Co-Funded by the Horizon 2020 programme of the European Union

Grant Agreement: 875358

FAITH

intelligent patient support

Deliverable D4.5

Explainable AI Framework

Work package:	WP4 – Data Analysis & Federated AI Services
Prepared By/Enquiries To:	Philip O’Brien (pobrien@tssg.org) – Waterford Institute of Technology Javier Rojo Lacal (jlacal@lst.tfo.upm.es) - UPM
Reviewers:	Stefanos Venios – Suite5
Status:	QA’d ready for release.
Date:	30/12/2020
Version:	1.0
Classification:	Public

Authorised by:

Gary McManus

Gary McManus
WIT

Authorised date: 08/01/2021

Disclaimer:

This document reflects only authors' views. Every effort is made to ensure that all statements and information contained herein are accurate. However, the Partners accept no liability for any error or omission in the same. EC is not liable for any use that may be done of the information contained therein.

FAITH Project Profile

Contract No H2020-ICT- 875358

Acronym	**FAITH**
Title	**a Federated Artificial Intelligence solution for moniToring mental Health status after cancer treatment**
URL	**https://h2020-faith.eu/**
Twitter	**https://twitter.com/H2020_Faith**
LinkedIn	**linkedin.com/company/faith-project**
Facebook	**https://fb.me/H2020.FAITH**
Start Date	**01/01/2020**
Duration	**36 months**

FAITH Partners

List of participants

Participant No	Participant organisation name	Short Name	Country
1 (Coordinator)	WATERFORD INSTITUTE OF TECHNOLOGY.	WIT	Ireland
2	UPMC Whitfield, Euro Care Healthcare Limited.	UPMC	Ireland
3	Universidad Politécnica de Madrid.	UPM	Spain
4	Servicio Madrileño de Salud.	SERMAS	Spain
5	UNINOVA, Instituto de Desenvolvimento de Novas Tecnologias.	UNINOVA	Portugal
6	Fundação D. Anna de Sommer Champalimaud e Dr. Carlos Montez Champalimaud.	CF	Portugal
7	Deep Blue.	DBL	Italy
8	Suite5 Data Intelligence Solutions Limited.	SUITE5	Cyprus
9	TFC Research and Innovation Limited.	TFC	Ireland

SC1-DTH-01-2019: Big data and Artificial Intelligence for monitoring health status and quality of life after the cancer treatment

H2020-SC1-DTH-2019

FAITH is co-funded by the European Commission - Agreement Number 875358 (H2020 Programme)

Document Control

This deliverable is the responsibility of the Work Package Leader. It is subject to internal review and formal authorisation procedures in line with ISO 9001 international quality management system procedures.

Version	Date	Author(s)	Change Details
0.1	08/12/20	Philip O'Brien (WIT)	Table of Contents defined.
0.2	08/12/20	Philip O'Brien (WIT)	Structure expanded.
0.3	08/12/20	Philip O'Brien (WIT)	References/links added to relevant sections.
0.4	09/12/20	Philip O'Brien (WIT)	Editing introduction.
0.5	14/12/20	Philip O'Brien (WIT)	Update all sections.
0.6	15/12/20	Philip O'Brien (WIT)	Update all sections.
0.7	15/12/20	Javier Rojo Lacal (UPM)	Update bias, uncertainty and tools sections.
0.8	21/12/20	Stefanos Venios (Suite5)	Internal review.
0.9	30/12/20	Philip O'Brien (WIT)	Final edits and formatting.
09.1	30/12/20	Tom Flynn (TFC)	QA review.
1.0	08/01/2021	Gary McManus (WIT)	Final release for submission to European Commission portal.

Executive Summary

Objectives:

To ensure AI is ethical, it must be transparent. It is prudent, therefore, for an AI to provide not only an output, but also a human understandable explanation that expresses the rationale of the machine. T4.3 will provide a FAITH library of ML and HCI modules that provide for more understandable AI implementations, which will give our healthcare stakeholders more insight into the decisions that were made by the FAITH framework, and therefore more confidence in the decision-making process. WIT is responsible for this task and will be supported by the FAITH project partners, Suite5 and UPM, in the implementation of this Explainable AI framework. This deliverable reflects the work undertaken as part of T4.3 and is released in three stages at M12, M24, and M39.

Results:

The primary results of this deliverable, and particularly this iteration (M12) are an overview of the area of Explainable AI, and also those related areas that we believe need to be treated in parallel e.g., model reproducibility.

TABLE OF CONTENTS

TABLE OF FIGURES

LIST OF TABLES

Not applied in this deliverable.

1 INTRODUCTION

As the reach of Artificial Intelligence (AI) grows, transforming industries such as medicine, transport and defence, we find ourselves entrusting our health, safety and security to intelligent machines. A worry for many, however, is that these machines are "black boxes" i.e., closed systems that receive an input, produce an output, and offer no clue why. As Cathy O'Neil explains in Weapons of Math Destruction, algorithms often determine what college we attend, if we get hired for a job, if we qualify for a loan to buy a house, and even who goes to prison and for how long. Unlike human decisions, these mathematical models are rarely questioned. They just show up on somebody's computer screen and fates are determined. Consider the case of Sarah Wysocki, a fifth-grade teacher who — despite being lauded by parents, students, and administrators alike — was fired from the D.C. school district because an algorithm judged her performance to be sub-par. *Why?* It's not exactly clear, because the system was too complex to be understood by those who fired her.[1]

Engineers may be able to deliver ever more accurate models, forecasting pandemic spread, classifying symptoms of mental disease etc. but if they cannot explain these models to the relevant decision-makers e.g., doctors, public health officials, politicians, then how can the models be trusted? Were something to go wrong, being unable to explain *why* could be the death knell to an otherwise transformative technology.

Explainable AI (XAI) was pinpointed as one of the project's four key areas of ambition. XAI is obviously of particular importance in healthcare, where the AI service might be making a clinical decision that could affect a person's life. The FAITH AI, however, isn't making an automated decision, it is providing the professional with an alert, leaving any diagnosis in their hands. That being said, we believe factoring in transparency and explainability from the start will strengthen FAITH for long-term adoption.

Marvin Minsky, one of the pioneers of AI, once said that "our mind contains processes that enable us to solve problems we consider difficult. Intelligence is the name for whichever of these processes we don't understand". An article[2] on XAI cleverly changes this to say "our model contains processes that enable us to solve problems we consider difficult. 'Black-box' is the name for whichever of these processes we don't yet understand."

