of recovery after penicillin has been administered.[58] Following Woodward, we cannot deduce that any given individual will recover following penicillin in accordance with the statistical law - the most we can say is that recovery is more or less probable. Consequently, an 'IS explanation will be good or successful to the extent that its *explanans* confers high probability on its *explanandum* outcome.'[59]

The statistically relevant model (S-R model) is different again. Broadly, the S-R model infers that if a property is statistically relevant, then that property is explanatorily relevant, the

converse also being true: statistically irrelevant properties are explanatorily irrelevant. On Salmon's account (as described by Woodward):

> 'Given some class of population *A*, an attribute *C* will be statistically relevant to another attribute *B* if and only if $P(B|A.C) \neq P(B|A)$ - that is if and only if the probability of *B* conditional on *A* and *C* is different from the probability of *B* conditional on *A* alone.'[60]

We consider D-S, I-S, and S-R explanation further at Section 8(d).

The above accounts of scientific explanation illustrate that there is no one form of explanation, even in the relatively narrow class of 'scientific explanations.' More importantly for our purposes, the above accounts may vary in their applicability to the set of techniques discussed in the Interpretable Machine Learning report. For instance, the S-R model seems a natural fit for many of the post hoc explanation techniques like partial dependence plots.[61] In this way, techniques like this seek to answer what features are statistically relevant. However, the logic-based D-N model may fit less well with these techniques. Artificial intelligence techniques once focused on rule-based methods of linguistic formalisation to simulate human intelligence.[62] However, the rise of machine learning and algorithmic modelling seems inherently inductive. Nevertheless, this general idea of fit aside, there are still methods and techniques that attempt to approximate rules.[63] In this sense, these rules, while not generated from an inductive source, might be inferred to fit under general laws akin to a D-S explanation.

c. Causal explanation

A leading theory of explanation is causal explanation. Lewis provides the seminal description of causal explanation: 'to explain an event is to provide some information about its causal history.'[64] In this way, the act of explaining consists in the provision of information on the causal history of an event (explanatory information) to someone else.[65] Lewis notes that causal explanation can be posited at different levels of generality. For instance, an explanation might be sought of a particular event or the causal history of a particular case. Alternatively, *general explanatory information* provides some 'general explanatory information' about events of that kind.[66] Moreover, sometimes a general explanation also explains a particular event or class of events - to take Lewis' example: 'explaining why struck matches light in general is not so very different from explaining why some particular struck match lit.'[67]

Causal explanation of machine learning might take various levels.[68] For instance, the explanation might provide casual information on how an algorithm trains a model: in this way, this information explains at a general level how models are constructed. Further, we might provide a global explanation of how the model functions overall. Finally, we might seek to provide causal information around how a particular output for a model was generated. Aside from explaining a machine learning model itself, we might use machine learning to enquire into the cause(s) of some other phenomena. In this case, we seek to use machine learning to explain, rather than explaining the model or algorithm itself.

i. Manipulability Conception

Woodward's manipulability conception of causal explanation emphasises this control and manipulation element of explanation. Notably, the account argues that the distinguishing features of causal explanations is that they provide 'information that is potentially relevant to manipulation and control: they tell us how, if we were able to change the value of one or more variables, we could change the value of other variables.'[69] In Woodward's words:

> '... an explanation ought to be such that it can be used to answer what I call a what-if-things-had-been-different question: the explanation must enable us to see what sort of difference it would have made for the *explanandum* if the factors cited in the *explanans* had been different in various possible ways.'[70]

While the manipulation or control derived from any explanation need not be actual - indeed, any hope of manipulation might be impossible - Woodward argues the process is helpful as a heuristic (a mental shortcut). In this way:

> 'the information that is relevant to causally explaining an outcome involves the identification of factors and relationships such that if (perhaps contrary to fact) manipulation of these factors were possible, this would be a way of manipulating or altering the phenomenon in question.'[71]

Certain explanations such as the provision of counterfactual explanation may inherently lend themselves to such interrogation. For instance, counterfactual explanations outline the 'closest possible world' where the smallest change leads to the desirable outcome.[72] Consequently, the provision of counterfactuals often satisfies the manipulability conception, at least in part.

Section 4 key messages:

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, the *explanandum* should be crafted to serve this purpose. In short, we should be clear exactly what we want to explain.

- **The proper account of explanation generally (explanation *simpliciter*) lies with general epistemology rather than the philosophy of science in particular.**
- **We should be clear what kind of explanation we seek - do we seek an everyday explanation, a scientific explanation, or an explanation of a more specific variety?**
- **Not all explanations of machine learning will be scientific but some may be.**
- **There are key contexts in which scientific explanation of machine learning, that is, explanation by reference to general laws, may be appropriate for healthcare or research.**
- **There are different accounts of scientific explanation, namely: deductive-nomological model, deductive-statistical explanation, inductive statistical explanation, and the statistical relevance model. Each may be relevant to the interpretability of machine learning.**
- **Many explanations sought of machine learning in the context of healthcare or research are likely causal, that is, they seek to demonstrate what caused a particular output, and to demonstrate the specific rules that lead to the generation of an output.**
- **In the machine learning context, data subjects often seek an explanation which provides information that is potentially relevant to manipulation and control, as proposed by Woodward's manipulability conception of causal explanation.**

5. Explanatory pragmatism

The question of what the concept of explanation includes is often held separate from the normative account of explanation. That is, the *what* of explanation is often separated from the *why* of explanation. This separation is thought to be a mistake by some theorists. Explanatory pragmatists are one such group of theorists. Loosely defined, *explanatory pragmatists* oppose traditional conceptual accounts on the basis that these accounts (erroneously) omit pragmatic and contextual elements.[73] Van Fraassen provides one of the best articulations of why objectivist approaches ought to be rejected, noting:

> 'The discussion of explanation went wrong at the very beginning when explanation was conceived of as a relation like description: a relation between a theory and a fact. Really, it is a three-term relation between theory, fact, and context. No wonder that no single relation between theory and fact ever managed to fit more than a few examples! Being an explanation is essentially relative for an explanation is an answer… it is evaluated vis-à-vis a question, which is a request for information. But exactly… what is requested differs from context to context.'[74]

In Van Fraassen's terms, context is a key feature of the concept of explanation. Explanation divorced from context is inherently ambiguous because it lacks the contrast class that context provides. For instance, consider Lipton's example of his three year old son:[75]

> 'When I asked my three year old son why he threw his food on the floor, he told me that he was full. This may explain why he threw it on the floor rather than eating it, but I wanted to know why he threw it rather than leaving it on his plate.'

