Can XAI methods satisfy legal obligations of transparency, reason-giving and legal justification?

ELSA Deliverable 3.3

Short Report

28 August 2024

Madeleine Waller PhD Candidate, UKRI Safe and Trusted AI Centre for Doctoral Training, King's College London.

Paul Waller University of Bradford and Thorney Isle Research.

Karen Yeung, Interdisciplinary Professorial Fellow in Law, Ethics and Informatics, Birmingham Law School and School of Computer Science.

Aims and objectives

The purpose of this report is to explore how available technical methods can (or cannot) be integrated with, and embedded into, legal and ethical governance regimes to ensure that the design and deployment of algorithmic decision-making systems (including those which utilise AI) will serve legal, democratic and ethical values, focusing on legal obligations pertaining to transparency and accountability. It focuses on algorithmic decision-making (ADM) systems that are deployed by an organisation that produce an output which is intended to inform, or to automate, the making of a 'decision' that can result in the imposition of a substantive intervention that produces legal or other significant effects on the life of an affected person (a 'Decision'). Our analysis proceeds on the basis that in real-world practice, an ADM system is typically embedded within a larger socio-technical system and is executed via an 'organisational decision-making system architecture' that its members are expected to follow in carrying out their tasks and duties. Such an architecture typically identifies the formal chains of decision-making authority through which responsibility for carrying out designated tasks and duties are assigned.

Terminology

Misunderstandings concerning the extent to which 'explainable AI' (XAI) methods can address concerns about the opacity and concomitant accountability deficits pertaining to ADM systems are partly due to terminological confusion. We therefore begin by clarifying the meaning of the terms and terminology used throughout this Report. We refer to the computer system that results from the development and testing of an algorithm, together with input mechanisms (which may also include hardware components such as sensors, cameras, microphones and so forth to collect input data) and output mechanisms, processing functions to convert inputs to numbers and to translate those inputs into outputs, as a 'model'. Since our concern is with ADM systems for real-world use, a model purports to offer a representation of some aspect of the world. The output of a model is referred to as a prediction for the specific case being processed. A 'decision' refers to an action that has a concrete impact on the real-world, which is taken (to a greater or lesser extent) based on that prediction, although the significance of that impact may vary considerably. Some decisions may be of little consequence, such as decisions concerning the colour of an automatically distributed ad-banner. Our report focuses, however, on a subset of decisions that may result in the imposition of a substantive intervention that produces legal or other significant effects on the life of an individual person, which we refer to as a **‘**Decision’.

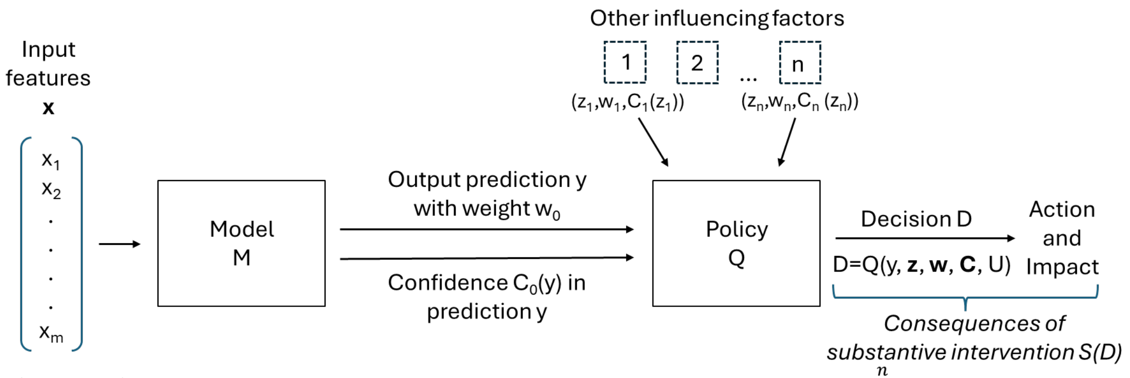
By separating the process by which the prediction is generated from a model, and the decision that is made based on that prediction, this facilitates the identification of any ‘explanations’ (or ‘functional reasons’) that may (or may not) be necessary to discharge one or more legal duties that may accompany the making of a Decision.   Interpretability is a property of an algorithm or model that reflects the degree to which a person can understand its workings, what each of its components means in real-world terms, and thus why it produces the outputs that it does. Interpretability is a product of the explainability of an algorithm, which refers to the degree to which functional reasons can be provided for either or both of the following:

- the general internal logic according to which the model converts inputs to outputs e.g. ‘It works like this…, and that is because…’ (a ‘global explanation’); and/or
- why a particular decision was arrived at, e.g. ‘in this case the following result is produced because…’ (a local explanation).

