A Survey of Explainable AI Terminology

Miruna A. Clinciu and Helen F. Hastie
Edinburgh Centre for Robotics
Heriot-Watt University, Edinburgh, EH14 4AS, UK
{mc191, H.Hastie}@hw.ac.uk

Abstract

The field of Explainable Artificial Intelligence attempts to solve the problem of algorithmic opacity. Many terms and notions have been introduced recently to define Explainable AI, however, these terms seem to be used interchangeably, which is leading to confusion in this rapidly expanding field. As a solution to overcome this problem, we present an analysis of the existing research literature and examine how key terms, such as *transparency*, *intelligibility*, *interpretability*, and *explainability* are referred to and in what context. This paper, thus, moves towards a standard terminology for Explainable AI.

***Keywords*—** Explainable AI, black-box, NLG, Theoretical Issues, Transparency, Intelligibility, Interpretability, Explainability

1 Introduction

In recent years, there has been an increased interest in the field of Explainable Artificial Intelligence (XAI). However, there is clear evidence from the literature that there are a variety of terms being used interchangeably such as *transparency*, *intelligibility*, *interpretability*, and *explainability*, which is leading to confusion. Establishing a set of standard terms to be used by the community will become increasingly important as XAI is mandated by regulation, such as the GDPR and as standards start to appear such as the IEEE standard in transparency (P7001). This paper works towards this goal.

Explainable Artificial Intelligence is not a new area of research and the term **explainable** has existed since the mid-1970s (Moore and Swartout, 1988). However, XAI has come to the forefront in recent times due to the advent of deep machine learning and the lack of transparency of "black-box" models. We introduce below, some descriptions of XAI collected from the literature:

- "Explainable AI can present the user with an easily understood chain of reasoning from the user's order, through the AI's knowledge and inference, to the resulting behaviour" (van Lent et al., 2004).
- "XAI is a research field that aims to make AI systems results more understandable to humans" (Adadi and Berrada, 2018).

Thus, we conclude that XAI is a research field that focuses on giving AI decision-making models the ability to be easily understood by humans. Natural language is an intuitive way to provide such Explainable AI systems. Furthermore, XAI will be key for both expert and non-expert users to enable them to have a deeper understanding and the appropriate level of trust, which will hopefully lead to increased adoption of this vital technology.

This paper firstly examines the various notions that are frequently used in the field of Explainable Artificial Intelligence in Section 2 and attempts to organise them diagrammatically. We then discuss these terms with respect to Natural Language Generation in Section 3 and provide conclusions.

2 Terminology

In this section, we examine four key terms found frequently in the literature for describing various techniques for XAI. These terms are illustrated in Figure 1, where we organise them as a Venn diagram that describes how a *transparent* AI system has several facets, which include *intelligibility*, *explainability*, and *interpretability*. Below, we discuss how *intelligibility* can be discussed in terms of *explainability* and/or *interpretability*. For each of these terms, we present the dictionary definitions extracted from modern and notable English dictionaries, quotes from the literature presented in tables and discuss how they support the proposed structure given in Figure 1. In every table, we emphasise related words and context, in order

Proceedings of the 1st Workshop on Interactive Natural Language Technology for Explainable Artificial Intelligence (NL4XAI 2019), pages 8–13,
Tokyo, Japan, October 29 2019.

to connect ideas and build up coherent relationships within the text.

In this paper, the first phase of the selection criteria of publications was defined by the relevance of the paper and related key words. The second phase was performed manually by choosing the papers that define or describe the meaning of the specified terms or examine those terms for ways in which they are different, alike, or related to each other.

