A Systematic Method to Understand Requirements for Explainable AI (XAI) Systems

Mark Hall[1] , **Daniel Harborne**[2] , **Richard Tomsett**[3] , **Vedran Galetic**[1] , **Santiago Quintana-Amate**[1] , **Alistair Nottle**[1] , **Alun Preece**[2]

[1]Airbus Central R&T
[2]Cardiff University
[3]IBM Research, UK

Abstract

This paper presents a five-step systematic method in the development of an explainable AI (XAI) system, to (i) understand specific explanation requirements, (ii) assess existing explanation capabilities and (iii) steer future research and development in this area. A case study is discussed whereby the method was developed and applied within an industrial context.

1 Introduction

Requirements for explainable artificial intelligence (XAI) systems are dependent upon the application and to whom the explanations are intended for [Tomsett *et al.*, 2018] [Bohlender and Köhl, 2019]. There are significant research efforts toward developing new techniques to make AI systems more explainable [Ribeiro *et al.*, 2016] [Lundberg and Lee, 2017]. Simultaneously, there is a growing body of research into metrics and ways in which such explainable methods and tools may be formally evaluated [Mohseni *et al.*, 2018] [Gunning, 2019]. In practice it is challenging to directly compare the effectiveness of these explainable techniques, without a formal set of requirements with respect to a given application [Doshi-Velez and Kim, 2017]. Furthermore, a key challenge exists in understanding unique requirements for explanations within AI systems [Wolf, 2019]. Thus, if research is to be undertaken that applies to real-word industry problems for developer and user communities alike [Preece *et al.*, 2018] [Rosenfeld and Richardson, 2019], then efforts need to be aligned to understand requirements for explainable AI systems.

The contribution of this research is to provide an industry-based engineering foundation to steer future AI research. It addresses a key gap in the literature towards understanding the requirements of stakeholders for XAI systems. The paper presents a systematic method that can be used in the development of an explainable AI system, to (i) understand specific explanation requirements, (ii) assess existing explanation capabilities and (iii) steer future research and development in this area. Section 2 provides an overview of the relevant literature in the field. Section 3 describes a case study within an engineering organisation, presenting a conceptual model in Section 3.1 alongside the five-step method. Sections 3.2 and 3.3 present the application and discussion. Finally, Section 4 provides a summary and proposes future work.

2 Related Work

[Bohlender and Köhl, 2019] stated that the notion of *explanation* is not absolute but relative, such as with respect to tasks and target groups. They argued that methodologies and guidelines within the context of requirements-engineering are needed to ensure that a system is indeed explainable.

[Schneider and Handali, 2019] observed that the personalised explanation in AI is largely absent within the existing literature, whereby knowledge is firstly extracted from the explainee and subsequently used to determine the explanation.

[Tomsett *et al.*, 2018] identified the following six distinct role categories within AI ecosystems:

- *Creators* - agents that create the system
- *Operators* - agents that interact directly with the system
- *Executors* - agents who make decisions that are informed by the system
- *Decision-Subjects* - agents who are affected by decisions made by the executors
- *Data-Subjects* - agents whose personal data has been used to train the system
- *Examiners* - agents auditing or investigating the system

[Tomsett *et al.*, 2018] also stated that this approach is intended to stimulate discussion and suggestions for improvements. Whilst distinguishing between roles is important, this is only the first step towards understanding requirements for explainable AI systems. In addition, [Doshi-Velez and Kim, 2017] claim that creating a shared language is essential for evaluation, citation and comparison of related work - which is currently challenging without a set of explanation characteristics.

The following section describes the prescriptive method that was developed to understand requirements for explainable AI systems.

3 Developing the Method: Case Study

3.1 Method

There is still debate surrounding the meaning of key terms such as explainability and interpretability within the litera-