2. Organisational decision-making systems which incorporate ADMs

We construct an analytical framework which we call an **‘**organisational decision architecture’, to depict the various social, technical and organisational components and steps through which a socio-technical system produces a Decision.   Our framework is intended to facilitate identification of the many and varied choices that must be made within the larger sequence of events that occur within any given socio-technical system, particularly when an ADM contributes in some way, to enable the making of a Decision. This organisational decision-making architecture distinguishes:

- the ‘model’ into which observed data for the case in question is fed to generate an output (a prediction) influencing the decision (either in the form of a recommendation or binding determination);
- a ‘policy; that takes in all influencing factors (from models or other sources) and processes them to arrive at a ‘decision’, taking account of any weighting of the various factors, the level of confidence in them, and any weighting (utility) of any decision options to reach a final decision; and
- the ‘substantive intervention’ that directly flows from the decision, and its consequences for the person subject to that intervention and possibly others indirectly affected.



$\{x_1, x_2, \ldots x_m\}$ are the m input features for the case in question

$\{w_0, w_1, \ldots w_n\}$ are the normalised weights of the influencing factors y and $\{z_1, \ldots z_n\}$ such that $\sum_{i=0}^{n} w_i = 1$

The "confidence" C_i in an influencing factor may be expressed as a rating, a probability distribution, the variance of a continuous variable, or probabilities e.g. of a true positive and false positive of a binary classifier

U represents a utility function that may be applied across prediction confidence and/or options for decisions to determine D; S(D) is the substantive intervention arising from the decision, from a set of possibilities $\{s_1, s_2, \ldots s_d\}$

Explaining how the model produces its output (its prediction) is not the same as explaining how a decision is reached (the process of administering and executing the policy). Explanations about how a model produces its output could include an explanation of the logic according to which the internal working of the model functions and how a specific output was generated, but may also require explanations for one or more of the many choices made during model development, such as the selection and preparation of training data, selection of algorithm for the model, configuration of the model, any bias mitigations, choice of cut-off points and error thresholds and so forth.

3. Explanations, reasoned justification and legal justification

Explanations offer a response to the question 'what happened here'? They provide an account of what was done or how a particular outcome was produced. Different kinds of explanation are possible. For example, a functional explanation provides an account of the various chain of events that produced a specific outcome. In contrast, a motivational explanation concerns my subjective intentions in pursuing some action, providing an account of why I acted in a particular way.

It is important to distinguish a functional reason or explanation from a reasoned justification. When someone demands a justification for a particular decision concerning them, they wish to know why that outcome is deemed to be normatively acceptable by reference to some underlying theory of moral or social acceptability — 'yes, we judge that it was acceptable for you to do that to me'. Note, however, that both a functional explanation and a motivational explanation (underlying reasons and/or motives) may be needed to produce a justification for a given outcome, decision or action.

If a person is required to provide a legal justification for their decision, action or omission, this requires that they provide an account of identifying why that action (or inaction) was legally acceptable or permissible. It is important to recognise the nature and scope of legal rights and duties are deeply contextual and context-specific. Hence, in any given set of circumstances, the law may require explanations and justifications arising from either or both:

i. Specific transparency and disclosure obligations arising, for example, under contemporary data protection law, or are imposed on public authorities as a matter of constitutional principle and general administrative law as safeguards against the abuse of governmental power; and

ii. To identify whether any legal wrongs have been committed (which may include the commission of one or more criminal offences, torts, violations of public law duties, or unlawful interferences with the legal and other fundamental rights of others (such as employment rights and rights to non-discrimination) and, if so, to identify an appropriate legal remedy.

