**Module 2: Explainability (2 lessons)**

A closely related value to transparency is explainability. Explainable AI (XAI), for example, is artificial intelligence which outputs can be understood by humans.

a. Examining the relation between transparency and explainability;

As discussed in the previous module, there are reasonable arguments to make on both sides that transparency is intrinsically and instrumentally valuable. Whichever category it falls under, it is stands close to another responsible AI value, namely explainability.

The goal of Explainable AI (XAI) is to ensure that users of AI tools come to understand algorithmic output in a meaningful and usable way. Explanation is not a one-way street. It requires interactivity between the explainer and the explainee. In order to achieve these goals of explainability, the purpose of the AI tool, its scope of use and its limitations must be transparent to the user.

The two concepts are linked closely enough for the OECD to include them in a single principle, Principle 1.3 Transparency and explainability.

They state that transparency carries multiple meanings:

* 1. Disclosure of when an AI is being used
  2. Transparency further means enabling people to understand how an AI system is developed, trained, operates, and deployed in the relevant application domain, so that consumers, for example, can make more informed choices.
  3. Transparency also refers to the ability to provide meaningful information and clarity about what information is provided and why.
  4. An additional aspect of transparency concerns facilitating public, multi-stakeholder discourse and the establishment of dedicated entities, as necessary, to foster general awareness and understanding of AI systems and increase acceptance and trust.

OECD defines explainability as:

Explainability means enabling people affected by the outcome of an AI system to understand how it was arrived at. This entails providing easy-to-understand information to people affected by an AI system’s outcome that can enable those adversely affected to challenge the outcome, notably – to the extent practicable – the factors and logic that led to an outcome.

However, these terms are not always used uniformly across all disciplines. Oftentimes, transparent AI is used to mean AI models which are self-evident in how they function. Explanations are then reserved for those algorithms that are not opaque. For example, black-box models require explanation because it is not clear from the outset how they function and how they arrive at their outputs.

One takeaway to keep in mind is that many of the values that constitute responsible AI mutually inform or even occasionally overlap with one another. This intertwining of values means that it is possible for conflicts among values to occur and it also means that identifying and understanding a particular value can be complex.

Reading: <https://oecd.ai/en/ai-principles>

b. Discussion on the difference between backend explainability and frontend explainability;

As mentioned previously, we might divide XAI into two main questions: How does the algorithm work? And, why, as an end user, is the algorithm’s output interesting for me?

Of course, there are many sub-questions that could fall under these questions, and there may be other questions that have to be asked (e.g. stakeholders might want to know how they can report an error), but these are principal questions to begin thinking about XAI.

These two questions correspond to different “ends” of an algorithm. The how question corresponds to the backend and targets data scientists as its principal audience. The why question corresponds to frontend explainability and targets end users and its principal audience.

Backend explainability answers the question of how the algorithm works. As sub-questions, we might ask: what kind of machine learning was used, is the logic reasonable, etc. The information developers need to provide to answer the how question include: data sources, where the data is stored methods of data exploration, labelling, cleaning, preprocessing, engineering and re-sampling, how the model runs, how it computes the data, where the results are stored. It Is important to measure the models performance through its precision and recall metrics and to document and communicate these metrics to end users.

The frontend component is where end users interact with the tool. It is not completely divorced from backened explainability and some of the explanations corresponding to the backend clearly inform the frontend as well. In this interface, it is important to include both global and local explanations. Global explanations refer to how or why the algorithm as a whole functions, whereas local explamnations focus on specific outcomes produced by the model and why the end user sees a particular outcome and not something else.

The goal of this module is not to focus on the technical aspects of these questions—that is covered in other modules—but the sociotechical aspects. That is, why are these different kinds of explainability important to the acceptability, uptake, use of an AI tools and for establishing it as trustworthy.

Reading: <https://graphite.page/explainable-ai-report/#read-full-article>

c. Explain different interpretations of explainability (the process of understanding vs the quality of understanding);

Explainability features aim to facilitate some level of understanding in the targeted audience. After all, we explain things to others because we want them to understand something or because they want to understand something. This objective immediately raises the questions of what does it mean to understand something and how can explainability features achieve this, or these kinds of, understanding.