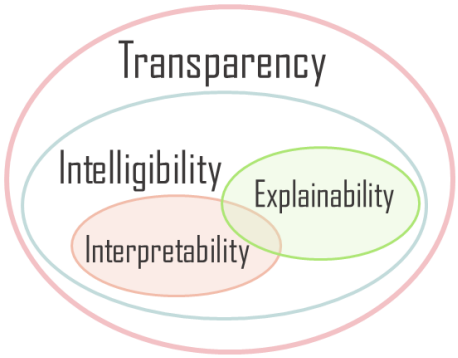

Figure 1: A Venn Diagram of the relationship between frequently used terms, that offers a representation of the authors' interpretation for the field, excluding post-hoc interpretation.

Transparency

Dictionary definitions: The word "*transparent*" refers to something that is "clear and easy to understand" (Cambridge Dictionary, 2019d); or "easily seen through, recognized, understood, detected; manifest, evident, obvious, clear" (Oxford English Dictionary, 2019d); or "language or information that is transparent is clear and easy to understand" (The Longman Dictionary of Contemporary English, 2019c).

Conversely, an opaque AI system is a system with the lowest level of transparency, known as a "**black-box**" model. A similar definition is given by Tomsett et al. (2018) in Table 1.

Tintarev and Masthoff (2007) state that *transparency* "explains how the system works" and it is considered one of the possible explanation facilities that could influence good recommendations in recommender systems.

In the research paper by Cramer et al. (2008), *transparency* aims to increase understanding and entails offering the user insight as to how a system works, for example, by offering explanations for system choices and behaviour.

"**Transparency** clearly describing the model structure, equations, parameter values, and assumptions to enable interested parties to understand the model" (Briggs et al., 2012).
Tomsett et al. (2018) defined **transparency** as a "level to which a system provides information about its internal workings or structure" and both "**explainability** and **transparency** are important for improving creator-interpretability".
"Informally, **transparency** is the opposite of opacity or blackbox-ness. It connotes some sense of understanding the mechanism by which the model works. We consider **transparency** at the level of the model (simulatability), at the level of individual components (e.g. parameters) (decomposability), and at the level of the training algorithm (algorithmic transparency)" (Lipton, 2016).

Table 1: Various notions of Transparency presented in recent research papers

Intelligibility

Dictionary definitions: An "*intelligible*" system should be "clear enough to be understood" according to Cambridge Dictionary (2019b); or "capable of being understood; comprehensible" (Oxford English Dictionary, 2019b); or "easily understood" (The Longman Dictionary of Contemporary English, 2019d).

The concept of *intelligibility* was defined by Bellotti and Edwards (2001) from the perspective of "context-aware systems that seek to act upon what they infer about the context must be able to represent to their users what they know, how they know it, and what they are doing about it" (Bellotti and Edwards, 2001).

As illustrated in Table 2, it is challenging to define how intelligible AI systems could be designed, as they would need to communicate very complex computational processes to various types of users (Weld and Bansal, 2018). Per the Venn diagram in Figure 1, we consider that an AI system could become intelligible in a number of ways, but also through *explanations* (e.g. in natural language) and/or *interpretations*. We discuss both of these in turn below.

"It remains remarkably hard to specify what makes a system **intelligible**; The key challenge for designing **intelligible AI** is communicating a complex computational process to a human. Specifically, we say that a model is **intelligible** to the degree that a human user can predict how a change to a feature" (Weld and Bansal, 2018).
"**Intelligibility** can help expose the inner workings and inputs of context-aware applications that tend to be opaque to users due to their implicit sensing and actions" (Lim and Dey, 2009).

Table 2: Various notions of Intelligibility presented in recent research papers

Interpretability

Dictionary Definitions: According to Cambridge Dictionary (2019c), the word "*interpret*" definition is "to decide what the intended meaning of something is"; or "to expound the meaning of (something abstruse or mysterious); to render (words, writings, an author, etc.) clear or explicit; to elucidate; to explain" (Oxford English Dictionary, 2019c); or "to explain the meaning of something" (The Longman Dictionary of Contemporary English, 2019b).

Considering a "black-box" model, we will try to understand how users and developers could define the model *interpretability*. A variety of definitions of the term *interpretability* have been suggested in recent research papers, as presented in Table 3.

