UPPSALA
UNIVERSITET

Explaining the output of a black box model and a white box model: an illustrative comparison

by

JOEL VIKLUND

15 June 2020
Thesis Project STS - Philosophy, 15.0 c
joel.wiklund2@gmail.com
Supervised by Sebastian Lutz
Department of Philosophy

1 Abstract

The thesis investigates how one should determine the appropriate transparency of an information processing system from a receiver perspective. Research in the past has suggested that the model should be maximally transparent for what is labeled as "high stake decisions". Instead of motivating the choice of a model's transparency on the non-rigorous criterion that the model contributes to a high stake decision, this thesis explores an alternative method. The suggested method involves that one should let the transparency depend on how well an explanation of the model's output satisfies the purpose of an explanation. As a result, we do not have to bother if it is a high stake decision, we should instead make sure the model is sufficiently transparent to provide an explanation that satisfies the expressed purpose of an explanation.

Contents

2 Introduction

In this essay, I will investigate how one can explain, respectively, a black box model and a white box model in an information processing system using the inductive-statistical explanation (IS explanation) and the deductive-nomological explanation (DN explanation). To begin with some terminology, there exist two types of black box models: The first type is a model whose prediction-generating-mechanisms in the model cannot be fully understood or accessed by humans. Due to the opacity of the black box model, in order to describe the prediction-generating-mechanisms, one has to create a second (post hoc) model which aims to approximate the actual prediction-generating-mechanisms of the model. The second type of a black box model is a model which is inaccessible to humans for property rights. Such a type of a black box model may have non-AI prediction-generating-mechanisms but remains opaque due to business confidentiality reasons. In contrary to a black box model, a white box model consists of prediction-generating-mechanisms which can be described in a way which is faithful to what the model actually predicts; no post hoc model is therefore needed. Black box models and white box models are both used in artificial intelligence (AI) and more specifically in machine learning applications.

In order to include non-AI black box models, i.e. in the case of the second type of a black box model, I will in this essay use the term information processing system (IPS). An IPS consist of three parts: data, model and prediction. Data are given as parameters to a model, black box or white box, which in turn generates a prediction based on the given data. However, for IPS, the complexity of the model's prediction-generating-mechanisms is not specified; they can be anything in between "a+b = c" to advanced machine learning, e.g. neural networks. Furthermore, a black box IPS is an IPS which uses a black box model as its choice of model and a white box IPS is an IPS which uses a white box model as its choice of model.

In real life, there exist problems that solely can be solved with black box IPS. To name one, constructing an artificial human brain will only be possible using a black box model because, as far as we know, the human brain is a highly complex neural network. However, problems that solely can be solved with black box IPS are not in the scope of this essay because we will only consider IPS where the creator can choose to implement a black box model or a white box model without losing performance. A topical example of when the accuracy does not depend on whether it is a black box IPS or a white box IPS is when predicting the recidivism risk for criminals in the US justice system. One model for predicting the recidivism risk, COMPAS (Correctional Offender Management Profiling for Alternative Sanctions), is a black box model and is used widely in the justice system for parole and bail decisions. The other model, COREL (Certifiably Optimal Rule Lists), is a white box model created by Rudin et al. (2020b). The COREL model has never been used in the justice system, but is a demonstrational example of the often missed opportunity to challenge the performance, given the same training data [1], of a black box model; the two models' accuracy of predicting the recidivism risk are very similar (Rudin 2019, p. 209).

In similar situations as in the case of predicting the recidivism risk in the US justice system, Rudin (2019) suggests that the creator of the IPS should always, if possible, implement a white box model for "high stake" decisions. It is not always clear what should fall under the category of high stake decision, but illustrative examples are ML-based pollution models stating that highly polluted air was safe to breathe and generally poor use of limited valuable resources in criminal justice, medicine, energy reliability, finance and in other domains. However, although these examples might be easy to label as high stake, it may be that a high stake decision for one person could be seen as a low stake decision for

[1] The training data is an initial set of data used to help a model optimize its prediction-generating-mechanisms in order to produce the most accurate results.

another and it is therefore hard to agree on why a decision should be considered high stake in general. Instead of motivating the choice of a white box model because the IPS generates a prediction that is used for high stake decisions, I will in this essay explore an alternative course of action to determine when one model is more appropriate than the other.

3 Background

When exploring the explainability of a model, a natural starting point is to give a brief overview of current work within the field (section 3.1). There is a large gap between what is called explainable AI (xAI) and what is normally considered an explanation within philosophy (Mittelstadt et al. 2018, p. 1). Due to this ambiguity, it makes sense to first give a background on what xAI is and how it relates to the topic of this essay, i.e. how one can explain an IPS's output using the IS explanation or the DN explanation (section 3.2). Secondly, it is important to motivate why an explanation for the output (prediction or classification) of an IPS is relevant in the first place. Due to the fact that this essay aims to compare a scientific explanation for a black box model and a white box model, I will in section 3.3 describe possible implications that might come with choosing a black box model instead of a white box model. In section 3.4 I present Floridi's receiver-contextualised explanation and discuss why such an explanation is important IPS based decisions.

3.1 xAI

Mittelstadt et al. (2018) summarize that xAI aims to "produce simplified approximations of complex decision-making [models]" (Mittelstadt et al. 2018, p. 1). From a philosophical perspective, it appears mistaken that "[producing] simplified approximations" should be considered an explanation; this is also what Mittelstadt et al. (2018) further conclude. The machine learning community has "re-purposed" the meaning of explanation and xAI should not be associated with scientific explanations in philosophy (Mittelstadt et al. 2018, p. 1). In philosophy, a scientific explanation is an "account that trace the causes of the events (states, conditions) explained" (Kitcher 1998, p. 1). But, within the machine learning community, xAI refers to a scientific model of the model without tracing the causes of a certain outcome of the model (Mittelstadt et al. 2018, p. 1).