Firstly, it is worth noting that understanding and knowing are not the same things. I can come to know things quite easily. For example, I can know that it is raining outside simply by opening my eyes or stepping outside and feeling the drops of water land on my skin. Furthermore, I can know that it is raining outside without knowing much about other things. I may have no knowledge about the water cycle, the saturation of moisture in clouds, or even the law of gravity(!) and yet I can know that it is raining outside.

Understanding seems to be different than knowing in several respects. It seems to involve more complexity and therefore more of an effort, and hence accomplishment, on the part of the individual who understands. Relatedly, understanding something also seems to be more interrelated to other facts. To continue with our example, if I understand that it is raining outside, this statement does seem to include an understanding of the water cycle, cloud saturation, etc.

This distinction is further reflected in how we ask questions about whether one knows something or understands something. One can know that the Glorious Revolution refers to the 1688 deposition of King James II and his replacement by Mary II and her husband William III of Orange. This is the kind of knowledge one needs to succeed at trivial pursuit or a tv quiz show. However, we can immediately see that this knowledge is different from understanding the interrelated and complex facts that led to the occurrence of the Glorious Revolution. Understanding this is both more complex and, hence, requires some effort or achievement on the part of the individual.

Thinking about understanding in this way, we might say that the objects of understanding are the interrelations of things. So in order to understand and not simply know why it is raining outside, it is necessary that I have interrelated information about the water cycle, cloud saturation and the law of gravity. It’s only insofar as I understand these connections or relations that I can be said to understand why it is raining outside.

Thinking of understanding in this way is essential to explainability. After all, if we want explainability features to facilitate understanding, and understanding focuses primarily on the interrelatedness of things or, to put it more simply, how things hang together, then explanations in and about AI tools ought to focus on how the different components of an AI “hang together”.

Even if this distinction between knowing and understanding is correct, we still don’t yet know (or understand 😊) how much interrelated information one must have to be said to understand. In other words, how do we know when an explanation is a good one and that the end users truly understand? We’ll return to this question in module (e).

Coming back to the idea that explanations aim for understanding, and understanding in distinction to knowing, requires a sense of how things hang together, or the interrelatedness of things, we want to know how XAI can strive to achieve this kind of understanding.

It is essential to note that AI tools are not technical artifacts that exist in isolation of human use. What point would there be in creating an AI tool that no one used! AI tools are created by humans, for humans. Even the decisions we take in deciding which tools to develop, manufacture and sell are based on human individual and societal interests and interactions. For example, developers make recommender systems because people watch tv and movies and purchase items online. This existing societal practice gives rise to the motivation to create an AI that enhances people’s experience. Facial recognition to unlock our mobile phones, autonomous vehicles and even a AI chess player show how human interests, activities and use influence the development and manufacture of AI tools.

But AI tools also influence human activities. The relationship is not linear but circular. We no longer have to spend time and movement shopping for things we like. This additional time can now be spent on other activities, some technical some not. For example, we might use our additional time to go hiking or we might spend it playing computer chess.

XAI ought to focus on explaining the interrelatedness of human activities and a given AI tool. It is this sociotech interrelatedness that is integral to successful XAI.

Pursuing this objective is not easy, and the characteristics of explanations will likely vary depending on context. It’s not difficult to observe that explanations for police end users can differ widely from those designed for health care providers. Owing to the variance in contexts in which XAI needs to be implemented, it is essential to identify the extant decision-making processes of the targeted end user. If you are designing and implementing XAI for police, then it is pivotal to understand the police decision-making processes, so that the algorithmic output can be explained in a manner that allows the police end user to take up and act upon the information and understanding they are gaining. In fact, UK police have a very strict decision-making process—called the National Decision Model—meaning that information that falls outside of this model risks being ignored or being incapable of uptake. Consequently, when designing and implementing XAI for UK police, the explanations must align with this extant process.

Again, we see that explanations are not isolated bits of information. They cannot be separated from the humans who want to achieve understanding based on the algorithmic output. In fact, the processes in which the end users engage in understanding must be embedded into the XAI work of the development team. To achieve this, it is recommended to establish an interdisciplinary team consisting of subject matter experts of the context of use to engage in dialogue with the data science and data engineering teams. Of course, ethicists also ought to be included on such teams to lend insight into the ethical aspects of XAI and how good XAI helps achieve ethical AI.