In this respect, 'adding structure' to the why-question asked resolves some of the ambiguities that would otherwise exist when considering explanation more generally.[76] We consider this process of refining why-questions further in Section 8(a) below.

This report remains neutral on the success of the explanatory pragmatist argument. We premise none of our further work on the proposition that explanation is 'essentially pragmatic.' Our proposition is more modest: we are interested in explanation of machine learning in healthcare and research. While it may not be generally true that an account of explanation differs according to context, because our research question concerns explanation in a particular context, we judge the success or failure of explanation according to that context. Accordingly, while we remain neutral in regards to the success of explanatory pragmatism, we recognise the importance of context and understanding why an explanation is being sought. To quote Putnam, 'to regard the 'pragmatics' of explanation as no part of the concept is to abdicate the job of figuring out what makes the explanation good.'[77] It is to the goods of explanation that we now turn.

Section 5 key message:

- **Explanatory pragmatism emphasises context, arguing that explanation is a 'three-term relation between theory, fact, and context.' On this account, explanation is evaluated with respect to the particular question being asked. Explanatory pragmatism suggests that we should reconsider any quest to find 'the' explanation appropriate to all machine learning applications: the appropriate explanation will depend on the context and the machine learning application at stake.**

6. The good(s) of explanation

What does an explanation give us? What are the possible goods - intrinsic or instrumental that explanation provides?[78] This section considers the intrinsic good of understanding and introduces some ideas as to what instrumental value explanation might provide. To reiterate earlier thoughts, explanation is to be distinguished from knowledge, the key question for us being what goods explanation provides over and above knowledge. In a phrase: what is the value of *understanding why*?

a. Intrinsic goods of explanation

Lipton tells us that the 'name for the intrinsic good of explanation is 'understanding.''[79] This begs the question: what is understanding? By way of answer, Lipton offers five possible conceptions of understanding: reason, familiarity, unification, necessity, and causation. We outline their broad contours below:[80]

- The *reason conception of understanding* argues that understanding is 'identified with having a good reason to believe.'
- The *familiarity conception of understanding* notes that understanding is merely 'reduction to the familiar.'
- The *unification conception of understanding* tells us that 'we understand a phenomenon when we see how it fits together with other phenomena into a unified whole.'
- The *necessity conception of understanding* asserts that 'explanations somehow show that the phenomenon in question had to occur.'
- The *causal conception of understanding* conceptualises explanation as giving information about causes.

Interpretable machine learning methods discussed in the Interpretable Machine Learning report may potentially generate all or perhaps none of the kinds of understanding outlined above. Literature on machine learning typically emphasises the instrumental goods that explanations offer.[81] Explanation is thought to be important because it facilitates accountability, underpins data subject rights, and facilitates trust. Although this limitation is important, a focus on the instrumental value of explanation should not be myopic, depriving us of thoughts on the intrinsic value of understanding in healthcare and research.

In healthcare, understanding is an important value beyond allowing healthcare professionals or patients to take informed action. We are familiar with the idea that diagnosis as a form of explanation labels a certain set of phenomena. Ideally, this labelling allows a healthcare professional or patient to better predict the patient's prognosis and, hopefully, prescribe the most effective treatment. However, there are situations where diagnosis provides little certainty by way of prognosis and where there is no effective treatment. In these latter scenarios, there is likely still some residual reason to consider understanding. Although understanding may not help the healthcare professional or patient predict the future or take action, it may still be important that they understand the patient's health status.

In research, the intrinsic good of understanding is often apparent - the main undertaking of research often being understanding itself. Indeed, as our analysis in the Machine Learning Landscape report indicated, common research applications for machine learning include the structuring of unstructured data and non-hypothesis research. Understanding how machine learning models reach their conclusions may assist with understanding any relationship the model found in the dataset. In short, understanding the model facilitates understanding of the

data it structures. Moreover, understanding a model also assists with the instrumental good of being able to iteratively improve on the machine learning algorithm.

b. Instrumental goods of explanation

Understanding is powerful - when we understand something, this understanding often allows us to control or manipulate that something.[82] For instance, Lombrozo emphasises the instrumental value of explanation, noting that explanation plays an important role in the 'discovery and confirmation of everyday beliefs and by extension of intuitive and scientific theories.'[83] Moreover, often the literature on interpretable machine learning frames the success or failure of an explanation solely by whether the explanation assists in the performance of the task or decision.[84] What instrumental goods does explanation provide? The following analysis provides some ideas of the instrumental goods explanation provides.

Lipton offers us one example of an instrumental good of explanation: inference, specifically inference to the best explanation. That is, explanation can be a tool to acquire true beliefs.[85] Lipton notes that self-evidencing explanations often operate in this fashion - the phenomenon that is explained also provides an essential part of the reason for believing the explanation is correct.[86] Following Lipton, a possible instrumental good of explanation is the ability to infer beliefs from our explanation, thereby acquiring these beliefs and strengthening our explanation.

As noted in Section 4(c)(i), Woodward's manipulability conception of causal explanation emphasises the hypothetical control or manipulation that explanation characteristically brings. In this way, one of the major goods that explanation might provide in the context of machine learning is the (often only hypothetical) understanding of what to change to produce a different result. Of course, knowledge of what to change to produce a different result does not guarantee that there is a feasible change to be made. Regardless, the knowledge of what to change to produce a different result, even if one is powerless to make that change is still valuable indeed.

As noted in the Interpretable Machine Learning report, the success of explanation in the context of interpretable machine learning is often measured by whether the explanation leads to better performance of the task in question. For instance, if a machine learning model in the context of radiology is rendered interpretable and this interpretability is successful, the resulting human reader interpretations will have increased predictive accuracy. As outlined in the Interpretable Machine Learning report counterintuitively to the idea that there is an accuracy to interpretability trade off, interpretability may in some instances produce better predictive accuracy.[87] This suggests that explanation to produce better predictive accuracy is important, however it does not exhaust the good of explanation in the context of machine learning.

A Salient Feature | Interviews

Some interviewees indicated that their criteria for success of an explanation would be whether the explanations assisted the performance of the task.