3.2 Relevance of an explanation for an IPS's output

Before attempting to explain a black box model and a white box model using the IS explanation or the DN explanation, it is justified to ask "why, in the context of IPS, is an explanation of the model's output relevant?". Depending if you ask a machine learning scientist or a philosopher, you will probably get a different answer. Lipton (2016) provides a useful summary of different conceptions of why xAI is important (Lipton 2016, pp. 3-4). Although Lipton (2016) considers xAI for machine learning models, I believe that some conceptions of why an explanation of a model is relevant are very transferable to a non machine learning context, e.g. explaining the output of an IPS. When summarizing existing literature in the field, there are five main reasons why xAI is important according to Lipton (2016, pp. 3-4): trust, causality [2], transferability [3], informativeness and lastly fair and ethical decision-making. Regarding trust, Lipton (2016, p. 3) concludes that a model's explanation is prerequisite for generating trust with the stakeholders. Stakeholders, i.e. persons affected by the models prediction or classification, will be more at ease with the model's outcome if they have an understanding of how the model came to its prediction or classification. However, and his will be discussed further in this essay, the stakeholders' need of trust might differ depending on how the stakeholders are affected by the model's prediction or classification.

When it comes to the importance of 'informativeness', it is often the case that the IPS's generated output "is used to provide information to human decision makers" (Lipton 2016, p. 4). In such cases, an explanation of the model could, depending on what decision is to be

[2] Term is too machine learning specific in the context of IPS, and will therefore be left out in this essay.

[3] Same as for footnote 2.

made, add valuable information to the decision-making process. Additionally, Lipton (2016, p. 4) also points out that depending on what decision is to be made, different levels of transparency could be desirable for the model: it may be the case that “an [explanation] may prove informative even without shedding light on a model’s inner workings” (Lipton 2016, p. 4), but it could also be the case that a model has to be more transparent to prove informative to the human decision-making-process.

Lastly, there is the motive of contributing to “fair and ethical decision-making”. The main argument is that without an explanation for the model’s prediction or classification, one can never confirm that the models’ prediction-generating-mechanisms conformed to ethical standards. However, fair and ethical decision-making is more important for specific branches of IPS than others. Rudin (2019, p. 206) labels these kinds of IPS’s predictions ”high stakes predictions”. As mentioned in the introduction, a typical example of high stakes predictions occurs in the US court system where “recidivism predictions are used to determine who to release and who to detain” (Rudin 2019, p. 208; Lipton 2016, p. 4). In the case of recidivism predictions being used in the US court system, the need of fair and ethical decision-making is palpable. To name one good reason how an explanation could contribute to fair and ethical decision-making in the case of recidivism predictions, consider the importance of the question “how can we be sure that predictions do not discriminate on the basis of race?” (Lipton 2016, p. 4).

3.3 The problem with black box models for IPS

Rudin (2019) adds an important dimension to current work within the field of xAI which is also relevant for the explainability of models in general. In contrast to the frequent assumption that a problem can only be solved with a black box model to reach a certain performance, Rudin suggest that one should always try to implement a solution that is “inherently interpretable” (white box) in the first place (Rudin 2019, p. 206). Regarding current work within the field of xAI, Mittelstadt et al. (2018) summarized that “the vast majority of work in xAI produces simplified approximations of complex decision-making functions” (Mittelstadt et al. 2018). The need of simplified approximations are, according to Rudin (2019), a direct consequence of the fact that the creators of the IPS choose to use a black box model instead of a white box model. The complex decision-making functions that cannot be fully understood by humans, can often be substituted by inherently interpretable functions, i.e. a white box model, without losing accuracy or desired performance (Rudin 2019, p. 206).

3.4 Receiver-contextualised explanation

One reason that supports Rudin’s main point, that one should “stop explaining black box machine learning models for high stakes decisions and use interpretable models instead” (Rudin 2019), is the importance of a receiver-contextualised explanation (Floridi et al. 2020, p. 11). Floridi points out that for fair and ethical decision making, we have to contextualize the explanation for the receiver of the explanation which in some cases requires a white box IPS. In contrast to xAI, which focus on scientific modelling rather generating an explanation for the prediction, Floridi et al. (2020) highlights the “importance of the right conceptualisation when explaining an [IPS-based] decision.” In order to find the right conceptualisation, Floridi et al. (2020) suggest that one should describe the IPS at a level of abstraction (LoA) which satisfies the purpose of the explanation. For now, LoA could be understood as “a lens through which [an IPS] is analysed” (Floridi et al. 2020, p. 12), but LoA will be more thoroughly described later in the text. The purpose of the explanation depends on who is the receiver of the explanation. Consider a customer-friendly vs. engineer-friendly explanation, then it becomes clear that “not every LoA is appropriate for a given receiver” (Floridi et al. 2020, p. 12).

3.5 Objectives of essay

Due to given background, it is favorably to first contextualize Floridi’s LoA for IPS (section 4). Then, I will present the DN account of explanation and the IS account of explanation and also motivate why these very similar accounts are adequate for the task of explaining the output of an IPS described by Floridi’s LoA (Section 5). Thirdly, I will give an example of a hypothetical scenario S, similar to the example described in the introduction, where a black box model and a white box model have the same overall performance (section 6).