[1] https://hbr.org/2019/10/we-need-ai-that-is-explainable-auditable-and-transparent

[2] https://towardsdatascience.com/explainable-ai-vs-explaining-ai-part-1-d39ea5053347

2 ABBREVIATIONS AND ACRONYMS

Abbreviation	Description
AI:	Artificial Intelligence
DL	Deep Learning
DVC:	Data Version Control
IG	Integrated Gradients
ISO:	International Organization for Standardization
LIME	Local Interpretable Model-Agnostic Explanations
LRP	Layer-wise Relevance Propagation
ISO9001-2015:	International Quality Management Systems.
ML:	Machine Learning
NTSB	National Transportation Safety Board
SHAP	Shapley additive explanations
XAI:	Explainable Artificial Intelligence

3 EXPLAINABLE AI BACKGROUND

This section will present an overview of the growing field of Explainable AI, explaining why it is important, and what it currently looks like. In terms of the FAITH project, we also include other relevant topics under the umbrella of this task e.g., reproducibility, bias, transparency etc. It is our opinion that this holistic approach is the most beneficial for the success of the project.

The first death on record involving a self-driving car occurred in Tempe, Arizona, in 2018.[3] Elaine Herzberg was killed as she wheeled a bicycle across the road and was struck by an Uber self-driving car. Although lengthy investigations by police and the US National Transportation Safety Board (NTSB) found that human error was mostly to blame for the crash (the back-up driver was found to have been using her phone), it brought to the fore new concerns around AI. As intelligent machines play an ever more important part in our lives, how can we trust them, particularly if they fail us? The vehicle's automatic systems failed to identify Ms Herzberg and her bicycle as an imminent collision danger in the way they were supposed to.

It is no wonder that explainability in machine learning is a very active topic, even receiving its own symposium at NIPS 2017. As Machine Learning (ML) systems become ever more widespread so does the fear of them being black boxes, i.e., closed systems that receive an input, produce an output, and offer no clue why. To ensure AI is ethical and trustable, it must be transparent. It is prudent, therefore, for an AI to provide not only an output, but also a human understandable explanation that expresses the rationale of the machine (Doran, Schulz and Besold 2017). As AI pushes into the mainstream this idea gets a lot of attention, even appearing as a section in the New York Times[4].

It's almost never enough to have a model that works well. We need to understand how a model works not just because we are scientifically curious, but also to make sure that it's not taking shortcuts.[5] We believe the real power of interpretability, however, lies in its correlation to trust. If we know something works well, and we know why it works well, then we are much more likely to trust it, and rely on it. Conversely, if we cannot rely on machine learning models then what opportunities are we missing out on?

[3] https://www.bbc.com/news/technology-54175359#:~:text=The%20back%2Dup%20driver%20of,Tempe%2C%20Arizona%2C%20in%202018.

[4] https://www.nytimes.com/2017/11/21/magazine/can-ai-be-taught-to-explain-itself.html

[5] https://www.nature.com/articles/s42256-020-00257-z.epdf?sharing_token=VMBI-Iokx50rvr-TxRie5NRgN0jAjWel9jnR3ZoTv0MXHH5Cwsvg3-c3drbqV45Glrtz54f1leY0_1VRr0DcLgLuXwOEXMOBDE-KQ4r08bibg25tSUnQA65dAMyEyDxN3Yz4AyD_ewFexhOiWWIvOcUGq6v7h6e1CH7WvBb7alk%3D

In truth, there are a range of reasons why some form of interpretability in AI systems might be desirable. These include:[6]

- Giving users confidence in the system.
- Safeguarding against bias.
- Meeting regulatory standards or policy requirements.
- Improving system design.
- Assessing risk, robustness, and vulnerability.
- Autonomy, agency, and meeting social values.
- Understanding and verifying the outputs from a system.

One of the most important things we must do is ensure those working in this area have shared context, i.e., that we have well-accepted definitions of the terms we use. The Royal Society[7] provide a list of definitions, useful at least as starting points in this work:

- Interpretable, implying some sense of understanding how the technology works.
- Explainable, implying that a wider range of users can understand why or how a conclusion was reached.
- Transparent, implying some level of accessibility to the data or algorithm.
- Justifiable, implying there is an understanding of the case in support of a particular outcome.
- Contestable, implying users have the information they need to argue against a decision or classification.

As (Lipton 2018) points out, *model interpretability* is often suggested as a remedy to the problem of humans unable to understand machine learning models, but few articulate precisely what interpretability means or why it is important. They break interpretability down into two types: transparency and post-hoc, a categorisation we find useful.

It is worth noting that the problem of interpretability is not exclusive to AI. Underlying concerns about human comprehensibility and generating explanations for decisions is a general issue in cognitive science, social science, and human psychology. (Miller 2018)

[6] https://royalsociety.org/-/media/policy/projects/explainable-ai/AI-and-interpretability-policy-briefing.pdf
[7] https://royalsociety.org/-/media/policy/projects/explainable-ai/AI-and-interpretability-policy-briefing.pdf

A subset of the many considerations in this rapidly growing field are captured in **Figure 1**. This taxonomy arranges models in terms of the kinds of explainability that are enabled (Miller 2018).