To summarise the above, explanations may be instrumentally valuable for a number of reasons. Principally, explanation may assist with inference to the best explanation, provide the knowledge to manipulate the system, and lead to better predictive accuracy. Notably, explanations may fail to produce any of these goods. Indeed, many of these goods require a

successful act of explanation - a transfer of explanatory information to another. The next section outlines the literature on explanatory pragmatism and the lessons it provides.

Section 6 key messages:

- **What does explanation give us? Explanation can be intrinsically valuable and/or instrumentally valuable.**
- **Explanation offers us the intrinsic good of understanding. Understanding may give us 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 is often not the focus of explanation in the context of interpretable machine learning. However, understanding may be especially important for machine learning for health research.**
- **Explanation is also instrumentally valuable. For instance, understanding often provides information relevant to manipulate and control phenomena. In the context of machine learning, explanation may mean noting what features or weighting of features to change to obtain a different outcome.**

7. Explaining as a speech act

Explanation as a concept is often analysed separately from the *act of explaining*. As a subset of explanatory pragmatism, some think this separation untenable. That is, the central case of explanation is as an illocutionary act, namely a statement which is bound up with the performance of an act. We do not take a position on whether explanation is necessarily always an illocutionary act. Indeed, there are those who think that not all explanations are acts or count as explaining, for instance, Lewis: 'One who explains may provide not another, but rather himself, with explanatory information.'[88] Nevertheless, analysing the concept in this way illuminates important aspects of what someone does when they explain.

a. Illocutionary acts

Speech acts are 'acts that can (though need not) be performed by saying that one is doing so.'[89] In this vein, we can promise, assert, and resign merely by saying so. What about explaining? Where does explaining fit? What does explaining as a speech act tell us?

It is characteristic of speech acts that they have particular aims. The act of *assertion* typically aims to produce a belief in an addressee.[90] This aim is typical of assertion but we can also imagine circumstances where something is asserted without the aim of producing a belief, not even in the person making the assertion.[91] Austin puts the act of explaining in the roughly drawn class of 'expositives', ' acts of exposition involving the expounding of views, the conducting of arguments, and the clarifying of usages and of references.'[92] However, we might be able to say something more specific, that the characteristic of the aim of explanation is to render something *understandable*.[93] In this way, most acts of explaining aim to render something understandable to someone else. However, we can also imagine instances of explaining that do not have that aim. For instance, a cynical take on the 'right to explanation' is that the chief aim of acts of explaining is the discharge of a legal requirement rather than the production of any understanding in the audience.

We may further subdivide speech acts into those acts that have direct and indirect force. Consider the difference between urging and persuading someone to shut a door.[94] Given the correct circumstances, we can 'urge' merely by saying "I hereby urge you to shut the door." Consequently, urging is capable of direct force and counts as an *illocution*. However, there are no circumstances in which we can 'persuade' merely by declaring "I hereby persuade you." In this way, the act of persuading is a *perlocution*.

Explaining is typically classed as an illocutionary act. Like warning or promising, explaining is reckoned to be 'performed by uttering words in certain contexts with certain intentions.'[95] Explaining differs from related but perlocutionary acts such as enlightening someone - acts which are effects that might arise from one's explaining.[96]

b. Explaining as an illocutionary act

If explaining is an illocutionary act, its characteristic aim is to render something understandable to someone. In light of this description, Achinstein distinguishes between correct explanations and good explanations. *Correct explanations* are explanations where the 'propositional member of the ordered pair is true.'[97] For example, Newton's explanation for tides consists in the pair: The tides occur because of the gravitational pull of the moon and this explains why tides occur.[98] However, correct explanations may nonetheless still fail to be *good explanations* as they may be

wholly inappropriate in a multitude of ways, being pitched beyond the abilities of the audience or failing to meet the interests of the audience.[99] For Achinstein, a good explanation reaches beyond a correct explanation in being further characterised by a set of instructions to align the explanation to the interests, beliefs, and capacities of the audience.[100]

It is a consequence of Achinstein's theory that there is no universal set of instructions for all audiences and contexts.[101] Audiences' interests in requesting an explanation differ. Capacities differ. Motivations differ. There may be universally 'correct' explanations but no archetypically good explanation exists beyond an explanation that is sensitive to these nuances. Indeed, the interests of the audience should influence the formulation of the why-question asked - perhaps the audience seeks an everyday explanation, perhaps a scientific, or a causal explanation.[102] To echo Achinstein, the only way we can judge which kind of explanation is appropriate is by reference to the interests, and capacities of the audience in question. To explain well, the ultimate yardstick is the audience we have in mind.

Section 7 key messages:

- **Characteristically, the act of explaining aims at rendering something explainable to someone.**
- **The act of explaining is an illocutionary act, in that it can be true that one has explained something, even if the act did not in fact have the effect of rendering that something explainable. In this way, explaining differs from concepts like enlightening that are more bound up in the effects upon the addressee's thoughts or beliefs.**
- **We can distinguish between 'correct explanations' where the 'propositional member of the ordered pair is true' and 'good explanations' that take into account broader ideas of being appropriate to the addressee in question.**
- **There is likely no one archetypically 'good' explanation. Good explanations are sensitive to the interests and requirements of their audience, these interests influencing the formulation of the specific why-question asked.**

8. Getting to *good* explanatory acts

Explanatory pragmatists tell us that explanations are inherently linked with the context in which they arise - the purpose for which the explanation is sought being a large part of this context. We noted that explanation has an intrinsic good: understanding. Further, we also noted that explanation often has instrumental value - explanation and any understanding produced helps us make inferences, gives us a degree of control, and allows us to manipulate the world around us. In the context of machine learning, explanation of a machine learning model or specific decision may produce none or all of these goods. This aside, how might our acts of explanation succeed or fail? How might our acts of explanation be more or less satisfactory?