The white box IPS and the black box IPS from scenario S will then be described at the same LoA, which will be assumed to be the right conceptualisation to satisfy the purpose of the explanation (section 7). For scenario S, I will then explain, respectively, the output for the black box IPS and the white box IPS using the DN explanation (section 8). Lastly, I hope that a comparison between the two explanations can be used as a good indication of which of the models should be chosen over the other (section 8.3).

4 Floridi's LoA for describing IPS

Floridi's Level of Abstraction (LoA from here on) consists of five key components and is a formal framework for describing systems in general. In the past, Floridi's LoA has been used in several areas and the most nearby application of the framework, in scope of this essay, is probably in the philosophy of AI, where LoA was used to provide a new model of telepresence [4]. The five components in LoA are typed variable, observable, level of abstraction (component), system behaviour and moderated level of abstraction (Floridi 2008, p. 305). In the following subsections, an attempt will be made to describe the components in the context of IPS.

4.1 Typed variable

Definition 4.1 A typed variable is a uniquely-named conceptual entity (the variable) and a set, called its type, consisting of all the values that the entity may take. Two typed variables are regarded as equal if and only if their variables have the same name and their types are equal as sets. A variable that cannot be assigned well-defined values is said to constitute an ill-typed variable. (Floridi 2008, p. 305)

Independently if it is a black box IPS or a white box IPS, the model will have some sort of data coming in from the outside world. The data, i.e. the parameters for the model, should be considered the variables for the IPS and the type should be seen as a set of values these parameters can take. For IPS, data is not the only entity that should be assigned typed variables: the same goes for the generated output, where the prediction should be considered the variable and all different possible predictions are the type. In addition to assigning typed variables for the data and the prediction, an IPS could also have internal states which, if the LoA requires it, should be assigned type variables.

4.2 Observable

Definition 4.2 An observable is an interpreted typed variable, that is, a typed variable together with a statement of what feature of the system under consideration it represents. Two observables are regarded as equal if and only if their typed variables are equal, they model the same feature and, in that context, one takes a given value if and only if the other does. (Floridi, 2008, p. 306)

In the case of using LoA to describe an IPS, an observable will simply refer to what the type of data or a prediction represent. As an illustrative example of an observable in the context of IPS, consider a simple weather forecast: data about winds, temperature etc. are taken into consideration by a model which predicts how the weather will be, let assume, tomorrow. However, the actual prediction is probably just a calculated likelihood; the type (i.e. a rational number in this case) for the prediction does only say something about tomorrow's weather when we combine it with a statement of what the value represents, namely "the probability of tomorrow's weather".

4.3 Level of Abstraction (component)

Definition 4.3 An observable is called discrete if and only if its type has only finitely many possible values; otherwise it is called analogue. (Floridi 2008, p.306)

[4] Telepresence refers to a set of techniques that allow a person to feel as if they were present, to prove themselves to be present, or to deliver an effect via telerobotics in a place other than their true place.

Depending on which IPS is being described, the assigned observables can be either discrete or analogue. However, in the context of IPS, it will in practice have no actual effect if it is discrete or analogue.

Definition 4.4 A level of abstraction (LoA) is a finite but non-empty set of observables. No order is assigned to the observables, which are expected to be the building blocks in a theory characterised by their very definition. A LoA is called discrete (respectively analogue) if and only if all its observables are discrete (respectively analogue); otherwise it is called hybrid. (Floridi 2008, p.309)

A bit confusing, one component in LoA is named "level of abstraction" itself. Coming back to the example of a simple weather forecast, which predicts tomorrow's weather based on winds and temperature, the observables would be that the value of variable "winds" represent how strong the wind is, the value of variable "temperature" represent how hot it is in the air and the value of variable "tomorrow's weather" represent how the weather will be tomorrow. When these three are the observables, the component level of abstraction would be winds (how strong the wind is), temperature (how hot it is in the air), tomorrow's weather (how the weather will be tomorrow). However, it is important to note that the component level of abstraction, which consist of observables, would have been completely different if one had chosen to initially assign other typed variables: the setting of typed variables for an IPS affects the composement of the component level of abstraction.

4.4 Behaviour and moderated level of abstraction

Definition 4.5 The behaviour of a system, at a given LoA, is defined to consist of a predicate whose free variables are observables at that LoA. The substitutions of values for observables that make the predicate true are called the system behaviours. A moderated LoA is defined to consist of a LoA together with a behaviour at that LoA. (Floridi 2008, p. 310)

It is convenient to study the weather forecast example one more time. At the chosen component level of abstraction, i.e. winds (how strong the wind is), temperature (how hot it is in the air), tomorrow's weather (how the weather will be tomorrow), it is known that winds and temperature are what determine tomorrow's weather. Describing the relations between winds, temperature and tomorrow's weather will at the chosen component level of abstraction represent the system behaviour. The described relations between winds, temperature and tomorrow's weather, i.e. the system behaviour, will together with the component level of abstraction form a moderated level of abstraction, which is the end product of Floridi's LoA.

5 Scientific explanations for IPS

In Philosophy of Science, there seems to be a general agreement that a scientific explanation, independently of the chosen LoA, consist of an explanans and an explanandum. Explanans could be understood as “the explaining” and explanandum as “what is being explained”. However, philosophers often disagree on how the explanations should be analysed and for that reason there exist many different accounts of scientific explanations. When explaining the output of an IPS, I believe the deductive-nomological explanation (DN explanation) and the inductive-statistical explanation (IS explanation) are both very applicable accounts.