Various techniques have been used to give insights into an AI model through interpretations, such as Feature Selection Techniques (Kim et al., 2015), Shapley Values (Sundararajan and Najmi, 2019); the interpretation of the AI model interpretation e.g. Hybrid AI models (Wang and Lin, 2019), by combining interpretable models with opaque models, and output interpretation (e.g. Evaluation Metrics Interpretation (Mohseni et al., 2018), and Visualisation Techniques Interpretation (Samek et al., 2017; Choo and Liu, 2018)). Thus in our model in Figure 1, we define *interpretability* as intersecting with *explainability* as some models may be interpretable without needing explanations.

"In model-agnostic **interpretability**, the model is treated as a black-box. **Interpretable models** may also be more desirable when interpretability is much more important than accuracy, or when interpretable models trained on a small number of carefully engineered features are as accurate as black-box models". (Ribeiro et al., 2016)
"An **explanation** can be evaluated in two ways: according to its **interpretability**, and according to its completeness" (Gilpin et al., 2018).
"We define **interpretable** machine learning as the use of machine-learning models for the extraction of relevant knowledge about domain relationships contained in data..." (Murdoch et al., 2019).

Table 3: Various notions of Interpretability presented in recent research papers

Explainability

Dictionary Definitions: For the word "*explain*" were extracted the following definitions: "to make something clear or easy to understand by describing or giving information about it" Cambridge Dictionary (2019a); or "to provide an explanation for something. to make plain or intelligible" (Oxford English Dictionary, 2019a); or "to tell someone about something in a way that is clear or easy to understand. to give a reason for something or to be a reason for something" (The Longman Dictionary of Contemporary English, 2019a).

Per these definitions, providing explanations is about improving the user's mental model of how a system works. Ribera and Lapedriza (2019) consider that we do not have a concrete definition for *explanation* in the literature. However, according to these authors, every definition relates "explanations with "why" questions or causality reasonings". Given the nature of the explanations, Ribera and Lapedriza (2019) proposed to categorise the explainees in three main groups, based on their goals, background, and relationship with the product, namely: developers and AI researchers, domain experts, and lay users. Various types of explanations have been presented in the literature such as "why" and "why not" (Kulesza et al., 2013) or Adadi and Berrada (2018)'s four types of explanations that are used to "justify, control, discover and improve". While it is out of scope

to go into detail here, what is clear is that in most uses of the term *explainability*, it means providing a way to improve the understanding of the user, whomever they may be.

"**Explanation** is considered closely related to the concept of **interpretability**" (Biran and Cotton, 2017).
"**Transparent** design: model is inherently **interpretable** (globally or locally)" (Lucic et al., 2019).
"I equate **interpretability** with **explainability**" (Miller, 2018).
"Systems are **interpretable** if their operations can be understood by a human, either through **introspection** or through a **produced explanation**" (Biran and Cotton, 2017).
In the paper (Poursabzi-Sangdeh et al., 2018), **interpretability** is defined as something "that cannot be manipulated or measured, and could be defined by people, not algorithms".

Table 4: Various notions of Explainability presented in recent research papers

3 The Role of NLG in XAI

An intuitive medium to provide such explanations is through natural language. The human-like capability of Natural Language Generation (NLG) has the potential to increase the intelligibility of an AI system and enable a system to provide explanations that are tailored to the end-user (Chiyah Garcia et al., 2019).

One can draw an analogy between natural language generation of explanations and Lacave and Diez's model of explanation generation for expert systems (Lacave and Díez, 2002); or Reiter and Dale's NLG pipeline (Reiter and Dale, 2000) with stages for determining "what" to say in an explanation (content selection) and "how" to say it (surface realisation). Lacave and Diez's model also emphasises the importance of adapting to the user, which is also a focus area in NLG (e.g. adapting styles (Dethlefs et al., 2014)).