Lewis suggests a non-exhaustive list on how an act of explaining may be more or less satisfactory.[103] This list includes: interest insensitivity, too little explanatory information, stale explanatory information, inaccurate explanatory information, tangled explanatory information, failure to correct misapprehensions of the recipient. The following elaborates on select items of this list.

a. Interest insensitivity

Explanatory acts can succeed or fail in relation to whether they address the specific question to which the addressee seeks an answer. Often what separates good explanations from poor explanations is that good explanations meet the interests and concerns of the relevant audiences.[104]

No amount of explanatory information, subtlety, or nuance of delivery can rescue an explanation that is misdirected.[105] Accordingly, we should consider the variety of purposes and interests that underpin the request for explanation and direct the explanation accordingly. How might we go about this? The philosophy of explanation, particularly contrastive accounts of explanation, may be instructive.

i. Contrastive diagnosis

Explanations are often said to be *contrastive*. That is, all forms of explanation implicitly or explicitly take the form "why *P* not *Q*?"[106] According to this account, explanations never simply ask why something occurs or obtains but always ask why something occurs or obtains instead of something else.[107] Following this, when posing why-questions, there is a *fact* to be explained and a *foil*, the foil being one of the many different contrastive ways in which we could state the why-question.[108] For example, consider an anecdote regarding the bank robber Willy Sutton. When asked by a priest why he robbed banks, Sutton replied: "it's where the money was kept."[109] One of the reasons why Sutton's response is humorous is that Sutton intentionally or unintentionally gives an odd contrastive foil.[110] Sutton's construction contrasts the fact of his bank robbing with the foil of why he does not rob other institutions. This is odd because (presumably) the foil the priest meant was why he robs banks instead of doing something else entirely.

The choice of foil is important as different foils often result in very different why-questions. To adapt Lipton's example, a preference for action movies may explain why I saw *Gladiator* instead of *Chocolat* last night, but does not explain why I went out rather than staying

home.[111] In this way, the choice of foil may radically alter what explanatory information is appropriate.

We can remain neutral on the proposition that all explanations are necessarily contrastive. However, we ought to note the pragmatic significance of having a clear fact-foil in mind. Lewis notes that where we cannot give a full explanation (which is likely - see Section 8(b) below) we can use contrastive why-questions to indicate what explanatory information is being sought.[112] In this way, we can approximate a more specific idea of what explanatory information best suits the addressee's needs. In the context of machine learning, an idea of the fact-foil combination key audiences have in mind may assist with selecting the appropriate interpretable machine learning tool and tailoring any emerging explanation.

b. Partial explanation

Explanatory acts may fail if they provide inadequate amounts of explanatory information. To quote Lewis: 'The explanatory information provided may be correct, but there may not be very much of it.'[113] Consider the concept of an *ideal explanatory text*.

> 'An ideal explanatory text contains all of the facts and all of the laws that are relevant to the *explanandum*-fact. It details all of the causal connections among those facts and all of the hidden mechanisms.'[114]

In almost all cases, the ideal explanatory text to explain a phenomenon is huge and complex, making 'full explanations' a poor fit for most audiences.[115] Consequently, most (if not all) explanations will attempt to tailor their explanation, to pick out the appropriate 'pages' of the explanatory text for different audiences. There are three elements we should bear in mind when considering the partiality of explanation. First, partiality is not falsity. If an explanation does not account for a phenomenon's full causal history, that explanation is not false, it is merely one part of the truth. Second, we might also think that what counts as a full explanation depends on the specific explanatory information sought by the addressee.[116] In this regard, while almost all explanations fall short of providing a full causal history, many are sufficient to satiate the specific why-question asked. Third, the strategy of carefully considering the fact-foil pair likely also assists us in figuring out what page of the ideal text best fits the interests of the addressee (see Section 8(a)(i) above). Perhaps this idea is best summed up by the originator of the ideal explanatory text, Railton:

> 'Needless to say, even if we did possess the ability to fill out arbitrarily extensive bits of ideal explanatory texts, and in this sense thoroughly understood the phenomena in question, we would not always find it appropriate to provide even a moderate portion of the relevant ideal texts in response to particular why-questions. On the contrary, we would tailor the explanatory information provided in a given context to the needs of that context; if we had the capacity to supply arbitrarily large amounts of explanatory information, there would be no need to flaunt it.'[117]

In the context of machine learning for healthcare, different audiences may (explicitly or implicitly) seek out different parts of the causal chain and have different requirements in terms of accessibility of this explanation. Accordingly, assessment of what part of the ideal explanatory text the addressee needs or wants may assist when tailoring the explanation.

c. Stale information

The *familiarity conception of understanding* tells us that understanding proceeds by making the unfamiliar familiar.[118] Accordingly, explanatory power arises from comparing unfamiliar phenomena, for example the kinetic theory of gasses, with the familiar, in this case, the movement of billiard balls.[119] The consequence of this theory is that understanding presupposes that something is already understood, the function of explanation being to relate the object of inquiry to knowledge that is already possessed.[120] The success of this conception of understanding is unclear and challenged. However, what we do know is that explanatory information that merely repeats information to an addressee and does not add anything fails to provide the good of understanding. That is, if an explanation merely repeats what the addressee already knows, no understanding is gained.

The familiarity conception of understanding requires us to consider our audience - their capacities, their current state of knowledge, the purpose for which the explanation might serve. Indeed, many of the elements that O'Neill links to intelligent trust likely apply here.

d. Accurate explanation as probabilistic explanation

Explanation is often said to be probabilistic. *Probabilistic explanation* is 'the explanation of things that happen by chance: the outcomes of irreducibly probabilistic processes.'[121] In the context of machine learning, interpretations of some machine learning algorithms may be necessarily and irreducibly probabilistic.

Some of the techniques discussed in the Interpretable Machine Learning report are described as being 'approximations' or 'estimations.' This is often because they seek to compress highly complex, highly variable models into a digestible idea of what the model finds significant.[122] Indeed, a neural network may have thousands of parameters, each with their own weighting, each weighting being dependent (or not) on other parameters. The explanation produced may be correct for a particular instance of processing, a local group, or be an accurate approximation of what the model generally finds significant. In many cases, the resulting explanation will be an approximation. In some instances, this kind of explanation may be the best we can hope for if we require a human-interpretable explanation. In a sense, these explanations are probabilistic explanations of an underlying model that is irreducibly probabilistic.

How does probabilistic explanation fit with philosophical theories of explanation? As covered in Section 4(b), probability is not a foreign concept to explanation. As noted in regards to D-S, I-S, and S-R explanation, there are multiple ways for probability to feature in accounts of explanation. Instead of framing interpretable machine learning processes as true or false then, we might instead think of these tools as containing, for example, statistically relevant information or not.

Section 8 key messages:

- **Acts of explaining can be more or less satisfactory. Broadly, the following non-exhaustive list from Lewis (1987) provides a good guide to how acts of explaining can succeed or fail:**
  - **Interest insensitivity - explanation as an act may be evaluated by respect to whether it answers the specific questions the audience had in mind.**
  - **One method to assist in tailoring explanations where a contrastive explanation is sought (Why *P* not *Q*?), is to ensure that the foil (the *Q*)**

chosen should assist in approximating the specific explanatory information the addressee requires.