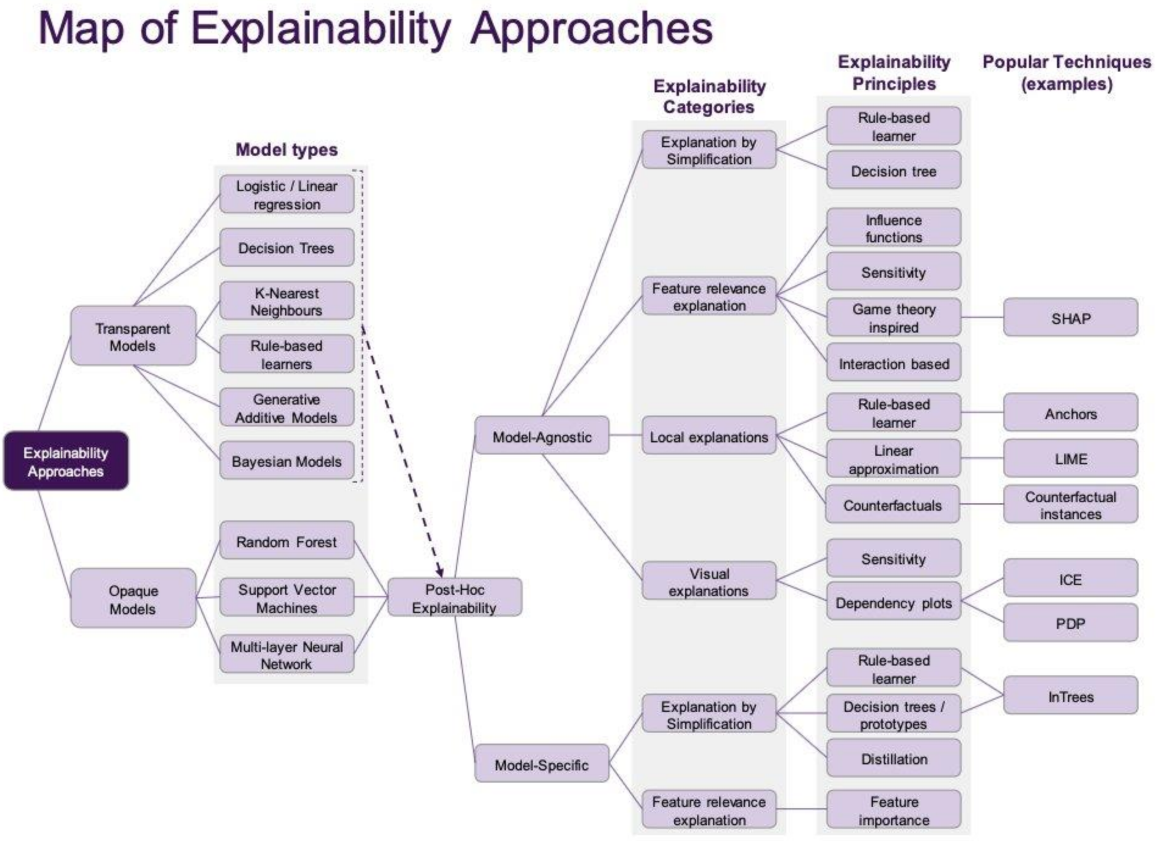


Figure 1 Sample Range of XAI Techniques[8]

3.1 Stakeholders

Different users require different forms of explanation in different contexts i.e., interpretability will look different depending on who is using it, the most likely stakeholders to consider are:

- Data scientists and developers, ML practitioners:
  - Benefited when debugging a model or when looking for ways to improve performance.
- Business owners:
  - Caring about the fit of a model with business strategy and purpose.
- Model risk analysts:

[8] https://arxiv.org/abs/2009.11698

  - Challenging the model, in order to check for robustness and approving for deployment.
- Regulators:
  - Inspecting the reliability of a model, as well as the impact of its decisions on the customers.
- Consumers:
  - Requiring transparency about how decisions are taken, and how they could potentially affect them.

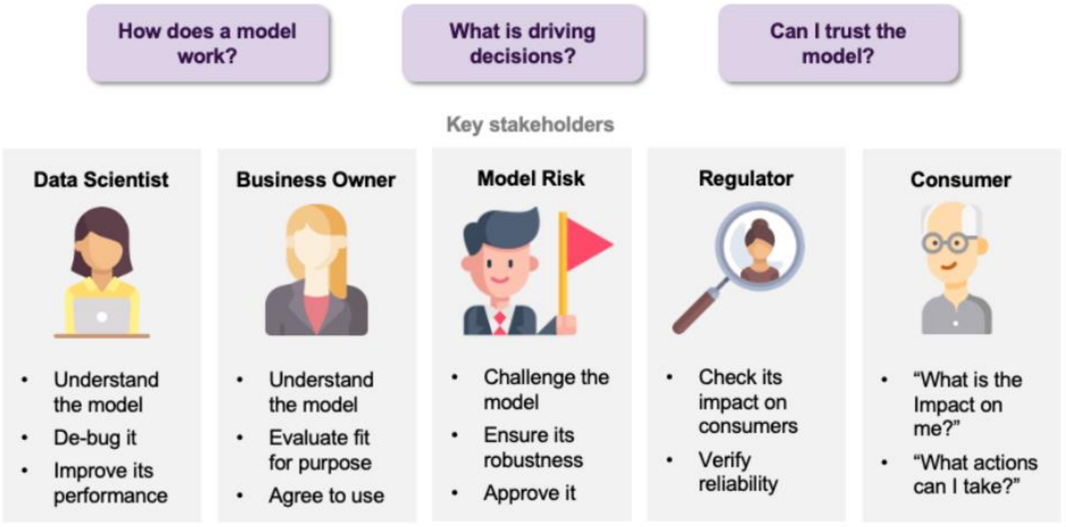
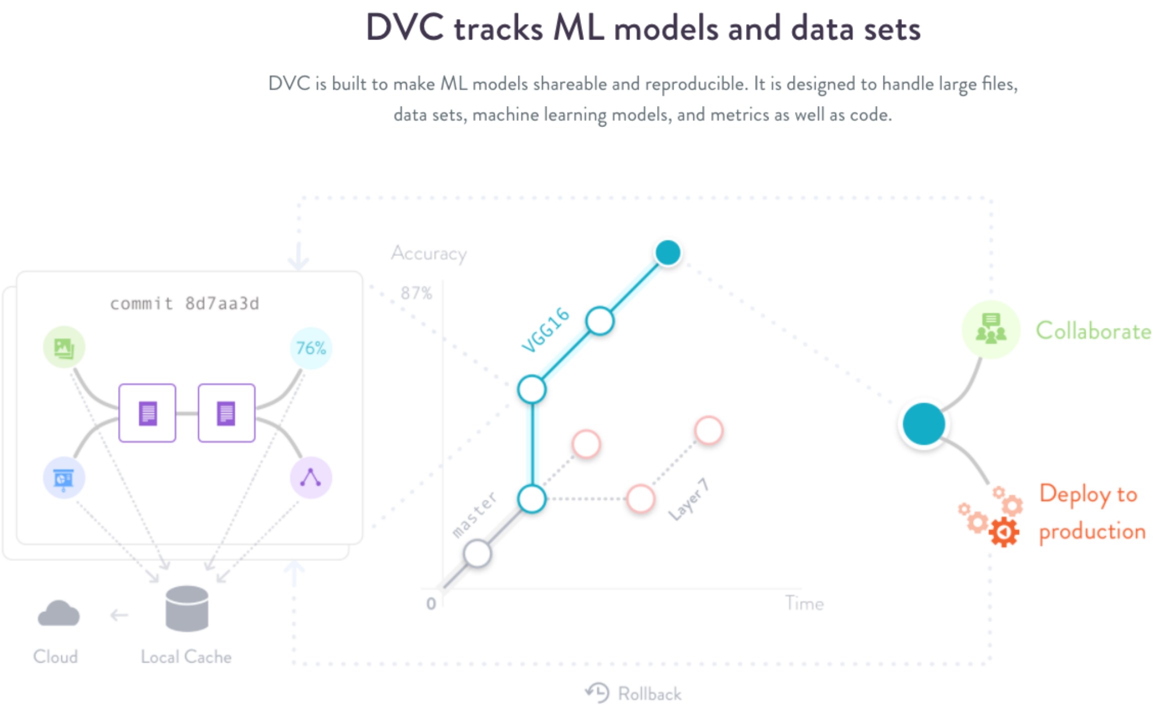


Figure 2 Concerns faced by various stakeholders (Belle and Papantonis 2020)

3.2 Reproducibility

When it comes to reproducibility a Nature survey found that more than 70% of researchers have tried and failed to reproduce another scientist's experiments, and more than half have failed to reproduce their own experiments. We believe making sure our research is reproducible is a cornerstone in making sure it is understandable. As stated in (The Royal Society 2019), data quality and provenance are part of the explainability pipeline.