Whether any given functional explanation and/or reasoned justification is considered sufficient to discharge any applicable legal duties will depend upon the scope and content of the applicable laws and will depend on multiple contextual factors. To provide useful and legally relevant information, the form and content of any explanation, interpretation or justification will need to be tailored to its intended audience, addressing the context, relevance and interpretability of the algorithm being used.

4. What must be explained or justified?

Within an organisational decision-making architecture which assigns a role to an ADM system, there are a wide variety of matters which may call for an explanation and justification. This includes choices about each of the specific components of the organisational decision system, depending on the circumstances in which these demands for explanations or justifications are made, including demands to explain and legally justify the substantive Decision and its consequences for the affected person which those choices contributed to producing. If a specific Decision is challenged, then it may be necessary to explain any part of the ADM system in the organisational decision-making architecture separately, or the ADM system as a whole, including salient choices made at each point: the same applies to justifications. A particular focus for justifying a decision will be on the policy and the process by which it is implemented (e.g. human involvement or not), including the role of a model. It may be that weights, confidence and utility are omitted, implicit and/or subjective rather than precisely defined: this is very often the case in fact, and their omission could lead to difficulties when called upon to provide a comprehensive explanation and justification for any given decision in a legal setting.

4.1 Model development choices

The following is an indication of the sorts of issues about which choices may have been made in the design of a model and which may thus be interrogated for explanation or justification.

(a) Choice of type of algorithm:

Frequently, the most appropriate algorithm type is not evident at the start of a development process. A model developer is likely to experiment with several different algorithm types before settling on the one which generates the ‘best’ results — in which case an explanation for the choice may be no more meaningful than ‘it gave the best fit to our data of all the algorithms we tried’. Whether or not there can be a reasoned justification for a choice in such a situation without reference to the specific context of application and substantive intervention intended to flow from that output is questionable and demands further research, but in other circumstances and contexts (such as choosing an expert system when there is an established logical process) reasoned justifications may be available.

(b) Data-related choices

The most fundamental questions concern whether the data is appropriate for the decision to be made based on a sound relationship between the input features and the output prediction (termed ‘construct validity’) and whether the output prediction is relevant and appropriate in making the decision at hand. Similarly, it is necessary for any training data that its characteristics and its statistical properties are appropriate for representing a particular reality. This may need explaining and justifying, as well as the provenance and quality of the data chosen and how it was prepared for processing e.g. for machine learning there will be steps for selecting, cleaning, encoding, partitioning and ordering of the data.

(c) Development process choices

Any or all of the choices made throughout the ADM development process might be subject to a request for explanation, but the most significant, and worthy of justification, are the choice of optimisation objectives ('what is the model designed to predict as best it can?'), and the performance metrics concerning accuracy ('how do you measure how accurately the model is predicting the target variable?') and fairness ('to what extent do outputs produced by the model discriminate between affected stakeholder groups?') as these — along with construct validity. These shed light on the extent to which the model is 'fit' for its specific and intended purpose and context of use. Explanations for the other, lower-level technical, choices may illuminate how much the development process was trial and error, in contrast to theory and/or evidence-based.

(d) Characteristics of outputs

The outputs from models based on different types of algorithms will have different properties. The outputs from those based on statistical methods will have uncertainty attached to them, expressible (in theory, though sometimes hard to achieve in practice) in terms of probabilities. One characteristic that has attracted a large and growing body of technical research concerns the 'fairness' or 'bias' of the output predictions. Within this field, there are a wide range of mathematical definitions of fairness. In seeing to explaining the workings of models, the approach to defining 'fairness' and assuring 'fairness' should be explicitly addressed, and sources of potential bias identified and addressed, either in the mechanics of the model development process or in the training data, and explanations and justifications for addressing sources of bias in a particular way explicitly documented and included in the model's accompanying technical documentation.