Other studies have looked at agents and robots providing a rationalisation of their behaviour (Ehsan et al., 2018) by providing a running commentary in language. Whilst this is not necessarily how humans behave, it is beneficial to be able to provide such *rationalisation*, especially in the face of unusual behaviour and, again, natural language is one way to do this. Defined as a process of producing an explanation for an agent or system behavior as if a human had performed the behaviour, *AI rationalisation* has multiple advantages to be taken into consideration: "naturally accessible and intuitive to humans, especially non-experts, could increase the satisfaction, confidence, rapport, and willingness to use autonomous systems and could offer real-time response" (Ehsan et al., 2018).

4 Conclusions and Future work

In this paper, we introduced various terms that could be found in the field of Explainable AI and their concrete definition. In Figure 1, we have attempted to define the relationship between the main terms that define Explainable AI. Intelligibility could be achieved through explanations and interpretations, where the type of user, their background, goal and current mental model are taken into consideration.

As mentioned previously, *interpretability* is defined as a concept close to *explainability* (Biran and Cotton, 2017). Our Venn diagram given in Figure 1 illustrates that transparent systems could be, by their nature interpretable, without providing explanations and that the activities of interpreting a model and explaining why a system behaves the way it does are fundamentally different. We posit, therefore, that the field moving forward should be wary of using such terms interchangeably. Natural Language Generation will be key to providing explanations, and rationalisation is one approach that we have discussed here.

Evaluation of NLG is challenging area (Hastie and Belz, 2014) with objective measures such as BLEU being shown not to reflect human ratings (Liu et al., 2016). How natural language explanations are evaluated will likely be based on, in the near term at least, subjective measures that try to evaluate an explanation in terms of whether it improves a system's *intelligibility*, *interpretability* and *transparency* along with other typical metrics related to the quality and clarity of the language used (Curry et al., 2017).

In future work, it would be advisable to perform empirical analysis of research papers related to the various terms and notions introduced here and continuously being added into the field of XAI.

Acknowledgements

The authors gratefully acknowledge the support of Dr. Inês Cecilio, Prof. Mike Chantler, and Dr. Vaishak Belle. This research was funded by Schlumberger Cambridge Research Centre Doctoral programme.

References

Amina Adadi and Mohammed Berrada. 2018. Peeking Inside the Black-Box: A Survey on Explainable Artificial Intelligence (XAI). *IEEE Access*, 6:52138–52160.

Victoria Bellotti and Keith Edwards. 2001. Intelligibility and accountability: Human considerations in context-aware systems. *Human-Computer Interaction*, 16(2-4):193–212.

Or Biran and Courtenay Cotton. 2017. Explanation and Justification in Machine Learning: A Survey. In *Proceedings of the 1st Workshop on Explainable Artificial Intelligence, IJCAI 2017*.

Andrew H. Briggs, Milton C. Weinstein, Elisabeth A. L. Fenwick, Jonathan Karnon, Mark J. Sculpher, and A. David Paltiel. 2012. Model parameter estimation and uncertainty: A report of the ispor-smdm modeling good research practices task force-6. *Value in Health*, 15(6):835–842.

Cambridge Dictionary. 2019a. *Explain*. Cambridge University Press. Accessed on 2019-08-25.

Cambridge Dictionary. 2019b. *Intelligible*. Cambridge University Press. Accessed on 2019-08-25.

Cambridge Dictionary. 2019c. *Interpret*. Cambridge University Press. Accessed on 2019-08-25.

Cambridge Dictionary. 2019d. *Transparent*. Cambridge University Press. Accessed on 2019-08-25.

Francisco Javier Chiyah Garcia, David A. Robb, Xingkun Liu, Atanas Laskov, Pedro Patron, and Helen Hastie. 2019. Explainable Autonomy: A Study of Explanation Styles for Building Clear Mental Models. In *Proceedings of the International Natural Language Generation (INLG)*.