- **Another way explanations may fail is if they provide too little information. Full explanations that encapsulate a phenomenon's entire casual history are fiction. Rather, we should consider what part of the 'page' of the ideal explanatory text best satiates the addressee in question.**
- **Explanations often work by making the unfamiliar familiar, using concepts and phenomena the audience is already familiar with to explain concepts or phenomena the audience is unfamiliar with. However, explanations should not be stale - they should add rather than just repeat information.**
- **Some explanations of machine learning may be irreducibly probabilistic. It is likely that these explanations are best analysed by considering, for example, whether the information they provide is statistically relevant to the output or not.**

9. Broad interests at stake

The literature on explanation tells us that, where there is a precise purpose in mind, we should precisely define what is to be explained. Further, we should be clear about the class of explanation we seek, for instance, do we seek an everyday, scientific, or causal explanation? Finally, explanatory pragmatism tells us to be sensitive to the interests and capacities of the addressees of the explanation: what kind of fact-foil pair do they have in mind? What does the addressee already know? What part of the causal process does the addressee want information on? This section seeks to elaborate on these details for machine learning in the context of healthcare and research, outlining key audiences and key interests in explaining or using machine learning to explain.

In general, it is important to distinguish between using a machine learning model to explain some phenomena and explaining the machine learning model itself, that is, machine learning models as the *explanans* and the *explanandum*.

In regards to the class of explanation we require and the interests and capacities of addressee, we note that there are multiple key audiences for explanation in the context of machine learning for healthcare or research. That is, each audience and perhaps even specific members within each audience often seek different classes of explanation and have different purposes in mind. We outline these key audiences, their primary purpose, and give an example of a relevant explanation in the table below.

Table 1: key audiences, primary purposes for explanation of machine learning models in healthcare and research

Audience	**Primary purpose**	**Example**
Developers themselves	To debug, understand the behaviour of, and iterate on their model. To verify, validate, and properly label the system	Semantic maps informing in the context of predictions based on automatic analysis of radiological images might assist in picking out confounding factors[123]
Regulatory bodies	To evidence the safety and effectiveness of a medical device	Intrinsically interpretable machine learning models or methods that transform models into decision rules (for example, RuleFit) may make it easier to link a model's reliance on features with supporting scientific literature[124]
Commissioning bodies	To evidence the system as a cost effective tool for use in a health system. For instance, the National	Global interpretability in the context of discharge management tools might demonstrate both

	Institute for Health and Care Excellence (NICE) Evidence standards framework for digital health standards[125]	conformance to policy but also demonstrate the return on investment for such a system. For example, it is known why the tools recommend discharge, we know the optimal patient pathway to follow
Healthcare professionals	To contextualise and interpret the output to make a clinically relevant action To contextualise and interpret the output for their patient To evidence the safety and effectiveness of the device	Global interpretability to understand what features the model generally finds significant to link to the healthcare professional's clinical judgment Local interpretability to understand what the model found significant for a particular patient
Health consumer / User	To contextualise and interpret for themselves the outputs of the model To contextualise and interpret for themselves to take an action related to their health or care To consider the system reliable or safe for their own use	Global interpretability to understand what features the model generally finds significant and link to the user's own understanding Local interpretability to understand what the model found significant for that particular user
Public	To assist in the public justification for the deployment of the system	Global interpretability to understand what features the model generally finds significant to consider the acceptability of reliance on these features Local interpretability to allow a human in the loop to be an effective checker
Scientific or academic	To assist in scientific discovery or assist in establishing causation To ensure conclusions of studies including machine learning are reproducible and benchmarked	For example, partial dependence plots show the marginal effect a feature has on the predictive outcome.[126] Accordingly, they may be useful for contextualising and interpreting a model

As the above table demonstrates, there are many key audiences for explanation in the context of machine learning for healthcare and research. It is unlikely that any given explanation will satisfy each audience and fulfil the purpose each audience has in mind. For example, the kind of interpretability (if any) required to test the safety and effectiveness of a medical device may be very different to the kind of interpretability sought by a patient in a clinical care setting.[127] It may be difficult to satisfy both simultaneously.

The above key audiences may be further distilled and combined with the literature outlined in the Interpretable Machine Learning report, giving us some distinct but overlapping purposes behind interpretability:

A. Interpretability to evidence the safety and effectiveness of a system

Interpretability here may assist with hunting for confounding features, debugging, and validation and verification of the system.

B. Interpretability to facilitate human-computer interaction

Interpretability here ensures the model is usable and conveys sufficient information to users to facilitate successful interaction between the system and the human in the loop.

C. Interpretability to assist in scientific or causal understanding

Interpretability here allows the model to illuminate the links between specific phenomena to general laws of nature or conveys casual understanding of the phenomena the model describes.

D. Interpretability as foundational to providing data subjects' control and a means to secure controllers' accountability

Interpretability here underpins data subjects, users, or consumers' ability to control the use of data or challenge decisions made concerning them using machine learning models. Moreover, interpretability here also facilitates accountability of the controller or manufacturer.

The importance of interpretability in relation to each purpose is heavily dependent upon context and the attributes of the machine learning model in question. As a consequence, the Interpretability by Design Framework provides a way to think through the centrality of interpretability in machine learning for healthcare and research.

Section 8 key messages:

- **The literature on explanation indicates that, where there is a precise purpose behind a request for explanation, we should precisely define what is to be explained. We should be clear about the class of explanation we seek, for instance, do we seek an everyday, scientific, or causal explanation? Finally, explanatory pragmatism tells us to be sensitive to the interests and capacities of the addressees of the explanation: what kind of fact-foil pair do they have in mind? What does the addressee already know? What part of the causal process does the addressee want information on?**
- **The above table noted the following key audiences for interpretability in healthcare and research: developers themselves, regulatory bodies, commissioning bodies, healthcare professionals, health consumers/patients/users, the public, and scientific or academic audiences.**
- **These audiences may be distilled into four key purposes for interpretability:**
    - **Interpretability to evidence the safety and effectiveness of a system**

- **Interpretability to facilitate human-computer interaction**
- **Interpretability to assist in scientific or causal understanding**
- **Interpretability as foundational to providing data subjects control and a means to secure controllers' accountability**

10. Transparency and explanation

What does transparency and explanation mean in the context of machine learning? The philosophical literature provides some answers.