I will describe the DN explanation in section 5.1 and the IS explanation in section 5.2. In the last subsection (section 5.3), a short motivation will be given on why these two accounts are adequate when explaining the output of an IPS described by Floridi’s LoA.

5.1 DN explanation

The relation between explanans and explanandum in the DN-explanation is as follows: ”the explanandum must be a logical consequence of the explanans”. Furthermore, ”the sentences constituting the explanans must be true” (Hempel, 1965, p. 248). Additionally, the explanans consist of certain premises and at least one of the premises is a “law of nature” (Woodward 2019); The explanandum is derived from the premises and the explanation is therefore a deductive inference in which the premises make up the explanans.

Premise 1
Premise 2
...
Premise n
$\Rightarrow$ Explanandum

The condition that at least one of the premises has to be a “law of nature” deserves to be further elaborated. To distinguish if a premise is a law or not may in some cases be difficult without an adequate account of lawhood. However, the discussion of different accounts of lawhood is beyond this essay, but it would be favorably to not solely rely on random guesses when trying to determine if a premise is a law or not. To at least have some idea what the difference could be about, consider Hempel’s own example on the topic (Woodward 2019):

The generalization

1. “All members of the Greensbury School Board for 1964 are bald” is, if true, only accidental. In contrast,
2. “All gases expand when heated under constant pressure”, is a law.

When explaining the output for an IPS using the IS or the DN explanation, I will rely on the relations to necessity for what distinguishes a true generalization (law) from an accidental true generalization (not a law). Van Fraassen (1989, p. 28) expresses the condition of relations to necessity with following analogies:

> Wood burns when heated, because wood must burn when heated. And it must burn because of the laws which govern the behaviour of the chemical elements of which wood and the surrounding air are composed. Bodies do not fall by chance; they must fall because of the law of gravity. In such examples as these we see a close connection between ‘law’ and ‘must,’ which we should stop to analyse.” (Van Fraassen 1989, p. 28)

Using the logic of relations to necessity on Hempel's example, it becomes intuitively clear that (2) is a law and (1) is not a law. A more in debt analysis of what should be considered a law is not in the scope in this essay.

5.2 IS explanation

When explaining the output of an IPS, the prediction, i.e. the explanandum, is not always an absolute multinomial value. In real life, a prediction often comes with a probability. In the case of a weather forecasts, the prediction is rather "there is a 75 % chance that it will be sunny tomorrow" instead of just "it will be sunny tomorrow". When the prediction of an IPS comes with a probability, a DN explanation will not be sufficient because one cannot deduce the explanandum from the premises if the law is a statistical law. For the DN explanation, we consider a law as when an outcome must happen because of that law. For the IS explanation, a statistical law is a law which specifies the probability of an outcome. Instead of requiring a deductive inference of the explanandum from the explanans, the IS explanation requires probabilistically inferring the explanandum from the explanans. However, as can be seen in the inference below, the IS explanation uses the same structure as the DN explanation.

Premise 1
Premise 2
...
Premise n
$\approx >$Explanandum

Due to the explanandum being a probabilistic inference from the explanans, the probability of the explanandum has to be larger than a certain value p. If the probability of the explanandum is low, we cannot use the premises to provide an IS explanation. The value of p has to be determined from context. However, if the probability of the explanandum happens to be equal to 1, the explanation is not only an IS explanation but also a DN explanation: all DN explanations are IS explanations and an IS explanation is a DN explanation if and only if p = 1.

5.3 DN- and IS explanation's adequacy for explaining the output of an IPS described by LoA

Describing an IPS with Floridi's LoA gives a very technical result and the DN- and the IS explanation are for this reason adequate accounts for explaining the output of IPS described by LoA. As for computer programs, entities in an IPS described by LoA are assigned typed variables. The final result of Floridi's LoA, i.e. a moderated level of abstraction, is a collection of observables (component level of abstraction) together with how the typed variables relate to each other (system behaviour). With the structure of the DN- and the IS explanation, where an explanandum is inferred from premises, it is a smooth transition from a LoA description to a scientific explanation: The explanandum in the DN- and the IS explanation will be represented by the IPS's output's assigned typed variable. After all, it is the prediction of the IPS we want to explain. Furthermore, the premises will be interchanged against the system behaviour. To exemplify, for the system behaviour of a weather forecast described by LoA, certain values for wind, temperature etc. would make up the premises of the explanans. Tomorrow's predicted weather, e.g. sunny, rainy etc., would make up the explanandum.

6 A black box IPS and a white box IPS that generate identical output: a hypothetical scenario

For some high stakes predictions it may be possible to construe two different models that generate the same output and also have the same accuracy, where one model is a black box model and the other is a white box model. I will give an example of one black box model and one white box model of above interchangeable kind. These two example models will then be described using the LoA framework and the models' generated output will be explained using the DN explanation at the chosen LoA. Although the COMPAS model and the COREL model from the introduction are very good candidates for playing the role of these two example models, I chose to construct hypothetical models instead (see Model W and Model B below). It should be noted that Model W ('W' as in white box) and Model B ('B' as in black box) are highly influenced by the COREL model and the COMPAS model, but the former were substituted for the following reasons:

- Constructing hypothetical models give the opportunity to get rid of mathematical and technical details that are not relevant in the context of this essay. Specifically, the focus should be the receiver-contextualised explanation for the prediction of the two different models; not the machine learning details. For the purpose of comparing the receiver-contextualised explanations, it is critical that the two example models meet the criteria for the definition of black box and white box. But including all technical details in the model description would possibly just draw attention from what is really needed to meet the criteria of the definition of black box and white box.