Reading: <https://qz.com/1123896/its-better-to-understand-something-than-to-know-it/>

**Part B.**

If you decided to sit down and read 5 articles about XA, you might encounter 5 different definitions of what it is. This lack of consistency can make it difficult to establish a coherent understanding with the concept of explainability, thereby making its implementation difficult.

Interpretability:

In some scholarly work, and even in some regulatory guidance documents, interpretability will be used interchangeably with explainability. Other times, these concepts are kept distinct. When they are kept distinct, it is often the case that interpretability will refer to what we called frontend explainability earlier in this module, and explainability will refer to what we called backend explainability. Using “interpretability” in this way is reasonable. The goal is to facilitate an interpretation of the algorithmic output on the part of the end users.

If the end users can “interpret” the output appropriately, then the goal has been achieved.

On the other hand, interpretation and explanation are very different things.

Reading: <https://arxiv.org/pdf/2108.03012.pdf>

Do we really even need XAI?:

The reason why we need to pursue XAI is because the algorithms we are using are not understandable without an explanation, or several explanations, of how they work and why their output is significant to end users. This inabaility to understand without explanations makes XAI a necessity.

But what if it were possible to use models that were directly understandable without explanation? Creating so-called “directly-interpretable, “transparent” or “white box” models would render XAI completely unnecessary. After all, we don’t need to explain what is already understood.

It is precisely this position that some scholars defend. For example [Cynthia Rudin](https://arxiv.org/pdf/1811.10154.pdf) argues that there is no need to use black box models such that even the data scientists and engineers who build and train the model can’t directly understand how it works. She argues that model developers will often select highly complex black box neural net approaches to AI tools, even when such approaches are unnecessary. Trying to explain black box models, rather than creating models that are interpretable in the first place, is likely to perpetuate bad practices and can potentially cause catastrophic harm to society. There is a way forward – it is to design models that are inherently interpretable.

Deep learning models, for instance, tend to be black boxes of the first kind because they are highly recursive. As the term is presently used in its most common form, an explanation is a separate model that is supposed to replicate most of the behavior of a black box (e.g., “the black box says that people who have been delinquent on current credit are more likely to default on a new loan”). Note that the term “explanation” here refers to an understanding of how a model works, as opposed to an explanation of how the world works.

She argues that using black box models is unnecessary because:

(i) It is a myth that there is necessarily a trade-off between accuracy and interpretability.

There is a widespread belief that more complex models are more accurate, meaning that a complicated black box is necessary for top predictive performance. However, this is often not true, particularly when the data are structured, with a good representation in terms of naturally meaningful features. When considering problems that have structured data with meaningful features, there is often no significant difference in performance between more complex classifiers (deep neural networks, boosted decision trees, random forests) and much simpler classifiers (logistic regression, decision lists) after preprocessing.

(ii) Explainable ML methods provide explanations that are not faithful to what the original model computes.

Explanations must be wrong. They cannot have perfect fidelity with respect to the original model. If the explanation was completely faithful to what the original model computes, the explanation would equal the original model, and one would not need the original model in the first place, only the explanation. (In other words, this is a case where the original model would be interpretable.) This leads to the danger that any explanation method for a black box model can be an inaccurate representation of the original model in parts of the feature space.

(iii) Explanations often do not make sense, or do not provide enough detail to understand what the black box is doing.

Even if both models are correct (the original black box is correct in its prediction and the explanation model is correct in its approximation of the black box’s prediction), it is possible that the explanation leaves out so much information that it makes no sense. I will give an example from image processing, for a low-stakes decision (not a high-stakes decision where explanations are needed, but where explanation methods are often demonstrated). Saliency maps are often considered to be explanatory. Saliency maps can be useful to determine what part of the image is being omitted by the classifier, but this leaves out all information about how relevant information is being used. Knowing where the network is looking within the image does not tell the user what it is doing with that part of the image.

(iv) Black box models are often not compatible with situations where information outside the database needs to be combined with a risk assessment.