This will be closely aligned with ongoing work in WP2 (Stakeholders identification, use cases definition, requirements specification, and architecture design), ensuring that the architecture being developed considers the full lifecycle of raw data to model output. There are a range of tools available that will be analysed in this context e.g., DVC to track models and data sets, TensorFlow Data Validation and Model Analysis to monitor the quality of deployed datasets and models, respectively.



DVC tracks ML models and data sets

DVC is built to make ML models shareable and reproducible. It is designed to handle large files, data sets, machine learning models, and metrics as well as code.

Figure 3 Tracking Models and Datasets[9]

3.3 Transparency

A human-level understanding of the inner workings of the model

Transparency is key because bias is embedded in our algorithmic world; it pervasively affects perceptions of gender, race, ethnicity, socioeconomic class, and sexual orientation. The impact can be profound, deciding for example who gets a job, how criminal justice proceeds, or whether a loan will be granted.

Bias from an estimator is classically defined as the difference between the expectation over the data and the true underlying value from the distribution (Goodfellow, Bengio and Courville 2016). In the context of fair and trustworthy AI, bias refers to the discrimination produced when some classes or results are more heavily weighted than others or the nature of the underlying distribution is poorly represented by the training sample.

There exist many different sources of introducing bias in an algorithm, from the dataset used in the training phase to unforeseen cases in model validation or extreme regularization. It is a significant risk in healthcare practice, for example, using of racial

[9] https://dvc.org/

categories as a proxy of genomics when dealing with personalized medicine (Bonham, Callier and Royal 2003) (Callier 2019) among other reasons such as historically under-represented populations. An infamous example of this kind of bias could be the application of Framingham's' risk score of any coronary heart disease event, fatal or non-fatal based on categories of age, sex, smoking status, total cholesterol and systolic blood pressure derived from the US population applied to the European populations. The studies carried out over European population evidenced that the reference score obtained from the US population overestimated absolute risk in populations with lower coronary heart disease rates (Conroy, et al. 2003).

Many different efforts have been done in identifying and classifying the bias in AI.

The mainly identify source are classified as (Barocas and Selbst 2016):

1. Skewed data: This bias comes from the source of acquisition.
2. Tainted data: The main cause of this bias comes from data modelling, categorization and labelling if the domain is not properly represented.
3. Limited features: It is a common practice to reduce dimensionality of input features of a model, but it has the counterpart of inducing some bias in the model, sometimes intentionally, for example, in the case of feature engineering.
4. Sample size disparities: The ideal dataset must provide balanced sets of the most sensitive features.
5. Proxy features: There may be correlated features with sensitive ones that can induce bias even when the sensitive features are not present in the dataset, an example of this bias could be the use of race categories previously commented.

Interpretability helps ensure impartiality in decision-making to correct from bias in the training dataset. Several research groups are currently focused on dealing with bias and fight against bad practices and biased datasets to avoid gender and racial discrimination. IBMs' AI Fairness 360 toolkit provides metrics to detect and remove bias in datasets and models that could be hidden of overseen unintentionally. Microsoft's Fairlearn is another open-source toolkit which provides interactive visualizations and bias mitigation algorithms to explore datasets prior to their use. Another example, could be the "Teach and Test" methodology framework from Accenture, which aims at helping decision making to overcome bias or other risks mainly focused on financial environments.

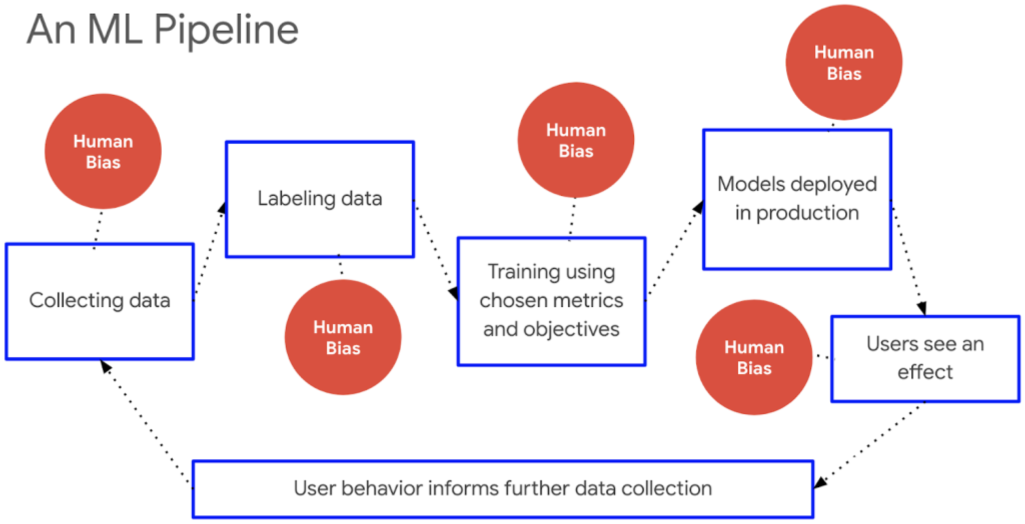

An ML Pipeline
Human Bias
Collecting data
Labeling data
Human Bias
Human Bias
Training using chosen metrics and objectives
Human Bias
Models deployed in production
Human Bias
Users see an effect
User behavior informs further data collection

Figure 4 Potential Sources of Bias[10]

There are various types of transparency in the context of human interpretability of algorithmic systems and several types and goals of transparency have been defined (Weller 2017) some of which have direct implications for FAITH (and healthcare AI in general):

1. For a developer, to understand how their system is working, aiming to debug or improve it: to see what is working well or badly, and get a sense of why.
2. For a user, to provide a sense for what the system is doing and why, to enable prediction of what it might do in unforeseen circumstances and build a sense of trust in the technology.
3. For society, broadly to understand and become comfortable with the strengths and limitations of the system, overcoming a reasonable fear of the unknown.
4. For a user to understand why one particular prediction or decision was reached, to allow a check that the system worked appropriately and to enable meaningful challenge.
5. To provide an expert (perhaps a regulator) with the ability to audit a prediction or decision trail in detail, particularly if something goes wrong.
6. To facilitate monitoring and testing for safety standards.
7. To make a user (the audience) feel comfortable with a prediction or decision so that they keep using the system.