- Using hypothetical models instead of trying to reconstruct the COMPAS model and the COREL model takes away the risk of incorrectly reconstructing them, which would not be fair towards the original creators of the two models.

- In the case of COMPAS, one option could be to just reuse the model description given by Rudin et al. (2020a, p. 7). However, the description in the article is based on the creator's documentation which is a manual at 59 pages (Northpoint 2015). Although the description is probably a good summary, there still have been a selection of what should be included in the description. For this reason, the right thing to do would be to work through the manual doing my own selection. With a hypothetical model, less time is spent filtering what is relevant from the creator's documentation when trying to only include relevant parts of the model.

These are the three main reasons why a hypothetical scenario S will be made up with Model W and Model B instead of the real life example of the US justice system. Nevertheless, it should be clearly stated that S, Model W and Model B are in many cases just a suitable framing of Rudin's et al. work regarding COMPAS and COREL (Rudin 2020a, p. 7; Rudin et al. 2020b).

6.1 A hypothetical scenario S

Person P has committed a crime, confessed the same crime, and is now facing time in prison. However, how many years judge J will send P to prison depends on an IPS's predicted recidivism risk for P. There are two different models that could be used in the IPS, Model W and Model B, which both have identical accuracy for predicting the recidivism risk for person P. Both models base their prediction on recidivism statistics on persons with different attributes (age, sex etc.), and the two models are using the same training data. In other words, which of the IPS's prediction J choose to base his decision on will not affect the length of P's imprisonment. The predicted recidivism risk for P, independently of what model was being used, is represented by an integer from 1 to 10, where a higher number suggests larger risk of criminal recidivism. In this made up scenario, judges' interpretation of the integer from 1 to 10, when deciding on how the number should affect the length of the prison sentence, is assumed to be consistent.

6.2 A white box model for S: Model W

Model W takes three parameters into consideration when predicting the recidivism risk for P: P's current age, P's sex and how many times P has been convicted in court in the past. When the parameters has been given a value, they are passed to the following conditional scheme:

```
IF (P's current age) is 18 and (P's sex) is male and (prior convictions) is $> 3$
THEN predict 10
ELSE IF (P's current age) is 19 and (P's sex) is male and (prior convictions) is $> 3$
THEN predict 9
ELSE IF (P's current age) is 20 and (P's sex) is male and (prior convictions) is $> 3$
THEN predict 8
ELSE IF (P's current age) is 21 and (prior convictions) is $> 2$
THEN predict 7
ELSE IF (P's current age) is 22 and (prior convictions) is $> 2$
THEN predict 6
ELSE IF (P's current age) is 23 and (prior convictions) is $> 2$
THEN predict 5
ELSE IF (prior convictions) is $> 3$
THEN predict 4
ELSE IF (prior convictions) is 3
THEN predict 3
ELSE IF (prior convictions) is 2
THEN predict 2
ELSE predict 1
```

As can be derived from above conditional scheme, an integer between 1 and 10 will always be generated as output, which is then taken into consideration by J who decides the length of P's imprisonment.

6.3 A black box model for S: Model B

In comparison with Model W, the number of features taken into consideration are much higher for Model B. First step in Model B is that P answers a questionnaire, which consist of 50 question divided into five different categories (10 questions in each category): criminal involvement, history of noncompliance, history of violence, educational and substance abuse. P's answers in each categories are then mapped, by an unknown function F, to corresponding scalar variables CI, HoN, HoV, E and SA.

P's answers in questionnaire category "criminal involvement" → CI
P's answers in questionnaire category "history of noncompliance" → HoN
P's answers in questionnaire category "history of violence" → HoV
P's answers in questionnaire category "educational" → E
P's answers in questionnaire category "substance abuse" → SA

In addition to these five variables, data is gathered on P's age at the time of the current offense (current age) and age at the time of the first offense (age-at-first-arrest). Now, when the model's parameters has been given a value, they are passed to the following equation, where w is an unknown weighted scalar:

$$\begin{aligned}&(\text{current age} \times -w) + (\text{age-at-first-arrest} \times -w) + (\text{CI} \times w) + (\text{HoN} \times w)\\ &+ (\text{HoV} \times w) + (\text{E} \times w) + (\text{SA} \times w)\\ &= \text{integer between 1 and 10}\end{aligned} \tag{1}$$

Analogously with the opacity of F, no information on how the value of w is assigned is provided. As for Model B, the predicted integer between 1 and 10 is taken into consideration by J who decides the length of P's imprisonment.

7 The right LoA in scenario S

In the search for a receiver-contextualised LoA of an IPS, one should aim to satisfy the purpose of the explanation (Floridi et al. 2020, p. 11). When formulating the purpose of the explanation, is it wise to study the three main motives why explaining the prediction of an IPS is important: trust, informativeness and fair and ethical decision-making (Lipton 2016,pp. 3-4). Additionally, it is also important to point out that it is likely that different stakeholders in an arbitrary scenario have different purposes for an explanation. In search for a receiver-contextualised LoA description of an IPS, one has to decide whose purpose for the explanation one should aim to satisfy. However, in the case of scenario S, it will be assumed that a receiver-contextualised LoA should help to satisfy the following purpose of an explanation:

> An explanation of the predicted recidivism value should make it possible for a judge J to decide whether he or she is applying the law correctly in his or her decision, it should make it possible for person P to understand how the verdict was reached, and it should make it possible for the prosecutors and the lawyers of the defense to address J's decision, say, in an appeal.