In high stakes decisions, there are often considerations outside the database that need to be combined with a risk calculation. For instance, what if the circumstances of the crime are much worse than a generic assigned charge? There are often circumstances whose knowledge could either increase or decrease someone’s risk. But if the model is a black box, it is very difficult to manually calibrate how much this additional information should raise or lower the estimated risk. This issue arises constantly; for instance, the proprietary COMPAS model used in the U.S. Justice System for recidivism risk prediction does not depend on the seriousness of the current crime. Instead, the judge is instructed to somehow manually combine current crime with COMPAS. Actually, it is possible that many judges do not know this fact. If the model were transparent, the judge could see directly that the seriousness of the current crime is not being considered in the risk assessment.

(v) Black box models with explanations can lead to an overly complicated decision pathway that is ripe for human error.

Typographical errors seem to be common in computing COMPAS, and these typographical errors sometimes determine bail decision outcomes. This exemplifies an important drawback of using overly complicated black box models for recidivism prediction – they may be incorrectly calculated in practice. The computation of COMPAS requires 130+ factors. If typographical errors by humans entering these data into a survey occur at a rate of 1%, then more than 1 out of every 2 surveys on average will have at least one typographical error. The multitude of typographical errors has been argued to be a type of procedural unfairness, whereby two individuals who are identical might be randomly given different parole or bail decisions. These types of errors have the potential to reduce the in-practice accuracy of these complicated models.

On the other hand, the long-standing understanding in the discipline has been that black box models are required to obtain an acceptable level of accuracy. In the quest for more accurate AI, the availability of compute resources coupled with increasing dataset sizes have fueled a trend towards more complex non-linear models.

Indeed, we've seen a progression from handcrafted rules and heuristics, to linear models and

decision trees, ensembles and deep models to, most recently, meta-learning or models that create other models.

On the flip side, these more complex models have become increasingly opaque. This, coupled with the fact that these models are still fundamentally built around correlation and association, have resulted in several challenges:

Spurious correlations can be learned from the data, often hampering the model's ability to

generalize and leading to poor real world results.

- Loss of debuggability and transparency leading to low trust as well as the inability to fix or

improve the models and/or outcomes. Furthermore, this lack of transparency impedes

adoption of these models, especially in regulated industries e.g. Banking & Finance or

Healthcare.

- Proxy objectives resulting in large differences between how models perform offline, often

on matching proxy metrics, compared to how they perform when deployed in the

applications.

- Loss of control due to model practitioners' reduced ability to locally adjust model behavior

in problematic instances.

- Undesirable data amplification reflecting biases that don't agree with our societal norms

and principles.

These challenges seem to highlight the need for explainability in order to keep the humans in the loop and empower them to develop and leverage AI responsibly.

d. Identify cases of insufficient explainability and transparency so that they can be addressed;

Google’s Gmail keeps getting smarter and smarter, using AI, but it doesn’t tell users exactly how it identifies spam, sorts emails into our “promotions” folder, or why its suggested responses are so invariably enthusiastic and submissive (i.e. “Got it, thank you!” and “This is great, thanks!”).

Netflix recommendations: What characteristics does Netflix base it recommendations on? How precisely is the recommender algorithm calibrated? Can I alter my recommendations if I don’t like them?

e. Discussing approaches for evaluation of explanation. (What makes a good explanation?)

As covered above, two factors determine understandability of an ML system: the features

of the ML system and the human’s capacity for understanding.

To evaluate whether explanations have been successful, there are some metrics that can be applied:

Application-grounded evaluation (experiments with end-users). This kind of evaluation

requires conducting end-user experiments using the explanation within a

real-world application. It directly tests the objective that the system is built for in a

real-world application, and, thus, performance with respect to that objective gives

strong evidence of the success of explanations. An important baseline for this is

how well explanations assist in humans trying to complete tasks, such as decision making

Tasks.

  Human-grounded evaluation (experiments with lay humans). It refers to conducting

simpler human–subject experiments that maintain the essence of the target application.

Compared with the application-grounded evaluation, the experiments in this kind

of evaluation are not carried out with the domain experts but with lay humans,

allowing a big subject pool and less expenses for the evaluation. Ideally, this evaluation

approach will depend only on the quality of the explanation, regardless of the types

of explanations and the accuracy of the associated prediction.

Functionality-grounded evaluation (proxies based on a formal definition of interpretability).