J. Choo and S. Liu. 2018. Visual analytics for explainable deep learning. *IEEE Computer Graphics and Applications*, 38(4):84–92.

Henriette Cramer, Vanessa Evers, Satyan Ramlal, Maarten van Someren, Lloyd Rutledge, Natalia Stash, Lora Aroyo, and Bob Wielinga. 2008. The effects of transparency on trust in and acceptance of a content-based art recommender. *User Modeling and User-Adapted Interaction*, 18(5):455.

Amanda Cercas Curry, Helen Hastie, and Verena Rieser. 2017. A review of evaluation techniques for social dialogue systems. In *ISIAA 2017 - Proceedings of the 1st ACM SIGCHI International Workshop on Investigating Social Interactions with Artificial Agents, Co-located with ICMI 2017*, pages 25–26. Association for Computing Machinery, Inc.

Nina Dethlefs, Heriberto Cuayáhuitl, Helen Hastie, Verena Rieser, and Oliver Lemon. 2014. Cluster-based prediction of user ratings for stylistic surface realisation. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics 2014, EACL 2014*, pages 702–711. Association for Computational Linguistics (ACL).

Upol Ehsan, Brent Harrison, Larry Chan, and Mark O. Riedl. 2018. Rationalization: A Neural Machine Translation Approach to Generating Natural Language Explanations. In *AIES 2018 - Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, pages 81–87. Association for Computing Machinery, Inc.

L. H. Gilpin, D. Bau, B. Z. Yuan, A. Bajwa, M. Specter, and L. Kagal. 2018. Explaining explanations: An overview of interpretability of machine learning. In *Proceedings of the 5th International Conference on Data Science and Advanced Analytics (DSAA) 2018 IEEE*, pages 80–89. IEEE.

Helen Hastie and Anja Belz. 2014. A comparative evaluation methodology for nlg in interactive systems. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, Reykjavik, Iceland. European Language Resources Association (ELRA).

Been Kim, Julie A Shah, and Finale Doshi-Velez. 2015. Mind the gap: A generative approach to interpretable feature selection and extraction. In *Proceedings of the Twenty-ninth Conference on Neural Information Processing Systems, NeurIPS 2015*, pages 2260–2268. Curran Associates, Inc.

T. Kulesza, S. Stumpf, M. Burnett, S. Yang, I. Kwan, and W. Wong. 2013. Too much, too little, or just right? ways explanations impact end users' mental models. In *Proceedings of the 2013 IEEE Symposium on Visual Languages and Human Centric Computing*, pages 3–10. IEEE.

Carmen Lacave and Francisco J. Díez. 2002. A Review of Explanation Methods for Bayesian Networks. *The Knowledge Engineering Review*, 17(2):107–127.

Michael van Lent, William Fisher, and Michael Mancuso. 2004. An explainable artificial intelligence system for small-unit tactical behavior. In *Proceedings of the 16th Conference on Innovative Applications of Artifical Intelligence*, IAAI'04, pages 900–907. AAAI Press.

Brian Y. Lim and Anind K. Dey. 2009. Assessing demand for intelligibility in context-aware applications. In *Proceedings of the 11th International Conference on Ubiquitous Computing*, UbiComp '09, pages 195–204, New York, NY, USA. ACM.

Zachary Chase Lipton. 2016. The mythos of model interpretability. *arXiv preprint arXiv:1606.03490*.

Chia-Wei Liu, Ryan Lowe, Iulian V. Serban, Michael Noseworthy, Laurent Charlin, and Joelle Pineau. 2016. How NOT To Evaluate Your Dialogue System: An Empirical Study of Unsupervised Evaluation Metrics for Dialogue Response Generation. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing, EMNLP 2016*, page 13.

Ana Lucic, Hinda Haned, and Maarten de Rijke. 2019. Contrastive explanations for large errors in retail forecasting predictions through monte carlo simulations. *arXiv preprint arXiv:1908.00085*.