In the following LoA descriptions of Model W and Model B, choices of the typed variables etc. will aim to satisfy the above purpose of an explanation in the best way.

7.1 Description of Model W with LoA

7.1.1 Typed variables

Using Floridi's own notation for a typed variable, x:X means that x is a variable of type X (Floridi 2008, p. 4). Starting with the data of the white box IPS in scenario S, a natural assignment of typed variables is the following: P's current age is encoded by ca:ℕ, P's sex is encoded by s:T1 and prior convictions are encoded by pc:ℕ:, where type T1 is {man, women}. In case of Model W, only one entity is left to assign a variable to and that is the actual prediction of the model. The assignment of the typed variable for the prediction is simply P's predicted recidivism risk, which is encoded by p:ℕ.

7.1.2 Observables

By giving the chosen typed variables meaning, i.e. combining a "variable together with a statement of what feature of the system under consideration it represents" (Floridi, 2008, p.306), they become observables. It is known from the original description that the variable ca:ℕ represents the current age of a person, the variable s:T1 represents the sex of person, the variable pc:ℕ represents how many times a person has been convicted in court in the past and the variable p:ℕ represent how likely a person is to commit a crime again.

7.1.3 Level of abstraction (component)

The component level of abstraction is a collection of the observables at chosen LoA, which, in vector form, in this case can be expressed as follows:

$$\begin{gathered} <\text{ca:}\mathbb{N}\text{ (age of a person), s:T1 (the sex of person),} \\ \text{pc:}\mathbb{N}\text{ (how many times a person has been convicted in court in the past),} \\ \text{p:}\mathbb{N}\text{ (how likely a person is to commit a crime again)} > \end{gathered} \tag{2}$$

7.1.4 System behaviour

Describing the system behaviour for IPS is a two step process: First, a restriction for each observable has to be made, e.g. $ca < 130$ (assuming humans cannot live longer than 130 years). The second step is to describe how the observables relate to each other, e.g. if $ca = 18$, $s = male$ and $pc > 3$ then it must follow that $p = 10$. Because Model W is a white box model, the system behaviour will, at this LoA, be very similar to the condition scheme described in the original model. With that said, the system behaviour of Model W could be expressed as follows:

```
$ca \leq 130$, $pc \leq 1000$, $0 \leq p \leq 10$ and
IF $ca = 18$, $s = male$ and $pc > 3$              THEN $p = 10$
ELSE IF $ca = 19$, $s = male$ and $pc > 3$         THEN $p = 9$
ELSE IF $ca = 20$, $s = male$ and $pc > 3$         THEN $p = 8$
ELSE IF $ca = 21$, $pc > 2$                        THEN $p = 7$
ELSE IF $ca = 22$ and $pc > 2$                     THEN $p = 6$
ELSE IF $ca = 23$ and $pc > 2$                     THEN $p = 5$
ELSE IF $pc > 3$                                   THEN $p = 4$
ELSE IF $pc = 3$                                   THEN $p = 3$
ELSE IF $pc = 2$                                   THEN $p = 2$
ELSE $p = 1$
```

7.1.5 Moderated level of abstraction (end result of LoA)

Finally, a moderated level of abstraction can be expressed by combining equation (2) and the system behaviour:

$<$ ca:$\mathbb{N}$ (age of a person), s:T1 (the sex of person),
pc:$\mathbb{N}$ (how many times a person has been convicted in court in the past),
p:$\mathbb{N}$ (how likely a person is to commit a crime again) $>$

and

```
$ca \leq 130$, $pc \leq 1000$, $0 \leq p \leq 10$ and
IF $ca = 18$, $s = male$ and $pc > 3$              THEN $p = 10$
ELSE IF $ca = 19$, $s = male$ and $pc > 3$         THEN $p = 9$
ELSE IF $ca = 20$, $s = male$ and $pc > 3$         THEN $p = 8$
ELSE IF $ca = 21$, $pc > 2$                        THEN $p = 7$
ELSE IF $ca = 22$ and $pc > 2$                     THEN $p = 6$
ELSE IF $ca = 23$ and $pc > 2$                     THEN $p = 5$
ELSE IF $pc > 3$                                   THEN $p = 4$
ELSE IF $pc = 3$                                   THEN $p = 3$
ELSE IF $pc = 2$                                   THEN $p = 2$
ELSE $p = 1$
```

7.2 Description of Model B with LoA

7.2.1 Typed variables

The assignment of variables for Model B is very similar to the assignment of variables for Model W. However, as we will see, it gets a bit more complicated regarding the remaining components for LoA due to the opacity of Model B in contrast to the transparency of Model W.

As for Model B, aiming to describe the black box IPS at the same LoA as for the white box IPS, the following typed variables can be assigned: P's current age is encoded by ca:$\mathbb{N}$ and P's age-at-first-arrest is encoded by aafa:$\mathbb{N}$. The remaining data in the model, i.e. CI, HoN, HoV, E, SA and w, are in a way already typed variables. The reason why an assignment of typed variables already have taken place is because P's answers are mapped by a black box function to a corresponding scalar variable. This is typical for black box models: a post hoc model is created to approximate the actual prediction-generating-mechanisms of the model and the result is the beginning of a LoA description of the original model.

Nevertheless, the assignment of typed variables for Model B can be finalized as CI is encoded by ci:$\mathbb{R}$, HoN is encoded by hon:$\mathbb{R}$, HoV is encoded by hov:$\mathbb{R}$, E is encoded by e:$\mathbb{R}$, SA is encoded by sa:$\mathbb{R}$ and w is encoded by w:$\mathbb{R}$. Finally, as for Model W, P's predicted recidivism risk is assigned to p:$\mathbb{N}$.