The pursuit of transparency can be pernicious where emphasis is put on disclosure without regard for the audience and their interests. Relatedly, the pursuit of trust can also be pernicious where those whose trust is sought are not given sufficient information to trust intelligently.

Explanation has two elements, the *explanandum* that which is to be explained and the *explanans* that which does the explaining. It is good to understand what kind of explanation is sought: scientific, causal, an everyday explanation? The kind of explanation sought will influence the *explanans* given.

Explanatory pragmatism tells us that the recipe of explanation is more than just an *explanandum* and an *explanans*, that context is an indispensable ingredient of what it is to explain. Accordingly, context narrows the why-question asked, resolving intractable ambiguities that would otherwise exist.

Why explain? What goods does explanation grant us? The intrinsic good of explanation is understanding. Understanding may be conceptualised in several ways. For instance, understanding may render otherwise unfamiliar phenomena familiar or link a phenomenon to our broader understanding, unifying our thoughts. Understanding is often not the focus of the interpretable machine learning literature. Despite this, understanding may be a good especially valuable in machine learning for health research, as understanding is often the primary good sought from research. Explanation provides instrumental goods to those who receive explanation. For instance, understanding often provides us with information relevant to manipulate or control the phenomenon in question.

Explanation is often an act that aims to render something explainable to someone else. Indeed, the standard case of explanation in the context of machine learning is an explanation of the machine learning system by the system itself or the developer to a user or data subject. As an act, explanation is an illocutionary act. That is, it can be true that one has explained something, even if the act did not have the effect of rendering something explainable. In this way, explanation characteristically aims at rendering something understandable to another, although the act may fail to achieve this purpose. We can distinguish between 'correct explanations' where the information given as the *explanans* is true and matches the *explanandum* and 'good explanations' that take into account broader sensitivities of the explanation being appropriate for the audience in question. Accordingly, on the explanation as a speech act account, there is no one archetypically good explanation – a good explanation meets the interests of those to whom it is addressed.

There may be no one archetypically good explanation, but are there some general characteristics of how acts of explanation may succeed or fail. Explanations should be sensitive to the interests of the audience, answering the specific why-question each has in mind. One

method to assist in tailoring explanations where a contrastive explanation is sought (why *P* not *Q*?), is to ensure that the foil (the *Q*) chosen approximates the specific explanatory information the addressee requires. Explanations should provide sufficient explanatory information for their given purpose. While 'full explanations' are a fiction, the addressee should receive the 'page' of the larger type of full explanation they need. Explanations often work by making the unfamiliar familiar. However, if they merely repeat information already known they fail to provide understanding, failing as much an explanation that is highfalutin, where no connection to familiarity is found. Some explanations, especially in the context of machine learning, are irreducibly probabilistic, involving the compression of complex processes. Instead of thinking of these explanations containing true or false premises, it may be better to consider the information they provide as being statistically relevant or not.

Transparency and explanation have a rich philosophical literature underpinning each concept. This literature is instructive, providing insight into how explanations of machine learning system in healthcare and research might succeed or fail. More often than not, the best question is ask when considering explanation in this sector will be: what work is my explanation doing? What exactly do I need to explain? Why do my addressees want an explanation?

References

[1] For example: Craglia M, Annoni A, Benczur P, et al. *Artificial Intelligence: A European Perspective*. Joint Research Centre of the European Commission. 2019.

[2] Baker McKenzie. *Outside the Comfort Zone: Building consumer trust in healthcare*. Baker McKenzie. 2019: 5.

[3] Institute of Electrical and Electronics Engineers. *Ethically Aligned Design: A Vision for Prioritizing Human Well-being with Autonomous and Intelligent Systems*. Version 2. 2017: 158-161.

High-Level Expert Group on Artificial Intelligence. *Ethics Guidelines for Trustworthy AI*. 2019.

[4] High-Level Expert Group on Artificial Intelligence. *Ethics Guidelines for Trustworthy AI*. 2019.

[5] Future Advocacy, NHSX. *Code of Conduct for data-driven health and care technology: Guidance to encourage best practice outlined in Principle 7*. 2019 Preprint: 1-27.

[6] O'Neill O. *Lecture 4: Trust and Transparency* [Lecture] Reith Lectures. 2002.

[7] Jobin A, Ienca M, Vayena E. Artificial Intelligence: the global landscape of ethics guidelines. *Nature Machine Intelligence*. 2019; 1(9): 7.

[8] High-Level Expert Group on Artificial Intelligence. *Ethics Guidelines for Trustworthy AI*. 2019: 18.

[9] Ibid.

[10] Ibid.

[11] Ibid.

[12] Jobin A, Ienca M, Vayena E. Artificial Intelligence: the global landscape of ethics guidelines. *Nature Machine Intelligence*. 2019; 1(9): 7.

[13] Craglia M, Annoni A, Benczur P, et al. *Artificial Intelligence: A European Perspective*. Joint Research Centre of the European Commission. 2019: 59.

[14] Institute of Electrical and Electronics Engineers. *Ethically Aligned Design: A Vision for Prioritizing Human Well-being with Autonomous and Intelligent Systems*. Version 2. 2017: 158-161.

[15] Heald D. Transparency as an Instrumental Value. In: Hood C, Heald D. (eds.) *Transparency: The Key to Better Governance*. London: British Academy Scholarship; 2006: 62-69.

[16] Larsson T. How Open Can Government Be? The Swedish Experience. In: Deckmyn V, Thomson I. (eds.) *Openness and Transparency in the European Union*. Maastricht: European Institute of Public Administration, 9–51.

Heald D. Varieties of Transparency. In: In: Hood C, Heald D. (eds.) *Transparency: The Key to Better Governance*. London: British Academy Scholarship; 2006: 26.

[17] Heald D. Transparency as an Instrumental Value. In: Hood C, Heald D. (eds.) *Transparency: The Key to Better Governance*. London: British Academy Scholarship; 2006: 62.

[18] O'Neill O. *Lecture 4: Trust and Transparency* [Lecture] Reith Lectures. 2002.

[19] Ibid.

[20] Spiegelhalter D. Should We Trust Algorithms? *Harvard Data Science Review*. 2020; 2(1).

[21] O'Neill O. Transparency and the Ethics of Communication. In: Hood C, Heald D. (eds.) *Transparency: The Key to Better Governance*. London: British Academy Scholarship; 2006: 85-86.