4.2 Explanatory AI (XAI) methods

If the model is interpretable, functional explanations for its outputs can be produced without XAI techniques. But for models that rely on machine-learning 'black boxes', XAI techniques might have a role to play in contributing to the generation of functional explanations. There is a large range of possible XAI methods, and the effectiveness of each, and the suitability of any one to a particular model, is not a well-established field of knowledge. This report provides a brief account of the following XAI Methods:

- Local Intepretable Model-agnostic Explanations (LIME) for image classification and) for tabular data classification
- Shapley Additive exPlanations (SHAP) for tabular data classifications
- Counter-factual explanations

Our analysis proceeds on the assumption that existing XAI methods provide adequate explanations concerning the model's operation: we make no attempt to evaluate their technical accuracy, although we recognise that this assumption might not be realistic. In other words, we do not examine how 'good' the explanations are at representing the true relationships between the input data and outputs from a technical perspective, and instead focus on the potential usefulness of the explanations to meet legal demands including the need to provide a legal justification for a Decision based on that output.

Explanations and justifications may apply at many levels where black box algorithms are use, including the choice of XAI method, the application of the method, the results of the method, the interpretation of the results to produce a functional explanation for the model's output, for any generalisation from a specific case, and for an account of how the application of a given XAI method contributes to a reasoned explanation or justification for the use of the model.

5. Can XAI methods satisfy legal obligations of transparency, reason-giving and legal justification?

Typically, when computer scientists use the term 'reason' they mean only 'functional explanation' (or 'functional reason'), and similarly 'explanation', i.e. 'this is how that output was technically arrived at from the perspective of the underlying mathematical function'. Within the XAI community, discussions about the 'interpretability' of a model and the ability to 'explain' its workings concern the extent to which functional explanations can be provided for the outputs. We argue that functional explanations of this kind are different from the concept of 'reasoned justification' and it is important to distinguish them. When someone demands a justification for a particular decision concerning them, they wish to know why that outcome is deemed to be normatively acceptable by reference to some underlying theory of moral or social acceptability — 'yes, we judge that it was acceptable for you to do that to me'. Our analysis demonstrates that a functional explanation such as derived by XAI is seldom (if ever) sufficient to provide a justification for a decision based on a model output. For it to do so would assume that its underlying rules can be defended in terms of some underlying 'theory of social acceptability'.

For example, consider a driver who exceeds the legal speed limit: what reasons would be considered 'acceptable' in order to justify this behaviour? The construction of a theory of social acceptability in relation to any decision will invariably require normative judgement and cannot be derived solely from mathematics or logic. Moral or social justifications are not, however, coterminous with legal justifications. Even if the justification offered might be considered socially acceptable, the law might not allow for the behaviour in question to be justified or excused. In other words, whether a particular decision, action or omission is legally justifiable is a matter for the law and the legal system to determine, rather than that of moral judgement alone.

Where one or more algorithm-based models are involved in producing a Decision, reasons for, and justifying the use of, the chosen model, must be provided if called the organisation making that Decision is called upon to explain and justify it. This should entail:

- whether the model was fit for purpose and appropriate for the use made of it,
- the specific output generated by the model in the specific case at hand, and
- the technical choices made in its design, development and testing leading to assessments of accuracy of and confidence in predictions.

It should also include an explanation of, and justifications for, the non-technical aspects of the model and its development, including the governance and assurance processes around its development and deployment. We reiterate that explaining how the model produces its output (its prediction) is not the same as explaining how a decision is reached (the process of administering and executing the policy). Of particular importance is the need to justify the design and deployment of the model given the substantive intervention that is intended to be applied based on the algorithmic output generated by the model. For decisions that have rights-critical and safety-critical consequences, the quality, scope and rigour of the explanations and justifications required are at their most demanding. Thus, decisions about whether to forcibly remove a child from his or her adult carers on the basis that the child has been predicted as at 'high risk' of carer neglect and abuse will entail far more exacting standards of reasoning, evidence and justification than decisions to automatically distribute consumer product advertisements to users on-line while surfing the internet. Yet the literature concerned with the explainability of algorithms displays a troubling failure to attend to the substantive intervention that is administered from the decision which the algorithmic output is intended to inform. When it comes to ADM systems, the developer of an algorithmic prediction model typically focuses solely on the task of creating an algorithmic model capable of identifying and scoring individuals with predictive accuracy, without reference to the substantive intervention imposed upon the individuals thereby identified. This prompts question for further research to understand whether, and under what conditions, it may be possible, legitimate and lawful to design a model for an ADM without knowing the specific domain, organisational context, policy and substantive intervention that it will inform.