7.2.2 Observables

As for Model W, ca:$\mathbb{N}$ represents the age of a person and the variable aafa:$\mathbb{N}$ represents the age of person at arrest. Furthermore, the variables ci:$\mathbb{R}$, hon:$\mathbb{R}$, hov:$\mathbb{R}$, e:$\mathbb{R}$ and sa:$\mathbb{R}$ measure how a person answered in corresponding questionnaire category. For example, a relatively high value for ci:$\mathbb{R}$ represents that the person has a high level of criminal involvement. As for Model W, the variable p:$\mathbb{N}$ represent how likely a person is to commit a crime again.

7.2.3 Level of abstraction (component)

In vector form, the component level of abstraction for Model B can expressed as follows:

$$\begin{aligned}
&< \text{ca:}\mathbb{N} \text{ (current age P), aafa:}\mathbb{N} \text{ (P's age at first arrest),}\\
&\text{ci:}\mathbb{R} \text{ (measurement of P's criminal involvement), hon:}\mathbb{R} \text{ (measurement of P's history}\\
&\text{of noncompliance), hov:}\mathbb{R} \text{ (measurement of P's history of violence),}\\
&\text{e:}\mathbb{R} \text{ (measurement of P's education), sa:}\mathbb{R} \text{ (measurement of P's substance abuse),}\\
&\text{w:}\mathbb{R} \text{ (weighted scalar in model's calculation), p:}\mathbb{N} \text{ (how likely a person is to commit}\\
&\text{a crime again)} >
\end{aligned} \tag{3}$$

7.2.4 System behaviour

Firstly, stating the obvious, we can restrict the typed variables as ca $\leq$ 130, 10 $\leq$ aafa $\leq$ 130 and 1 $\leq$ p $\leq$ 10. Unfortunately, for the remaining typed variables, it is hard to do any restriction because of the opacity of given post hoc model; we know that 1 $\leq$ p $\leq$ 10 but due to that w is unknown it is impossible to estimate the magnitude ci:$\mathbb{R}$, hon:$\mathbb{R}$, hov:$\mathbb{R}$, e:$\mathbb{R}$ and sa:$\mathbb{R}$.

Secondly, in contrary to the system behaviour for Model W, Model B does not necessarily have a deterministic output. Due to the black box function which maps P's answers in the questionnaire to scalar values for ci:$\mathbb{R}$, hon:$\mathbb{R}$, hov:$\mathbb{R}$, e:$\mathbb{R}$ and sa:$\mathbb{R}$, we cannot say with certainty that "if P answers the questionnaire in this way, the predicted recidivism risk must be that". For this reason, the system behaviour of Model B has to be expressed as an interpretation of equation (1). First, it is favorably to rewrite equation (1) using the newly assigned typed variables:

$$\begin{aligned} \mathrm{p} \ &= (\mathrm{ca} \times -w) + (\mathrm{aafa} \times -w) + (\mathrm{ci} \times w) + (\mathrm{hon} \times w) \\ &\quad + (\mathrm{hov} \times w) + (\mathrm{e} \times w) + (\mathrm{sa} \times w) \end{aligned} \tag{4}$$

Then, it is true to say for the system behaviour of Model B that relative high values of ci:ℝ, hon:ℝ, hov:ℝ, e:ℝ and sa:ℝ will give a high probability for person P committing a crime again.

7.2.5 Moderated level of abstraction (end result of LoA)

As for Model W, a moderated level of abstraction is expressed by combining equation (3) and the system behaviour:

< ca:ℕ (current age P), aafa:ℕ (P's age at first arrest),
ci:ℝ (measurement of P's criminal involvement), hon:ℝ (measurement of P's history of noncompliance), hov:ℝ (measurement of P's history of violence),
e:ℝ (measurement of P's education), sa:ℝ (measurement of P's substance abuse),
w:ℝ (weighted scalar in model's calculation), p:ℕ (how likely a person is to commit a crime again) >

and

$$\begin{aligned} \mathrm{p} \ &= (\mathrm{ca} \times -w) + (\mathrm{aafa} \times -w) + (\mathrm{ci} \times w) + (\mathrm{hon} \times w) \\ &\quad + (\mathrm{hov} \times w) + (\mathrm{e} \times w) + (\mathrm{sa} \times w) \end{aligned}$$

8 Explaining IPS's prediction with the DN explanation

In scenario S, the right purpose of an explanation was assumed to be that:

> An explanation of the predicted recidivism value should make it possible for a judge J to decide whether he or she is applying the law correctly in his or her decision, it should make it possible for person P to understand how the verdict was reached, and it should make it possible for the prosecutors and the lawyers of the defense to address J's decision, say, in an appeal.

Based on this assumed purpose of explanation, Model W and Model B were described by applying LoA's five components to the models which as an end result gave two different moderated level of abstractions.

For both the IS explanation and the DN explanation, the task of explaining the output from an IPS described with LoA consist of two steps: Firstly, from the moderated level of abstraction, identify premises, a premise that also is a law and what the explanandum should be. Secondly, infer the explanandum from the premises.

8.1 Explaining the recidivism risk in the case of Model W

So, in an attempt to explain the output from the white box IPS in scenario S, a good starting point is to study the final moderated level of abstraction of Model W in section 7.1.5. Based on the purpose of an explanation, it follows that p:$\mathbb{N}$ (how likely a person is to commit a crime again) should be the explanandum in the explanation. Now, due to the low level system behavior of Model W, we will be able to explain all possible values for p:$\mathbb{R}$: Suppose person P asks "why was my recidivism risk predicted to a likelihood of 7?". Then, a DN explanation can simply be given as "your current age is 21 and your number of prior of convicts are 2, then it must follow that your recidivism risk is 7 out of 10".