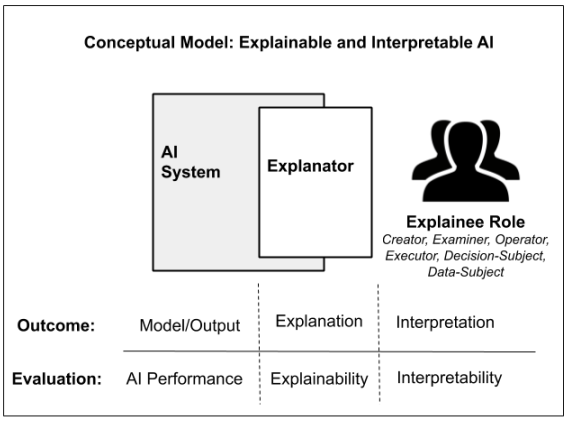

Figure 1: Explainable & Interpretable AI Conceptual Model

ture, with terms often used interchangeably. For the purposes of this research, a simplistic and pragmatic conceptual model was developed from the literature to provide clarity to apply this to an industrial case. This is shown in Figure 1 and described below. In this context, an *outcome* refers to the result of an action or process. For example, an outcome could be an explanation artifact or the resultant change in the mental state of an explainee. In addition, *evaluation* here refers to the assessment of each of these outcomes.

- The outcome of an *AI system* is a *model* or some decision *output*, which is evaluated in terms of *AI system performance*.
- An *explanator* forms part of the AI system which generates *explanation* artifacts, which in turn are evaluated in terms of *explainability* effectiveness.
- An *explainee agent* (human or machine) will consume *explanation* artifacts, and their ability to understand this is evaluated in terms of explanation *interpretability*.

[Bohlender and Köhl, 2019] characterise an explanation that is successfully interpreted, which demonstrates this process: *'A representation E of some information I is an explanation of explanandum X with respect to target group G iff the processing of E by any representative agent A of G makes A understand X.'*. Distinguishing between the explanation artifact and the (human/machine) interpretation of that explanation is useful to conceptualise the problem. Once this conceptual model was adopted, the following five-step method was then developed to be applied within the industrial context, in order to understand requirements for an explainable AI system.

Five-Step Method:

1. Determine the relevant explainee roles within the ecosystem [Tomsett *et al.*, 2018].
2. Determine the relevant explanation characteristics (see Table 1 for the characteristics that were derived).
3. Capture explanation requirements from individuals related to the specific roles. This can be undertaken by a more traditional requirements engineering process, or applying agile methods for requirements elicitation.
4. Assess the ability of appropriate explainable methods to meet these requirements (see Tables 2 & 3 for examples for both LIME and SHAP).
5. Map existing explainable techniques to XAI system requirements. This identifies which existing techniques address specific requirements, and highlight any gaps in capabilities to steer further research and development for the given application.

This systematic method was developed and followed within an industrial research project, to understand the requirements for designing an explainable AI system. The AI system was developed as a proof-of-concept by data science researchers alongside domain experts within the engineering discipline.

3.2 Application

End users for an AI system within the company were identified and their role labels within the ecosystem were established. Semi-structured interviews then took place with these subject matter experts, which enabled initial insights to be gained into the differing types of requirements. A total of 12 individuals were formally interviewed, supported by various informal discussions that took place. The individuals were all experienced aerospace engineers with a variety of engineering responsibilities including system design, testing, verification and validation. The formal interviews ranged between 45-60 minutes, and the discussion was facilitated by the researcher. It became clearer through these discussions what the differing needs would be. For example, in these initial interviews it became evident that the role of a *creator* required some degree of transparency of the AI system when undertaking a debugging task. Whereas, this appeared less relevant for the *executor* role in an operational context for the task of reviewing a single decision retrospectively. Through the course of these discussions, this activity helped to provide some clarity to the situation. However, it became clear that there needed to be a common lexicon to develop a shared understanding if formal requirements were to be developed for each role type. This observation was addressed by deriving a set of explanation characteristics (shown in Table 1), both from the literature and characteristics that emerged from these interviews.

When investigating the most appropriate explainable techniques that could be applied to the problem, it was challenging to formally assess them, especially to compare them against requirements for the AI system. Thus, the explanation characteristics were used as the basis for comparison to be able to map requirements to explainable techniques. Tables 2 and 3 show the assessment for LIME [Ribeiro *et al.*, 2016] and SHAP [Lundberg and Lee, 2017], representing examples of two widely-used techniques. From undertaking this work it was possible to understand the limitations in current techniques to address through research and development.