[10] https://blog.tensorflow.org/2019/12/fairness-indicators-fair-ML-systems.html

We intend to follow the examples of (Weller 2017) and (Doshi-Velez and Kim 2017) by striving for *global* interpretability (a general understanding of how an overall system works, as in the transparency types 2-3) and *local* interpretability (an explanation of a particular prediction or decision, as in types 4, 5, and 7.

Informally, transparency is the opposite of opacity or blackbox-ness. It connotes some sense of understanding the mechanism by which the model works. Transparency as interpretability refers to the model's properties that are useful to understand and can be known before the training begins. (Lipton 2018) consider transparency at the level of the entire model (simulatability), at the level of individual components (e.g., parameters) (decomposability), and at the level of the training algorithm (algorithmic transparency).

3.3.1 Simulatability

Can a human walk through the model's steps?

This property addresses whether or not a human could go through each step of the algorithm and check if each step is reasonable to them. Could they predict its behaviour on new inputs? Linear models and decision trees are often cited as interpretable models using such justifications; the computation they require is simple, and it is relatively easy to interpret each of the steps executed when a prediction is made. This, however, isn't a guarantee. A decision tree with ten nodes is easy to interpret but make that ten-thousand nodes and understanding may become a challenge.

3.3.2 Decomposability

Is the model interpretable at every step or with regards to its sub-components?

Can a model be broken down into parts e.g., input, parameters, and computations, and can these parts then be explained.

3.3.3 Algorithmic Transparency

Does the algorithm confer any guarantees?

Understanding the procedure the model goes through in order to generate its output.

This question asks if our learning algorithm has any desirable properties which are easy to understand. For example, we might know that the algorithm only outputs sparse models, or perhaps it always converges to a specific type of solution. In these cases, the resulting learning model can be more amenable to analysis.[11]

[11] https://thegradient.pub/interpretability-in-ml-a-broad-overview/

3.4 Post-Hoc Interpretability

(Lipton 2018) pose four questions on post-hoc interpretability, which refers to things we can learn from the model after training has finished.

- Text Explanations:
  - Can the model explain its decision in natural language, after the fact?
- Visualisation:
  - Generating visualisations that facilitate the understanding of a model.
- Local Explanations:
  - Can the model identify what is/was important to its decision-making?
- Two properties that any good feature attribution method should follow[12]:
  - Consistency. Whenever we change a model such that it relies more on a feature, then the attributed importance for that feature should not decrease.
  - Accuracy. The sum of all the feature importances should sum up to the total importance of the model.
- Explanation by Example:
  - Can the model show what else in the training data it thinks are related to this input/output?

3.5 Model Uncertainty

When applying artificial intelligence algorithms to clinical practice it is important to provide clinicians with values that they can use as a reference for better informed decision making. A reliable system must accompany its predictions with a measure of uncertainty based on the premise that there is no such thing as a perfect system. Therefore, looking at uncertainty provides robustness to a system by allowing it to assess, for example, whether the system in question is basing its predictions on characteristics that can be considered artefacts. For example, the study by (Zech, et al. 2018) used an uncertainty measure to check how the accuracy of their x-ray imaging algorithm would lose precision if no brightness artifacts caused by the radiation shielding were seen. (Begoli, Bhattacharya and Kusnezov 2019) highlighted the importance of uncertainty quantification in guiding decisions based on Deep Learning algorithms. To help bound the overall confidence in predictions of medical applications epistemic uncertainty is usually determined using Bayesian neural networks; in theory, this uncertainty can be modelled (Natekar, Kori and Krishnamurthi 2020). However, a more practical and computationally simple approach is to approximate this Bayesian inference; it is typically performed by using dropout layers while testing the models, known as Test Timed Dropout (Gal and Ghahramani 2016). The use of uncertainty maps is another practice to provide better explainability of the models.

[12] https://towardsdatascience.com/interpretable-machine-learning-with-xgboost-9ec80d148d27

3.6 Pitfalls to Avoid

(Molnar, et al. 2020) provides a useful summary of ML model interpretation pitfalls. There reasoning is that a number of techniques can actually lead to wrong conclusions if applied incorrectly. ML models usually contain non-linear effects and higher-order interactions. Therefore, lower-dimensional or linear approximations can be inappropriate and misleading masking effects can occur. As interpretations are based on simplifying assumptions, the associated conclusions are only valid if we have checked that the assumptions underlying our simplifications are not substantially violated. In classical statistics this process is called "model diagnostics" (Fahrmeir, et al. 2013) and (Molnar, et al. 2020) argue that a similar process is necessary for interpretable machine learning based techniques.

The pitfalls are:

- Bad model generalisation
- Unnecessary Use of Complex Models
- Ignoring Feature Dependence
    - Interpretation with Extrapolation
    - Confusing Correlation with Dependence
    - Misunderstanding Conditional Interpretation
- Misleading Effect due to Interactions
- Ignoring Estimation Uncertainty
- Ignoring Multiple Comparisons
- Unjustified Causal Interpretation

4 CURRENT TOOLS AND TECHNIQUES

While Section 3 gave an overview of Explainable AI and related topics, in this section we discuss current tools and techniques that are used to facilitate Explainable AI.

As can be seen in Figure 5 there are many techniques that have become popular in an attempt to make the models more explainable. In the same way, tools have emerged that allow them to be integrated both in the model creation phase and in the validation phase so that they can contribute to the current machine learning frameworks.

Regarding those tools, the most popular ones are those that refer to the training phase. In this category we can find tools for the implementation of techniques such as Layer-wise Relevance Propagation (LRP), explainable embeddings and Integrated Gradients (IG). These techniques seek to understand the importance of the characteristics, identify the deviation of the data and debug the performance of the model.

4.1 TF-Explain

TF-Explain[13] is the most common in practice library. TF-Explain provide tools for models trained in TensorFlow which are also compatible with TensorBoard which is useful since the smooth integration with TFs' visualization interface/kit provides quick hands on with this framework library for local interpretability.