8.2 Explaining the recidivism risk in the case of Model B

As when explaining the prediction of Model W, we will start by looking at the corresponding moderated level of abstraction of Model B from section 7.2.5. Completely analogously with Model W, it follows that p:$\mathbb{N}$ (how likely a person is to commit a crime again) should be the explanandum of the explanation of the black box IPS. When formulating the explanans, it becomes clear that the description of Model B will lead to different premises than for Model W. As for Model W, we presume that person P asks "why was my recidivism risk predicted to a likelihood of 7?". The only premise expressed in the system behaviour, that also is a law, is that "if the sum of ci:$\mathbb{R}$, hon:$\mathbb{R}$, hov:$\mathbb{R}$, e:$\mathbb{R}$ and sa:$\mathbb{R}$ is relatively high then must p:$\mathbb{N}$ also take a high value". This poor description of the system behaviour is a result of the opacity of Model B because we cannot describe how P's answers in the questionnaire truly affects the predicted recidivism risk. A poor description yields a poor explanation, and to answer P's question we can only give a DN-explanation as "your answers in the questionnaire lead to that the sum of ci:$\mathbb{R}$, hon:$\mathbb{R}$, hov:$\mathbb{R}$, e:$\mathbb{R}$ and sa:$\mathbb{R}$ became relatively high and therefore must p also take a high value".

9 Conclusion

When comparing the two different DN explanations in scenario S, it is obvious that they did not end up the same. It is no surprise that two models with different described system behaviour will have different explanations for their output. However, it is not necessarily the case that one model should be preferred over the other model just because the explanations differ. What is critical though, is how well the explanations satisfy the purpose of the explanation. Choosing between Model W and Model B in scenario S, one should therefore study the purpose for an explanation, the explanation of Model W and the explanation of Model B. In the case of scenario S, I believe it is safe to say that Model W outperforms Model B regarding satisfying the purpose for an explanation. Furthermore, for arbitrary IPS where the white box version performs equally to the black box version in terms of accuracy, I hope that investigating how well the two models' explanations satisfy the purpose for an explanation will be less subjective than motivating the choice of a certain model because it is a high stake decision or not.

Throughout the essay, I have avoided to touch on two problematic aspects that comes with the initially suggested method of letting the transparency of the IPS depend on how well an explanation of the model's output satisfies the purpose of an explanation. The first problem is that it may not be clear what the purpose of an explanation should be. Floridi et. al. (2020) suggest that one should describe the IPS at a LoA which satisfies the purpose of the explanation. However, Floridi et al. (2020) do not point out that it might be hard to define the purpose of the explanation.

The second problem with the suggested method is that it can be hard to verify that the given explanation actually satisfies the purpose of an explanation. In scenario S it was obvious that the explanation of Model W outperformed the explanation of Model B, but that does not always have to be the case.

In this essay I have suggested an alternative method to let a model's transparency depend on how well an explanation of the model's output satisfies the purpose of an explanation. The reason behind this method was to get away from the non-rigorous criterion that the model's transparency should depend on if its prediction contributes to a high stake decision or not. However, although the method succeeded regarding escaping the debate what should be considered a high stake decision, new questions arose such as ”what is the right LoA?” and ”how to decide if the explanation satisfies the purpose of an explanation?”.

10 References

Floridi, Luciano. (2008). The Method of Levels of Abstraction. Minds & Machines, 18:303–329.
doi: 10.1007/s11023-008-9113-7

Floridi, Luciano., Cowls, J., King, T.C. et al. (2020). How to Design AI for Social Good: Seven Essential Factors. Sci Eng Ethics (2020).
https://doi.org/10.1007/s11948-020-00213-5

Hempel, Carl, (1965). Aspects of Scientific Explanation and Other Essays in the Philosophy of Science, New York: Free Press.

Kitcher, Philip. (1998). Explanation. Routledge Encyclopedia of Philosophy, Version 1.0. London, New York: Routledge (1998)
doi:10.4324/9780415249126-Q034-1.

Lipton, Zachary. (2016). The Mythos of Model Interpretability. Communications of the ACM, 61.
doi: 10.1145/3233231.

Mittelstadt, Brent Russell, Chris Wachter, Sandra. (2018). Explaining Explanations in AI.
doi: 10.1145/3287560.3287574

Northpoint. (2015). Practitioner's Guide to COMPAS Core.
http://www.northpointeinc.com/downloads/compas/Practitioners-Guide-COMPAS-Core-_031915.pdf

Rudin, Cynthia. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. Nat Mach Intell 1, 206–215 (2019).
https://doi.org/10.1038/s42256-019-0048-x

Rudin, Cynthia, Wang, C., Coker, B. (2020a). The Age of Secrecy and Unfairness in Recidivism Prediction. Harvard Data Science Review, 2(1).
https://doi.org/10.1162/99608f92.6ed64b30

Rudin, Cynthia et al.. (2020b). "About CORELS".
https://corels.eecs.harvard.edu/corels/index.html

Van Fraassen, Bas. (1989). Laws and Symmetry. Oxford University Press Inc., New York.

Woodward, James. (2019). Scientific Explanation. The Stanford Encyclopedia of Philosophy (Winter 2019 Edition), Edward N. Zalta (ed.).
https://plato.stanford.edu/archives/win2019/entries/scientific-explanation