CHARACTERISTICS	DESCRIPTION
Effectiveness	
Explainability	*Explanation:* The answer to a why-question [Miller, 2019]. The information provided by a system to outline the cause and reason for a decision or output for a performed task [Tomsett *et al.*, 2018] [Lipton, 2016]. It seeks to answer questions such as: 'what, why, why not, what if, and how to' [Lim et al., 2009] *Explainability:* The level to which a system can both *accurately* and *comprehensively* provide the cause of its decisions/outputs [Tomsett *et al.*, 2018]. Unfortunately, there is currently no established measure for explainability that allows the verification of explainability of explanations (e.g. something that would parallel accuracy, precision and recall from model performance). This has led some to suggest "explainability" methods should not be used for some cases, and inherently interpretable AI must be developed instead [Rudin, 2018].
Interpretability	*Interpretation:* The understanding gained by an agent with regard to the cause for a system's decision when presented with an explanation [Tomsett *et al.*, 2018]. *Interpretability:* The degree to which a human can understand the cause of a decision [Miller, 2019] [Guidotti et al., 2018]. The degree to which a human can consistently predict the model's result [Kim, et al., 2016].
Versatility	
Generalisability	The range of model to which the explanation method can be applied [Ras *et al.*, 2018].
Explanatory Power	The scope of questions that the explanation can answer [Ras *et al.*, 2018]
Constraints	
Privacy	The degree to which information on the explainee is extracted, stored and used [Schneider and Handali, 2019], or to which information about the AI and its training data are revealed.
Resources	The computational limitations in providing an explanation.
Timeliness	The appropriate time required to usefully produce and interpret an explanation.
Information Collection Effort	The effort that an individual needs to undertake to provide additional information needed for personalisation [Schneider and Handali, 2019]
Conceptualisation	
Explanator Categories	
Transparency / Post-Hoc	*Transparency:* The level to which a system provides information about its internal workings or structure, and the data it has been trained with [Tomsett *et al.*, 2018]. The opposite of opacity or blackbox-ness [Lipton, 2016]. *Post-hoc:* Presents a distinct approach to extracting information from learned models. While post-hoc interpretations often do not elucidate precisely how a model works, they may nonetheless confer useful information for practitioners and end users of machine learning [Lipton, 2016].
Local / Global	*Global (Holistic/Modular):* Global model explainability/interpretability helps to understand the distribution of your target outcome based on the features. To explain the global model output, you need the trained model, knowledge of the algorithm and the data. This level of interpretability is about understanding how the model makes decisions, based on a holistic view of its features and each of the learned components such as weights, other parameters, and structures [Molnar, 2019]. *Local (Single/Group):* Helps to understand a single instance or group of instances that the model predicts [Molnar, 2019].
Explanator Types	
Feature Importance	The importance of a feature is the increase in the prediction error of the model after we permuted the feature's values, which breaks the relationship between the feature and the true outcome [Molnar, 2019].
Component Data	Returning data points along with ML outcomes [Schneider and Handali, 2019].
Model Internals	Returning the decision maker's internal representations of data points (component data) [Schneider and Handali, 2019].
Feature Visualisation	Refers to the graphical representation of feature importance.
Explanation by Example	Refers to the selection of particular instances of the dataset to explain the behavior of machine learning models or to explain the underlying data distribution [Molnar, 2019].
Contrastive	Explanations that describe how the input would have had to differ from the actual input to produce a specific, alternative model output [McGill and Klein, 1993][Miller, 2018].
Counterfactual	Explanations describing what model output would have been produced, had the input features been different in a specific way [McGill and Klein, 1993][Wachter *et al.*, 2018].
Explanation Properties	
Complexity	Refers to both the size of an explanation, e.g. rule length or decision tree depth, and relationships between features presented in an explanation, e.g. correlation [Paulheim, 2012] or conjunction [Fürnkranz et al., 2018] [Schneider and Handali, 2019].
Prioritization of decision information	Refers to selection of features and feature relationships to present in an explanation.
Representation of Explanation	The form in which the explanation is presented.
Interactivity	The degree to which an explanation may be interacted with.
Personalisation Considerations	
Information Collection	Indicates how information from the explainee, is obtained. Information can refer to knowledge, e.g. how a user solves the task, or to user preferences, e.g. preferred colours in displays [Schneider and Handali, 2019].
Personalization Granularity	Personalization granularity focuses on "to whom to personalize", i.e. a category of individuals or a specific individual. Findings on social identity indicate [Fan and Poole, 2006] that people might behave more according to values and concerns associated with a social group in certain situations. Categorization might be a crude form of personalization, e.g. we might simply categorize users into experts or non-experts, rather than assessing different dimensions related to expertise and customizing along each dimension. However, given sufficiently many categorical attributes on an individual, such as age, gender, origin etc. might also lead to personalization to a specific individual.
Personalisation Automation	Focuses on "who does the personalization" [Fan and Poole, 2006], i.e. a manual personalization done by the explainee or an automatic one by the explanation system [Schneider and Handali, 2019].