This kind of evaluation does not require human experiments. In this type

of evaluation, some formal definition of interpretability serves as a proxy to evaluate

the explanation quality, e.g., the depth of a decision tree.

Application-grounded and human-grounded evaluations depend on the selected pool

of humans for various experimental tasks. In these evaluations, ML systems with explanations

are repeatedly updated and evaluated with humans. Within human experiments,

both qualitative and quantitative metrics can be used to evaluate explanation qualities.

Qualitative metrics include asking about the usefulness of, satisfaction with, confidence

in, and trust in provided explanations by means of interviews or questionnaires.

Quantitative metrics include measuring human-machine task performance in terms of accuracy, response time needed, likelihood to deviate, or ability to detect errors, and

even physiological responses from humans during experimental tasks.

Functionality grounded evaluation does not require human experiments and uses a formal definition of interpretability as a proxy for the explanation quality.

The advantages of application-grounded and human-grounded evaluations are that

they can provide direct and strong evidence of success of explanations. However,

these evaluations are usually expensive and time-consuming because of invitation of

humans and necessary approvals (e.g., Human Research Ethics Committees’ review) as

well as additional time for experimental conductions. Most importantly, these evaluations

are subjective. Comparatively, functionality-grounded evaluation can provide objective

quantitative metrics without human experiments.

Application-grounded and human-grounded evaluations use human experiments to

assess the effectiveness of ML explanations. Doshi-Velez and Kim mentioned four task- related factors for explanations:

  Global vs. local explanation. Global explanation means explaining why patterns are

present in general, while local explanation implies knowing the reasons for a specific

decision (such as why a particular prediction was made).

  Area and severity of incompleteness. This relates to incompletely specified inputs,

constraints, domains, internal model structure, costs, or even the need to understand

the training algorithm. The types of explanations needed may vary depending on

whether the source of concern is due to different incompleteness. The severity of the

incompleteness may also affect explanation needs.

  Time constraints. This refers to the time that the user can afford to spend to understand

the explanation. Different applications may have different times end-users can spend

on the explanation. Such constraints may affect approaches and effectiveness of

Explanations.

  Nature of user expertise. This is related to how experienced the user is in the task.

The nature of user expertise will affect the level of sophistication of explanation,

the organisation of explanation information, and others. For example, domain-users,

machine learning researchers, and layman users have different background knowledge

and communication styles.

These factors can impact the explanation goal of understandability and efficiency.

Each of these factors can be isolated in human-grounded experiments in simulated tasks to

determine which methods work best when they are presented.

In the current studies, there are at least two types of human studies on ML systems:

Studies using actual tasks to evaluate the performance of human and the system on

the ML-informed decision-making tasks. In these studies, participants are told to focus on making good decisions, but are flexible in terms of using ML to

accomplish tasks.

  Studies using proxy tasks to evaluate how well users are able to simulate the model’s

decisions or decision boundaries. In such studies, participants are specifically instructed to pay attention to the ML to evaluate human’s mental model of the ML system with the system’s predictions and explanations, but do not necessarily evaluate how well users are able to perform real decision-making tasks with the ML system.

Besides evaluation tasks for ML explanations, the choice of evaluation metrics plays a

critical role in the correct evaluation of ML systems. Two types of evaluation metrics can

be found in explainable ML research

  Subjective metrics. Subjective questionnaires are designed for users on tasks and explanations, and are asked during or after task time to obtain user’s subjective responses

on tasks and explanations. Examples of subjective metrics are user trust, confidence and preference, which have been largely embraced as the focal point for

the evaluation of explainable systems.

For example, Hoffman et al. presented metrics for explainable systems that are grounded in the subjective evaluation of a system (e.g., user trust, satisfaction, and understanding). Zhou, et al. investigated factors such as uncertainty and correlation that affect user confidence in ML-informed decision-making. Zhou et al. found that the explanation of influence of training data points significantly affected user trust in ML-informed decision-making. Yu et al. investigated the trust dynamics corresponding to ML performance changes. Holzinger et al. developed a scale named System Causability Scale to quickly determine whether and to what extent an explanation or an explanation process itself is suitable for the intended purpose.

  Objective metrics. This refers to objective information on tasks or humans before,

after, or during the task time. Examples include human metrics, such as physiological

and behaviour indicators of humans, during ML-informed decision-making, or task-related

metrics, such as task time length and task performance.