The principal tools provided by this library are based on:

Activations Visualization: More oriented to vision algorithms and convolutional layers, this component allows the visualization of the activations of hidden layers in order to visually interpretate the activations given a particular input in the network.

Grad CAM: As an extension of the previous tool, it allows to visualize how parts of an image affects the output of the network, so it allows to provide some interpretation on which parts of the image are activating a classifier.

Occlusion Sensitivity: This tool allows to study how partial occlusions of the image or input features, in the similar way that a dropout layer is applied, affects the confidence in predictions.

SmoothGrad: allows the visualization of sensitivity maps which are based on gradients. This simple technique aims at explaining the inputs that triggers a decision by adding and removing noise through a sensitivity analysis.[14]

Integrate Gradients: This is the most widely known techniques and it is already included in frameworks such as Keras. This approach integrates some call to the standard gradient operator and relies on an attribution methodology to assess sensitivity and implementation invariance to extract rules from them.[15]

[13] https://pypi.org/project/tf-explain/

[14] https://arxiv.org/abs/1706.03825

[15] https://arxiv.org/pdf/1703.01365.pdf

4.2 Captum

The analogous library **Captum**[16] provides the same tools for the Pytorch framework covering a set of gradient-based and perturbation-based attribution algorithms. Figure 5 shows the attribution algorithms provided by this library.

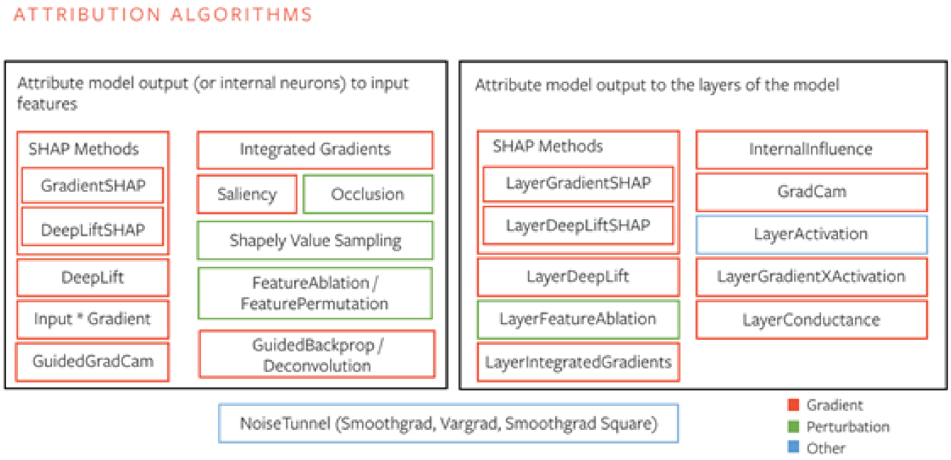


Figure 5 Sample Attribution Algorithms from Captum[17]

4.3 What-If

What-If Tool[18] integrated in the aforementioned TensorBoard from Tensorflow provides validation tools to test if a model follows some fairness constrains. This is a very visual tool oriented to analyze the behavior of the algorithms.

4.4 SHAP

SHAP (SHapley Additive exPlanations)[19] is developed by Scott Lundberg at the University of Washington. It is a unified framework for interpreting predictions. It assigns each feature an importance value for a particular prediction.

SHAP computes Shapley values from game theory, by assuming that each feature value of the instance is a “player” in a game where the prediction is the payout. Then a prediction can be explained by computing the contribution of each feature to the prediction. Note SHAP has these desirable properties:

[16] https://captum.ai

[17] *https://medium.com/pytorch/introduction-to-captum-a-model-interpretability-library-for-pytorch-d236592d8afa*

[18] https://pair-code.github.io/what-if-tool/

[19] https://github.com/slundberg/shap

1. **Local accuracy**: the sum of the feature attributions is equal to the output of the model we are trying to explain

2. **Missingness**: features that are already missing have no impact

3. **Consistency**: changing a model so a feature has a larger impact on the model will never decrease the attribution assigned to that feature.

SHAP supports tree ensemble, deep learning and other models. It can be used for both global and local explanation and it can be integrated with other techniques based on their modules: DeepExplainer, GradientExplainer and KernelExplainer.

In the example in Figure 6 each customer has one dot on each row. The x position of the dot is the impact of that feature on the model's prediction for the customer, and the colour of the dot represents the value of that feature for the customer. Dots that don't fit on the row pile up to show density (there are 32,561 customers in this example).

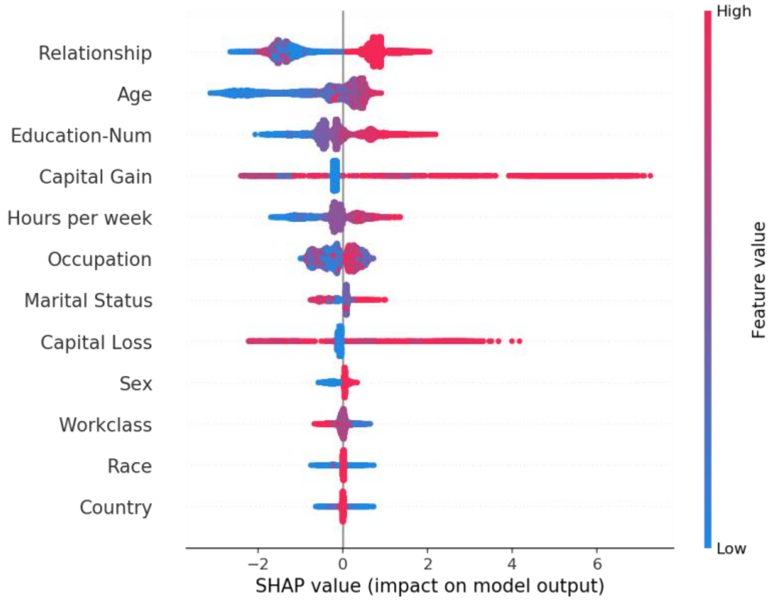


Figure 6 SHAP Example Plot

A similar approach to SHAP is also taken by InterpretML, with a comparison of their outputs shown in Figure 7**Error! Reference source not found.**