[22] Ibid

[23] Ibid, 89.

[24] O'Neill O, Ethics for Communication? *European Journal of Philosophy*. 2009; 17(2): 173.

[25] O'Neill O. Transparency and the Ethics of Communication. In: Hood C, Heald D. (eds.) *Transparency: The Key to Better Governance*. London: British Academy Scholarship; 2006: 85.

[26] O'Neill O, Ethics for Communication? *European Journal of Philosophy*. 2009; 17(2): 172-173.

[27] O'Neill O. Transparency and the Ethics of Communication. In: Hood C, Heald D. (eds.) *Transparency: The Key to Better Governance*. London: British Academy Scholarship; 2006: 84.

[28] Ruben D. *Explaining Explanation*. London: Routledge; 1992: 14.

[29] Miller T. Explanation in artificial intelligence: Insight from the social sciences. *Artificial Intelligence*. 2019; 267: 1-38

[30] Salmon WC. Scientific explanation. In: Salmon MH, Earman J, Glymour C, et al (eds.) *Introduction to the Philosophy of Science*. Indianapolis: Hackett; 1992: 10.

[31] Ruben DH. *Explaining Explanation*. London: Routledge; 1992: 23.

[32] Salmon WC. Scientific explanation. In: Salmon MH, Earman J, Glymour C, et al (eds.) *Introduction to the Philosophy of Science*. Indianapolis: Hackett; 1992: 10.

[33] Ibid.

[34] Ruben DH. *Explaining Explanation*. London: Routledge; 1992: 23.

[35] Salmon WC. Scientific explanation. In: Salmon MH, Earman J, Glymour C, et al (eds.) *Introduction to the Philosophy of Science*. Indianapolis: Hackett; 1992: 10.

[36] Ruben DH. *Explaining Explanation*. London: Routledge; 1992: 13-14.

[37] Salmon WC. Scientific explanation. In: Salmon MH, Earman J, Glymour C, et al (eds.) *Introduction to the Philosophy of Science*. Indianapolis: Hackett; 1992: 8.

[38] Ruben DH. *Explaining Explanation*. London: Routledge; 1992: 13-14.

[39] Salmon WC. Scientific explanation. In: Salmon MH, Earman J, Glymour C, et al (eds.) *Introduction to the Philosophy of Science*. Indianapolis: Hackett; 1992: 8.

[40] Ruben DH. *Explaining Explanation*. London: Routledge; 1992: 13.

[41] Miller T. Explanation in artificial intelligence: Insight from the social sciences. *Artificial Intelligence*. 2019; 267: 1-38

[42] Ibid, 3.

[43] Mittelstadt B, Russell C, Wachter S. Explaining Explanations in AI. *Conference on Fairness, Accountability, and Transparency*. 2019: 2.

[44] Lipton P. What Good is An Explanation? In: Hon G, Rakover SS (eds.) *Explanation: Theoretical Approaches and Applications*. Netherlands: Kluwer Academic; 2001: 2.

[45] Salmon WC. Scientific explanation. In: Salmon MH, Earman J, Glymour C, et al (eds.) *Introduction to the Philosophy of Science*. Indianapolis: Hackett; 1992: 7.

[46] Lipton P. What Good is An Explanation? In: Hon G, Rakover SS (eds.) *Explanation: Theoretical Approaches and Applications*. Netherlands: Kluwer Academic; 2001: 2.

[47] Ibid, 2-3.

[48] Ibid.

[49] Ibid, 3.

[50] Ibid.

[51] Pearl J. Theoretical Implications to Machine Learning With Seven Sparks from the Causal Revolution. *arXiv*. 2018.

[52] Hempel C. *Aspects of Scientific Explanation and other Essays in the Philosophy of Science*. New York: The Free Press; 1965: 248.

[53] Woodward J. Scientific Explanation. *Stanford Encyclopedia of Philosophy*. 2014; 3-4.

[54] Deduction differs from induction: deduction is the process of reasoning from one or more statements (premises) to reach a logically certain conclusion.

[55] Ibid.

[56] ibid, 5.

[57] Railton P. A Deductive-Nomological Model of Probabilistic Explanation. *Philosophy of Science*. 1978; 45(2): 209-210.

[58] Woodward J. Scientific Explanation. *Stanford Encyclopedia of Philosophy*. 2014; 5.

[59] Ibid.

[60] Woodward J. Scientific Explanation. *Stanford Encyclopedia of Philosophy*. 2014; 10.

[61] Molnar C. Interpretable Machine Learning: A Guide for Making Black Box Models Interpretable. Learnpub; 2019. Available from: https://christophm.github.io/interpretable-ml-book/pdp.html [Accessed 13th February 2020].

[62] Lipton Z. *From AI to ML to AI: On Swirling Nomenclature & Slurried Thought*. Available from: http://approximatelycorrect.com/2018/06/05/ai-ml-ai-swirling-nomenclature-slurried-thought/ [Accessed 16 February 2020].

[63] Friedman JH, Popescu BE. Predictive learning via rule ensembles. The Annals of Applied Statistics. 2008; 2(3): 916-54.

[64] Lewis D. *Philosophical Papers Volume II*. Oxford: Oxford University Press; 1987: 217-218.

[65] Ibid.

[66] Ibid, 224-225.

[67] Ibid.

[68] Guidotti R, Monreale A, Ruggieri S, et al. A Survey of Methods for Explaining Black Box Models. ACM Computing Surveys. 2018; 51(5): 1-42.

[69] Woodward J. *Making Things Happen: A Theory of Causal Explanation*. Oxford: Oxford University Press; 2004: 6.

[70] Ibid, 11.

[71] Ibid.

[72] Wachter S, Mittelstadt B, Russell C. Counterfactual Explanations Without Opening the Black Box: Automated Decisions and the GDPR. *arXiv*. 2017: 6.

[73] Woodward J. Scientific Explanation. *Stanford Encyclopedia of Philosophy*. 2014; 25.

[74] van Fraassen BC. *The Scientific Image*. Oxford: Oxford University Press; 1980: 156.

[75] Lipton P. Contrastive Explanation. In: Knowles D (eds.) *Explanation and its Limits*. Cambridge: Cambridge University Press; 1991: 249-250.

[76] Ibid.

[77] Putnam, H. *Meaning and the Moral Sciences*. London: Routledge. 2010: 42.