For example, Zhou et al. found that physiological signals such as Galvanic Skin

Response (GSR) and Blood Volume Pulse (BVP) showed significant differences with

the explanation presentation influencing training data points on ML predictions,

and these physiological responses can be used as indicators of user trust to assess the

quality of ML explanations. Based on approaches that quantify response times in user

studies as well as agreement between human labels and model predictions,

Schmidt and Biessmann proposed that faster and more accurate decisions indicate

intuitive understanding of explanations. A trust metric was then derived based on

these explainability metrics.

Functionality-Grounded Evaluation Metrics

The ML explanation methods can also be divided into three types :

  Model-based explanations. It refers to explanations that use a model to explain original task models. In this category, either the task model itself (e.g., decision tree) is used as an explanation or more interpretable models are generated to explain the task model.

  Attribution-based explanations. This kind of explanation ranks or measures the

explanatory power of input features and use this to explain the task model. For example, feature importance or influence based explanation approaches belong to this category.

  Example-based explanations. This kind of method explains the task model by selecting

instances from the training/testing dataset or creating new instances. For example, selecting instances that are well predicted or not well predicted by the model as

explanations, or creating counterfactual examples as explanations.

Reading: <https://hbr.org/2022/08/when-and-why-you-should-explain-how-your-ai-works>

Reijers, W., Wright, D., Brey, P., Weber, K., Rodrigues, R., O’Sullivan, D., & Gordijn, B. (2018). Methods for Practising Ethics in Research and Innovation: A Literature Review, Critical Analysis and Recommendations. *Science and Engineering Ethics*, *24*(5), 1437-1481. <https://doi.org/10.1007/s11948-017-9961-8>

Turilli, M., & Floridi, L. (2009). The ethics of information transparency. *Ethics and Information Technology*, *11*(2), 105-112. <https://doi.org/10.1007/s10676-009-9187-9>

Worthington, R. (1982). The Social Control of Technology. By David Collingridge. (New York: St. Martin's Press, 1980. Pp. i + 200. $22.50.). *American Political Science Review*, *76*(1), 134-135. <https://doi.org/10.2307/1960465>

Pre-reading:

<https://www.hiig.de/en/explaining-ai-explain-the-unexplainable/>

<https://arxiv.org/abs/1706.07269>

Reflection Sheet

1. Why is XAI considered to be valuable? Are there any disadvantages to doing XAI?
2. List and describe 3 of the technical methods described in the module for doing XAI.
3. Is explainability a subcategory of transparency or its own category? Explain your answer.
4. Provide 2 reasons from the module why XAI may not be necessary at all.

Multiple Choice questions:

1. What is one of the differences between knowing and understanding?
   1. Interrelatedness of the subject matter
   2. How convinced you are that you are correct
   3. If you are able to tell other people about it
2. Are interpretability and explainability the same thing?
   1. Yes
   2. No
   3. Sometimes
3. How can we best explain algorithms?
   1. By isolated bits of information
   2. Using generalisable models
   3. Aligning explanations with the decision processes of the end users
4. Are there metrics available for evaluating the quality of an explanation?
   1. Yes, and they are simple to apply
   2. No, we cannot know whether an explanation is good or bad.
   3. Yes, but the field is still new meaning we should be cautious regarding our conclusions.

Task:

A) Make notes to give a five-minute talk explaining to doctors the advantages and disadvantages of using XAI for their diagnostic tools, ending with practical steps they can take to ready themselves for using AI tools.

B) Read the following article and think about the example of the parolee at the top of the page. <https://itrexgroup.com/blog/explainable-ai-principles-classification-examples/>

How could this have been avoided?

Using the principles learned in this module, what could you have done from a technical perspective to help achieve a different outcome?

c) Group Exercise

From the previous article about the parolee in NY, imagine the company that built the AI used to adjudicate the man’s aprole is being taken to court. Appoint a third of the group to argue for the company, a third of the class to argue against, and the final third to act as the court, marshalling those arguments into a final judgment.

At the end, is the class satisfied with the judgment? Do they think that the judgment is realistic and would happen today? Is a company responsible for poor explainability even for examples it does not foresee?