Table 1: Description of explanation characteristics framework

CHARACTERISTICS	LIME
Effectiveness	
Explainability	Work has shown that the base implementation of LIME can be unstable but that with modification this can be accounted for and the instability can be reduced [Lee *et al.*, 2018].
Interpretability	When working with imagery, highlighted regions of pixels tend to be large and defined by edges. It is therefore easier for a semantic feature to be identified from each region. As the explanation purely highlights input features (with no accompanying explanation of why those features were important to the model), the information in the explanation can be open to interpretation. This could lead to scenarios where the user projects their ideas on to the explanation rather than taking away the true explanation being conveyed within the LIME output. For tabular or text-based data, the outputs are more clearly interpretable.
Versatility	
Generalisability	Model agnostic. Explanations can be generated using model as black box.
Explanatory Power	Explanation aims to answer: "Which features of the input were evidence for a given class, and which features were evidence against it?" — The given class can be the predicted class or another of the model's output classes. Answers to other questions can only be inferred from this explanation output.
Constraints	
Privacy	For some modalities(imagery, text) input data is revealed as part of the explanation.
Resources	A new input is created for each input feature (super pixel in the case of imagery) and each of these is passed to the model for a forward pass prediction. (This has implications on memory requirements in the current implementation of the LIME library as they are all created, and processed at once).
Timeliness	Time to generate an explanation is based on the time taken to perform a forward pass using the original model. Compared to other explanations, this is relatively quick.
Information Collection Effort (for personalisation)	Parameters that can be personalised: Colours used for highlighting, size of super pixel regions (imagery). The colours used for highlighting can be set by default to options which cater for factors like colour-blindness. Super pixel size can often be left as default. Adding an interface to allow for a user to configure these items is possible.
Conceptualisation	
Explanator Categories	
Transparency / Post-Hoc	Post-Hoc - use of surrogate models to gauge local impact of regions of the input image on the actual models prediction.
Local / Global	Local
Explanator Types	
Feature Importance	Yes
Componant Data (Data Points)	No
Model Internals	No
Feature Visualisation	No
Explanation by Example	No
Contrastive	Yes
Counterfactual	No
Explanation Properties	
Complexity	Low - shaded input features (and large shaded regions for imagery). Does not require Machine Learning knowledge to attempt to interpret.
Prioritization of decision information	All information generated by the explanation is returned.
Representation of Explanation	Modality: Values returned by technique are visualised as highlighted input features. Resultant modality is often an image.
Interactivity	None.
Personalisation Considerations	
Information Collection	Minimal personalisation is offered - colours used could be changed based on a record of the user or by option present in interface.
Personalization Granularity	Personalisation available does not target the role of the user making use of the explanation.
Personalisation Automation	Personalisation could be automated by application (though is not offered by default).

Table 2: Assessment of the explanation technique LIME using the explanation characteristics framework.