Tim Miller. 2018. Explanation in Artificial Intelligence: Insights from the Social Sciences. *arXiv preprint arXiv:1706.07269*.

Sina Mohseni, Niloofar Zarei, and Eric D. Ragan. 2018. A survey of evaluation methods and measures for interpretable machine learning. *arXiv preprint arXiv:1811.11839*, abs/1811.11839.

J.D. Moore and W.R. Swartout. 1988. *Explanation in Expert Systems: A Survey*. Number no. 228 in Explanation in Expert Systems: A Survey. University of Southern California, Information Sciences Institute.

W James Murdoch, Chandan Singh, Karl Kumbier, Reza Abbasi-Asl, and Bin Yu. 2019. Interpretable machine learning: definitions, methods, and applications. *arXiv preprint arXiv:1901.04592*.

Oxford English Dictionary. 2019a. *explain, v.* Oxford University Press. Accessed on 2019-11-10.

Oxford English Dictionary. 2019b. *intelligible, adj. (and n.)*. Oxford University Press. Accessed on 2019-11-10.

Oxford English Dictionary. 2019c. *interpret, v.* Oxford University Press. Accessed on 2019-11-10.

Oxford English Dictionary. 2019d. *transparent, adj. (and n.)*. Oxford University Press. Accessed on 2019-11-10.

Forough Poursabzi-Sangdeh, Daniel G. Goldstein, Jake M. Hofman, Jennifer Wortman Vaughan, and Hanna M. Wallach. 2018. Manipulating and measuring model interpretability. *arXiv preprint arXiv:1802.07810*.

Ehud Reiter and Robert Dale. 2000. *Building Natural Language Generation Systems*. Cambridge University Press, New York, NY, USA.

Marco Túlio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. Model-agnostic interpretability of machine learning. *arXiv preprint arXiv:1606.05386*.

Mireia Ribera and Agata Lapedriza. 2019. Can we do better explanations? A proposal of user-centered explainable AI. In *Proceedings of the CEUR Workshop*, volume 2327. CEUR-WS.

Wojciech Samek, Thomas Wiegand, and Klaus-Robert Mller. 2017. Explainable artificial intelligence: Understanding, visualizing and interpreting deep learning models. *ITU Journal: ICT Discoveries - Special Issue 1 - The Impact of Artificial Intelligence (AI) on Communication Networks and Services*, 1:1–10.

Mukund Sundararajan and Amir Najmi. 2019. The many shapley values for model explanation. *arXiv preprint arXiv:1908.08474*, abs/1908.08474.

The Longman Dictionary of Contemporary English. 2019a. *explain*. Pearson Longman. Accessed on 2019-11-10.

The Longman Dictionary of Contemporary English. 2019b. *interpret, v.* Pearson Longman. Accessed on 2019-11-10.

The Longman Dictionary of Contemporary English. 2019c. *transparent, adj. (and n.)*. Pearson Longman. Accessed on 2019-11-10.

The Longman Dictionary of Contemporary English. 2019d. *transparent, adj. (and n.)*. Pearson Longman. Accessed on 2019-11-10.

Nava Tintarev and Judith Masthoff. 2007. Effective explanations of recommendations: User-centered design. In *Proceedings of the 2007 ACM Conference on Recommender Systems*, RecSys '07, pages 153–156, New York, NY, USA. ACM.

Richard Tomsett, Dave Braines, Dan Harborne, Alun D. Preece, and Supriyo Chakraborty. 2018. Interpretable to whom? A role-based model for analyzing interpretable machine learning systems. *arXiv preprint arXiv:1806.07552*, abs/1806.07552.

Tong Wang and Qihang Lin. 2019. Hybrid predictive model: When an interpretable model collaborates with a black-box model. *arXiv preprint arXiv:1905.04241*.

Daniel S. Weld and Gagan Bansal. 2018. Intelligible artificial intelligence. *arXiv preprint arXiv:1803.04263*.