6. What do our case studies reveal about the role of XAI methods?

This report includes three case studies (based on, but not identical to, real world application cases) to draw attention the potential contribution of technical XAI methods to provide local or global explanations for the predictions intended to inform and 'assist' the decision-maker in question.

- **Case A:** A white box algorithm to help a public welfare authority identify individual benefit fraudsters;
- **Case B:** A black box model embedded into live facial recognition technology (FRT) deployed by the police to match images taken of individuals as they pass in front of a video camera located in an open public setting to those stored in the watchlist, comprised of images of 'wanted' individuals; and
- **Case C:** A black box hiring algorithm used by a private firm to screen individual job applicants.

Case A, involving a white box algorithmic tool, demonstrates that obligations of transparency, accountability and due process apply if those tools are deployed by a public authority to inform decisions that have rights-critical impacts on individuals. XAI techniques provide no assistance in discharging those obligations. Good governance and obligations arising under administrative law highlight the need to ensure that legal and organisational responsibility for the deployment of algorithmic decision-making tools is properly allocated and that those responsible can explain and justify its design, development, deployment and review.

In Case B, a black box real-time facial recognition system deployed by a law enforcement authority, it is evident that many of the questions raised by the case are not concerned with the operation of the algorithmic tool per se, but with organisational and human decisions concerning its specific deployment: and thus, for which XAI methods are of no assistance. It also shows that functional explanations are seldom capable of providing legal justifications for specific decisions or actions. Yet the rule of law demands that the decisions and actions of public authorities must be lawfully authorised. This places the onus on the public authority to identify the legal authority upon which their substantive interventions were taken, and the processes by which those interventions were administered, and to justify their action in those terms. Nevertheless, XAI methods may be helpful in cases where individuals wish to understand why the algorithm identified their face as 'matched' to an image on a watchlist, and to demonstrate that an automated system was used to flag individuals for police intervention, rather than arbitrary decisions by police officers in the field.

In the case of a recruitment tool that incorporates black box model which uses a machine learning algorithm to predict which job applicants would be the best fit for a specific job (Case C) we show that explanations from XAI tools for ML models based on tabular data can help in providing functional explanations concerning why the model produced an output in a given case. However, these do not, in and of themselves, justify any given decision taken on the basis of the model prediction. Nevertheless, they are useful in helping to explain the logic through which the prediction was arrived at and therefore prompting further questions about whether that logic was justified, depending on the context. This case indicates that AI tools can be useful where an ML model is used to fully automate decisions, by providing functional explanations of the underlying logic of a fully automated system and thus help demonstrate compliance with GDPR Art 15.

Taken together, these cases clearly demonstrate that many of the questions raised by these decision-making systems are *not* concerned with the operation of the algorithmic tool per se, but with organisational and human decisions concerning its specific deployment: and thus, for which XAI methods are of no assistance.

Conclusion

In real-world practice, an ADM system is typically embedded within a larger socio-technical system and is executed via an 'organisational decision-making system architecture' which serves as a framework that is intended to govern how its members to carry out their tasks and duties, depicting the various social, technical and organisational components through which a socio-technical system produces a Decision. It demonstrates that there are many and varied choices that must be made within the larger sequence of events that occur within any given socio-technical system, particularly when an ADM contributes in some way, in order to enable the making of a Decision.

Functional explanations must be distinguished from 'reasoned justifications'. A functional explanation provides an account of how a particular outcome was produced. But when someone demands a justification for a particular decision concerning them, they wish to know why that outcome is deemed to be normatively acceptable by reference to some underlying theory of moral or social acceptability — 'yes, we judge that it was acceptable for you to do that to me in these circumstances'. While technical methods known as XAI can, in some circumstances, help to provide functional explanations of why an algorithmic decision-making system which includes a black box model has contributed to the making of a decision that produces legal or other significant effects on the life of an affected person. However, a functional explanation such as those derived by XAI is seldom (if ever) sufficient to provide a justification for a decision based on a model output. Further research is needed to understand whether, and under what conditions, it may be possible, legitimate and lawful to design a model for an ADM without knowing the specific domain, organisational context, policy and substantive intervention that it will inform.