CHARACTERISTICS	SHAP
Effectiveness	
Explainability	Output is stable.
Interpretability	Explanations consist of highlighting input features. For imagery this is done at the pixel level. This can create abstract patterns across the image which can be hard to identify semantic features from. In addition, explanations can often be generated in which parts of the same object in the image are shaded both as evidence for a given class and against it. This makes it difficult to interpret and make definitive conclusions about what was important to the model.For tabular or text-based data, the outputs are more clearly interpretable.
Versatility	
Generalisability	Need to use dedicated implementation of SHAP for the type of model.For Deep Learning models, SHAP needs access to their hidden layers.
Explanatory Power	Explanation aims to answer: "which features of the input were evidence for a given class, and which features were evidence against it?" - the given class can be the predicted class or any of the output classes. Answers to other questions can only be inferred from this explanation output.
Constraints	
Privacy	For some modalities (imagery, text) input data is revealed as part of the explanation. A proportion of the training data must be provided as background to generate expectation values. These aren't revealed during explanation but the movement and processing of these images should be managed appropriately if they are sensitive.
Resources	Requires a catalogue of input examples to be passed as background to build the baseline expectation values used in the explanation. Where input examples are large (e.g. in the case of imagery), batching will need to take place. Batch size will be determined by resources available and a small batch size will have implications on explanation generation time.
Timeliness	As SHAP requires expectation values to be calculated before generating an explanation, there is a "warm-up" time for the technique to be ready. The time needed is based on the number of examples from the training data provided as background. Providing more examples allows for a more accurate explanation but takes more time. After the background probabilities are generated, they can be reused for further explanations for the same dataset/domain. The background can be calculated ahead of the first explanation being required.
Information Collection Effort	Parameters that can be personalised: Number of background examples used to calculate expectations, Colours used for highlighting. Number of images used to generate expectation value can be offered to the user to customise but changing it will cause the values to be recalculated which takes time and resources. The colours used for highlighting can be set by default to options which cater for factors like colour-blindness.
Conceptualisation	
Explanator Categories	
Transparency / Post-Hoc	Post-Hoc - SHAP values are calculated based on a generated expectation of values for each pixel. Whilst the technique uses the model internals, it provides an estimate of what was important to the model in the form of shapley value [Winter, 2002] rather than revealing the internals of the model.
Local / Global	Local.For tabular data, a summary indicating feature importance across a dataset can be generated.
Explanator Types	
Feature Importance	Yes
Componant Data (Data Points)	No
Model Internals	No (They are used but not revealed).
Feature Visualisation	No
Explanation by Example	No
Contrastive	Yes
Counterfactual	No
Explanation Properties	
Complexity	Low (most modalities) Medium to High (imagery) — Does not require Machine Learning knowledge to attempt to interpret. For imagery, highlighting takes place at the pixel level. This often leads to uninterpretable explanations consisting of scattered colours.
Prioritization of Decision Information	All information generated by the explanation is returned.
Representation of Explanation	Modality: Values returned by technique are visualised as highlighted input features. Resultant modality is often an image.
Interactivity	None.
Personalisation Considerations	
Information Collection	Minimal personalisation is offered - colours used could be changed based on a record of the user or by option present in interface.
Personalization Granularity	Personalisation available does not target the role of the user making use of the explanation.
Personalisation Automation	Personalisation could be automated by application (though is not offered by default).

Table 3: Assessment of the explanation technique SHAP using the explanation characteristics framework.

3.3 Discussion

Method Strengths:

- Identifying key stakeholders and their roles within the AI ecosystem, as outlined by [Tomsett *et al.*, 2018] proved to be beneficial in practice to recognise differing end-user needs early in the process. This distinction between roles enabled the AI system to be *designed for explainability*, recognising that explanations are relative to specific roles and tasks. This step also provided clarity when establishing certain stakeholders that may require some level of explanation, but may not directly interact with the system themselves.
- The explanation characteristics proved to be effective for framing discussion between data science researchers and subject matter experts to understand explanation requirements. These characteristics supported the facilitation of the discussion, to enable the problem to be understood more clearly with unambiguous terminology. The initial semi-structured interviews that were conducted without these characteristics did not result in clear requirements, demonstrating the need for a set of established explanation characteristics that may be systematically addressed.
- The explanation characteristics also provided the means by which to assess the capability of existing explainable techniques in XAI. Without a form of explanation characteristics, it would not be possible to compare and evaluate the effectiveness of these techniques.

Method Limitations:

- Throughout the discussions with individuals representing different role types, it became evident that requirements differed depending on the following, which would benefit from further consideration:
  - AI Application – The requirements would differ if the system were to be a decision-support machine learning system, or embedded as an autonomous component within wider engineering systems.
  - Criticality - The level of risk associated with the AI system influences the explainability requirements, such as time-critical or safety-critical applications.
  - System life cycle stage - The discussions highlighted that a specific role may need to undertake tasks at different points in the life cycle, for example a system *creator* may have different requirements when performing a debugging task during model building compared to maintaining a model in operation.
- Explanation characteristics - These would benefit from being developed further to provide greater granularity and support a more comprehensive assessment of explainable techniques.

4 Conclusion

This paper presented a five-step systematic method for understanding XAI requirements, that was developed and applied for an industrial AI system. This research has established an approach to understanding explainable requirements in a given AI system. However, it is recommended that future work should include application of this method to a variety of scenarios, and further refinement in an iterative manner.