[78] Lipton P. What Good is An Explanation? In: Hon G, Rakover SS (eds.) *Explanation: Theoretical Approaches and Applications*. Netherlands: Kluwer Academic; 2001: 1.

[79] Ibid.

[80] Ibid, 4-8.

[81] Kim B, Khanna R, Koyejo OO. Examples are not enough, learn to criticize!. *Advances in Neural Information Processing*. 2016: 2280-2288.

[82] Woodward J. Scientific Explanation. *Stanford Encyclopedia of Philosophy*. 2014; 25.

[83] Lombrozo T. The Instrumental Value of Explanations. *Philosophy Compass*. 2011; 6(8): 541-546.

[84] Kim B, Khanna R, Koyejo O. Examples are not Enough, Learn to Criticize! Criticism for Interpretability. *30th Conference on Neural Information Processing Systems*. 2016: 8.

[85] Lipton P. What Good is An Explanation? In: Hon G, Rakover SS (eds.) *Explanation: Theoretical Approaches and Applications*. Netherlands: Kluwer Academic; 2001: 18.

[86] Ibid, 19.

[87] Rudin C. Please stop explaining black box models for high stakes decisions. *arXiv*. 2018

[88] Lewis D. *Philosophical Papers Volume II*. Oxford: Oxford University Press; 1987: 219.

[89] Green M. Speech Acts. *Stanford Encyclopedia of Philosophy*. 2014: 2.

[90] Ibid, 9.

[91] Ibid, 9.

[92] Austin JL. *How to do things with words: The William James Lectures delivery at Harvard University in 1955*. Oxford: Oxford University Press; 1975: 161.

[93] Achinstein P. What is an explanation? *American Philosophical Quarterly*. 1977; 14(1): 1-2.

[94] Green M. Speech Acts. *Stanford Encyclopedia of Philosophy*. 2014: 9-10.

[95] Achinstein P. What is an explanation? *American Philosophical Quarterly*. 1977; 14(1): 1.

[96] Ibid.

[97] Achinstein, P. *Evidence, Explanation, and Realism: Essays in Philosophy of Science*. Oxford: Oxford University Press; 2010: xi.

[98] Woodward J. Scientific Explanation. *Stanford Encyclopedia of Philosophy*. 2014; 29-30.

[99] Ibid, 30.

[100] Ibid.

[101] Achinstein, P. *Evidence, Explanation, and Realism: Essays in Philosophy of Science*. Oxford: Oxford University Press; 2010: 137.

[102] Woodward J. Scientific Explanation. *Stanford Encyclopedia of Philosophy*. 2014; 30.

[103] Lewis D. *Philosophical Papers Volume II*. Oxford: Oxford University Press; 1987: 226-227.

[104] Ruben DH. *Action and its Explanation*. Oxford: Oxford University Press; 2003: 186.

[105] Lewis D. *Philosophical Papers Volume II*. Oxford: Oxford University Press; 1987: 227.

[106] Lipton P. Contrastive Explanation. In: Knowles D (eds.) *Explanation and its Limits*. Cambridge: Cambridge University Press; 1991: 256.

[107] Lipton P. A real contrast. *Analysis*. 1987; 47(4): 207.

[108] Lipton P. Causation and Explanation. In: Beebee H, Hitchcock C, Menzies P (eds.) *The Oxford Handbook of Causation*. Oxford: Oxford University Press; 2010: 624.

[109] Ruben DH. Explaining Contrastive Facts. *Analysis*. 1987; 47(1): 35.

[110] Ruben DH. *Explaining Explanation*. London: Routledge; 1992: 32.

[111] Lipton P. Contrastive Explanation. In: Knowles D (eds.) *Explanation and its Limits*. Cambridge: Cambridge University Press; 1991: 251.

[112] Lewis D. *Philosophical Papers Volume II*. Oxford: Oxford University Press; 1987: 229-230.

[113] Ibid, 226.

[114] Salmon WC. Scientific explanation. In: Salmon MH, Earman J, Glymour C, et al (eds.) *Introduction to the Philosophy of Science*. Indianapolis: Hackett; 1992: 36-37.

[115] Railton P. Probability, explanation, and information. *Synthese*. 1981; 49(2): 243-244.

Ruben DH. *Action and its Explanation*. Oxford: Oxford University Press; 2003: 186.

[116] Putnam, H. *Meaning and the Moral Sciences*. London: Routledge. 2010: 41-43.

Scriven M. Explanations, Predictions, and Laws. *Minnesota Studies in the Philosophy of Science*. 1972: 202.

[117] Railton P. Probability, explanation, and information. *Synthese*. 1981; 49(2): 244.

[118] Bridgman PW. *The Logic of Modern Physics*. New York: 37.

[119] Friedman M. Explanation and Scientific Understanding. *The Journal of Philosophy*. 1974; 71(1): 9-11.

[120] Scriven M. Explanations, Predictions, and Laws. *Minnesota Studies in the Philosophy of Science*. 1972: 202.

[121] Railton P. Probability, explanation, and information. *Synthese*. 1981; 49(2): 233.

[122] Molnar C. Interpretable Machine Learning: A Guide for Making Black Box Models Interpretable. Learnpub; 2019. Available from: https://christophm.github.io/interpretable-ml-book/scope-of-interpretability.html [Accessed 13th February 2020].

[123] Zech JR, Badgeley MA, Liu M, et al. Confounding variables can degrade generalization performance of radiological deep learning models. *PLoS Medicine*. 2015: 1-15.

[124] Friedman JH, Popescu BE. Predictive learning via rule ensembles. *The Annals of Applied Statistics*. 2008; 2(3): 916-54.

Regulation (EU) 2017/745 of the European Parliament and of the Council on medical devices [2017] OJ L117/1, art 61.

[125] National Institute for Health and Care Excellence (NICE), Evidence standards framework for digital health technologies. 2019. Available from: https://www.nice.org.uk/about/what-we-do/our-programmes/evidence-standards-framework-for-digital-health-technologies [Accessed 13th February 2020].

[126] Molnar C. Interpretable Machine Learning: A Guide for Making Black Box Models Interpretable. Learnpub; 2019. Available from: https://christophm.github.io/interpretable-ml-book/pdp.html [Accessed 13th February 2020].

[127] Ordish J, Murfet H, Hall A. *Algorithms as medical devices*. PHG Foundation. 2019: 23-29.

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