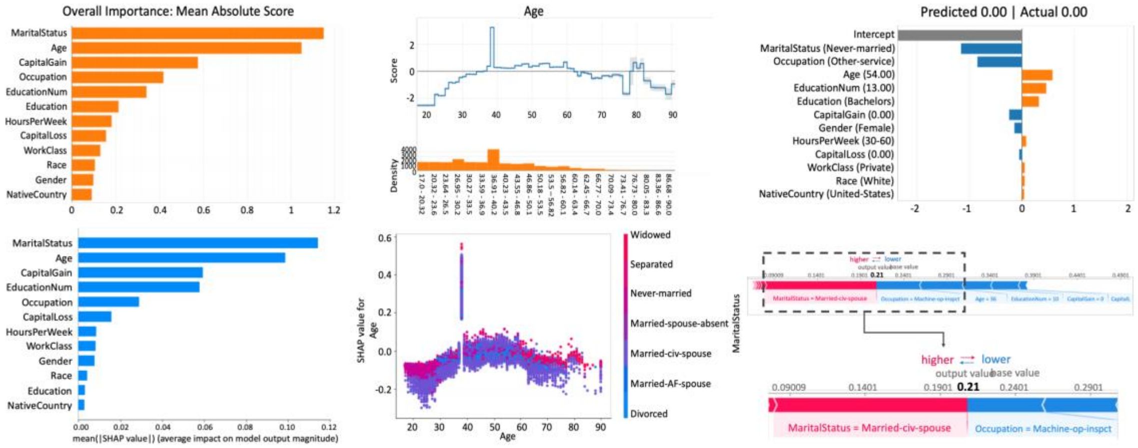


Figure 7 Comparison of InterpretML and SHAP Plots

4.5 LIME

Local Interpretable Model-Agnostic Explanations (LIME)[20] is based on the concept of surrogate models. When interpreting a black box model, LIME tests what happens to the predictions with variations of data, and trains local surrogate models with weighted features. Finally, individual predictions for "black box" models can be explained with local, interpretable, surrogate models.

The LIME library provides the code to apply it to categorical, text and image datasets. It provides native support for SciKit Learn and can be easily integrated in TensorFlow through their text and tabular explainer classes or Pytorch through the Captum library.

4.6 Anchors

High-precision model-agnostic explanations. A method for learning rule lists that predict model behaviour with high confidence.[21]

[20] https://arxiv.org/abs/1602.04938

[21] https://homes.cs.washington.edu/~marcotcr/aaai18.pdf

4.7 TreeExplainer

Tree-based machine learning models are among the most popular non-linear predictive learning models in use today, with applications in a variety of domains. TreeExplainer provides a novel set of tools rooted in game theory that enables exact computation of optimal local explanations for tree-based models.

It is the first tractable method capable of quantifying an input feature's local importance to an individual prediction while simultaneously measuring the effect of interactions among multiple features using exact fair allocation rules from game theory.[22] It produces local explanations by assigning a numeric measure of credit to each input feature, such as factors that contribute to mortality risk as shown in Figure 8.

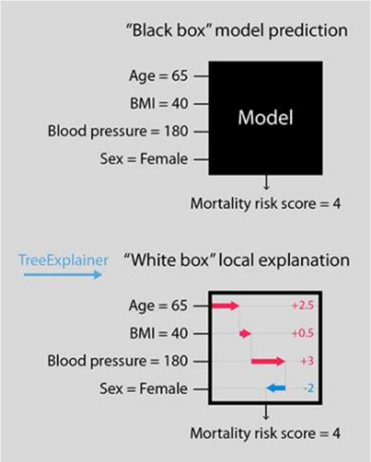
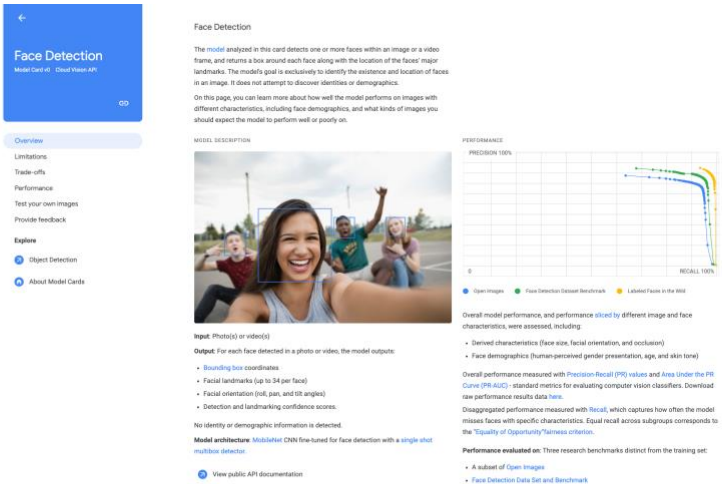


Figure 8 TreeExplainer Example

4.8 Model Cards

Model cards are short documents accompanying trained machine learning models that provide benchmarked evaluation in a variety of conditions that are relevant to the intended application domains. Model cards also disclose the context in which models are intended to be used, details of the performance evaluation procedures, and other relevant information.

[22] https://news.cs.washington.edu/2020/01/17/seeing-the-forest-for-the-trees-uw-team-advances-explainable-ai-for-popular-machine-learning-models-used-to-predict-human-disease-and-mortality-risks/



Face Detection

The model analyzed in this card detects one or more faces within an image or a video frame, and returns a box around each face along with the location of the faces' major landmarks. The model's goal is exclusively to identify the existence and location of faces in an image. It does not attempt to discover identities or demographics.

On this page, you can learn more about how well the model performs on images with different characteristics, including face demographics, and what kinds of images you should expect the model to perform well or poorly on.

MODEL DESCRIPTION

Input: Photo(s) or video(s)

Output: For each face detected in a photo or video, the model outputs:

- Bounding box coordinates
- Facial landmarks (up to 34 per face)
- Facial orientation (roll, pan, and tilt angles)
- Detection and landmarking confidence scores.

No identity or demographic information is detected.

Model architecture: MobileNet CNN fine-tuned for face detection with a single shot multibox detector.

Overall model performance, and performance sliced by different image and face characteristics, were assessed, including:

- Derived characteristics (face size, facial orientation, and occlusion)
- Face demographics (human-perceived gender presentation, age, and skin tone)

Overall performance measured with Precision-Recall (PR) values and Area Under the PR Curve (PR-AUC) - standard metrics for evaluating computer vision classifiers. Download raw performance results data here.

Disaggregated performance measured with Recall, which captures how often the model misses faces with specific characteristics. Equal recall across subgroups corresponds to the "Equality of Opportunity"fairness criterion.

Performance evaluated on: Three research benchmarks distinct from the training set:

- A subset of Open Images
- Face Detection Data Set and Benchmark

Figure 9 Google Model Card Example

5 CONCLUSIONS

This deliverable serves as an overview of the most interesting trends in the field of Explainable AI. What is clear from the range of research discussed is that there is no approach suitable for each and every scenario. This is in tune with how humans perceive explainability as well, since we know that there is not a single question whose answer would be able to communicate all the information needed to explain any situation (Belle and Papantonis 2020).