References

[Bohlender and Köhl, 2019] Dimitri Bohlender and Maximilian Köhl. Towards a characterization of explainable systems. *arXiv:1902.03096*, 2019.

[Doshi-Velez and Kim, 2017] Finale Doshi-Velez and Been Kim. Towards a Rigorous Science of Interpretable Machine Learning. *arXiv:1702.08608*, 2017.

[Gunning, 2019] David Gunning. Darpa's explainable artificial intelligence (xai) program. In *Proceedings of the 24th International Conference on Intelligent User Interfaces*, IUI '19, pages ii–ii, New York, NY, USA, 2019. ACM.

[Lee *et al.*, 2018] Eunjin Lee, David Braines, Mitchell Stiffler, and Daniel Harborne. Developing the sensitivity of LIME for better machine learning explanation. In *Artificial Intelligence and Machine Learning for Multi Domain Battle Applications*. SPIE Defense + Commercial Sensing, 2018.

[Lipton, 2016] Zachary Chase Lipton. The Mythos of Model Interpretability. *Proceedings of the ICML Workshop on Human Interpretability in Machine Learning*, pages 96–100, 2016.

[Lundberg and Lee, 2017] Scott M Lundberg and Su-In Lee. A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems 30*, pages 4765–4774. 2017.

[McGill and Klein, 1993] Ann L. McGill and Jill G. Klein. Contrastive and counterfactual reasoning in causal judgment. *Journal of Personality and Social Psychology*, 64(6):897–905, 1993.

[Miller, 2018] Tim Miller. Contrastive explanation: A structural-model approach. *arXiv:1811.03163*, 2018.

[Miller, 2019] Tim Miller. Explanation in artificial intelligence: Insights from the social sciences. *Artificial Intelligence*, 267:1–38, 2019.

[Mohseni *et al.*, 2018] Sina Mohseni, Niloofar Zarei, and Eric D. Ragan. A survey of evaluation methods and measures for interpretable machine learning. *arXiv:1811.11839*, 2018.

[Molnar, 2019] Christof Molnar. *Interpretable Machine Learning*. 2019.

[Preece *et al.*, 2018] Alun Preece, Dan Harborne, Dave Braines, Richard Tomsett, and Supriyo Chakraborty. Stakeholders in Explainable AI. In *AAAI FSS-18: Artificial Intelligence in Government and Public Sector*. 2018.

[Ras *et al.*, 2018] Gabrielle Ras, Marcel van Gerven, and Pim Haselager. Explanation Methods in Deep Learning: Users, Values, Concerns and Challenges. *arXiv:1803.07517*, 2018.

[Ribeiro *et al.*, 2016] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. Why should I trust you?: Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 1135–1144. ACM, 2016.

[Rosenfeld and Richardson, 2019] Avi Rosenfeld and Ariella Richardson. Explainability in human–agent systems. *Autonomous Agents and Multi-Agent Systems*, pages 1–33, 2019.

[Rudin, 2018] Cynthia Rudin. Please stop explaining black box models for high stakes decisions. *arXiv:1811.10154*, 2018.

[Schneider and Handali, 2019] Johanes Schneider and Joshua Handali. Personalized explanation in machine learning: a conceptualization. In *27th European Conference on Information Systems (ECIS 2019)*. 2019.

[Tomsett *et al.*, 2018] Richard Tomsett, Dave Braines, Dan Harborne, Alun Preece, and Supriyo Chakraborty. Interpretable to whom? A role-based model for analyzing interpretable machine learning systems. *Proceedings of the ICML Workshop on Human Interpretability in Machine Learning*, pages 1–8, 2018.

[Wachter *et al.*, 2018] Sandra Wachter, Brent Mittelstadt, and Chris Russell. Counterfactual explanations without opening the black box: automated decisions and the gdpr. *Harvard Journal of Law and Technology*, 31(2):841–887, 2018.

[Winter, 2002] Eyal Winter. Chapter 53: The Shapley value. In *Handbook of Game Theory with Economic Applications Volume 3*, pages 2025–2054. 2002.

[Wolf, 2019] Christine T. Wolf. Explainability scenarios: Towards scenario-based xai design. In *Proceedings of the 24th International Conference on Intelligent User Interfaces*, IUI '19, pages 252–257, New York, NY, USA, 2019. ACM.