As the clinical study protocol is finalised, and the FAITH platform architecture formalised, we now have a clear vision of where this task needs to go, and the areas we must tackle e.g.

- Find better ways to formalise what we mean by interpretability, and through consultations with our clinical partners understand what explanations are desired. We can then design the architecture of the learning method(s) to give results that pertain to these explanations.
- Ensuring that these interpretability approaches are actually providing value i.e., in the context of the project do they offer anything useful to patients and doctors. This point is important and maybe not so intuitive. A large study from Microsoft Research (Poursabzi-Sangdeh, et al. n.d.) found that there was no significant difference between a transparent model with few features and a black-box model with many features in terms of how closely participants followed the model's predictions.
- Defining what is important for FAITH, considering things such as the question of decreased performance and adoption
    - It is clear that black box models dominate in terms of results for many areas. Any additional work to induce a more interpretable model or derive a post-hoc explanation brings an additional cost. At this point, all the approaches towards improving model interpretability we have seen either increase training/processing time, reduce accuracy, or do some combination of both.
    - In applications where explainability is of utmost importance, it is worth considering using a transparent mode. The downside of this, is that these models often compromise performance for the sake of explainability, so it is possible that the resulting accuracy hinders their employment in crucial real-world application, so we need to figure out where FAITH lies on this spectrum.
- At this point there is no established way of combining techniques (in a pipeline fashion) to produce a more complete explanation, so there is room for experimenting and adjusting them, according to the explanation at hand (Belle and Papantonis 2020). The question for us is not simply whether our ML is explainable, or whether one model is more explainable than other, but whether the FAITH system can provide the type of explainability that is necessary for our specific tasks and user groups. We believe there is real opportunity here for FAITH, particularly through our involvement in the Health & Care Cluster Working Group.

6 Bibliography

Aquasmart Project. 2015. *Redmine.* Πρόσβαση March 19th, 2015. https://www.aquasmartdata.eu/redmine.

Barocas, S., και A. D. Selbst. 2016. «Big data's disparate impact.» *Calif. L. Rev.*

Beck, Kent et al. 2001. *Manifesto for Agile Software Development.* Πρόσβαση March 18th, 2015. http://agilemanifesto.org/.

Begoli, E., T. Bhattacharya, και D. Kusnezov. 2019. «The need for uncertainty quantification in machine-assisted medical decision making.» *Nature Machine Intelligence* 20-23.

Belle, Vaishak, και Ioannis Papantonis. 2020. «Principles and Practice of Explainable Machine Learning.» *https://arxiv.org/pdf/2009.11698.pdf.*

Bonham, V. L., S. L. Callier, και C. D. Royal. 2003. «Will precision medicine move us beyond race?» *The New England Journal of Medicine* 374(21).

Callier, S. L. 2019. «The use of racial categories in precision medicine research.» *Ethnicity & Disease* 651.

Conroy, R. M., K. Pyorala, A. P. Fitzgerald, S. Sans, A. Menotti, G. De Backer, D. De Bacquer, και συν. 2003. «Estimation of ten-year risk of fatal cardiovascular disease in Europe.» *European Heart Journal* 987-1003.

Doran, Derek, Sarah Schulz, και Tarek R Besold. 2017. «What Does Explainable AI Really Mean? A New Conceptualization of Perspectives.» *arXiv.*

Doshi-Velez, Finale, και Been Kim. 2017. «Towards A Rigorous Science of Interpretable Machine Learning.» *arXiv.*

Fahrmeir, L., T. Kneib, S. Lang, και B. Marx. 2013. *Regression: Models, Methods and Applications.* Berlin: Springer-Verlag.

Gal, Y., και Z. Ghahramani. 2016. «Dropout as a bayesian approximation: Representing model uncertainty in deep learning.» *ICML.* 1050-1059.

Git contributors. 2015. *git --distributed-is-the-new-centralized.* Πρόσβαση March 19th, 2015. http://git-scm.com/.

Goodfellow, I., Y. Bengio, και A. Courville. 2016. *Deep Learning.* MIT PRess.

Goodman, Bryce, και Seth Flaxman. 2017. «European Union Regulations on Algorithmic Decision-Making and a "Right to Explanation".» *AI MAgazine.*

Lipton, Zachary C. 2018. «The Mythos of Model Interpretability: In machine learning, the concept of interpretability is both important and slippery.» *ACM Queue*, June.

Lundberg, Scott M., και Su-In Lee. 2017. «A Unified Approach to Interpreting Model Predictions.» *https://arxiv.org/pdf/1705.07874.pdf.*

McLaughlin, John. 2015. *Creating User Stories.* TSSG. 19th March. Πρόσβαση March 19th, 2015. https://www.aquasmartdata.eu/redmine/projects/aquasmart/wiki/UserStory.

—. 2015. *How to Use Git.* Aquasmart Project. 19th March. Πρόσβαση March 19th, 2015. https://www.aquasmartdata.eu/redmine/projects/aquasmart/wiki/UsingGit.

Miller, Tim. 2018. «Explanation in Artificial Intelligence: Insights from the Social Sciences.» *https://arxiv.org/pdf/1706.07269.pdf.*

Molnar, Christoph, Gunnar Konig, Julia Herbinger, Timo Freiesleben, Susanne Dandl, Christian A. Scholbeck, Giuseppe Casalicchio, Moritz Grosse-Wentrup, και Bernd Bischl. 2020. «Pitfalls to Avoid when Interpreting Machine Learning Models.» *https://arxiv.org/abs/2007.04131.*

Natekar, P., A. Kori, και G. Krishnamurthi. 2020. «Demystifying Brain Tumor Segmentation Networks: Interpretability and Uncertainty Analysis.» *Frontiers in Computational Neuroscience.*

Poursabzi-Sangdeh, Forough, Daniel G. Goldstein, Jake M. Hofman, Jennifer Wortman Vaughan, και Hanna Wallach. χ.χ. «Manipulating and Measuring Model Interpretability.» https://arxiv.org/pdf/1802.07810.pdf.

The Royal Society. 2019. «Explainable AI: the basics.» Policy briefing.

Weller, Adrian. 2017. «Challenges for Transparency.» *arXiv.*

Zech, J. R., M. A. Badgeley, M. Liu, A. B. Costa, J. J. Titano, και E. K. Oermann. 2018. «Variable generalization performance of a deep learning model to detect pneumonia in chest radiographs: a cross-sectional study.» *PLoS Medicine.*