1st Workshop on Interactive Natural Language Technology for Explainable Artificial Intelligence
October 29, 2019

A Survey of Explainable AI Terminology

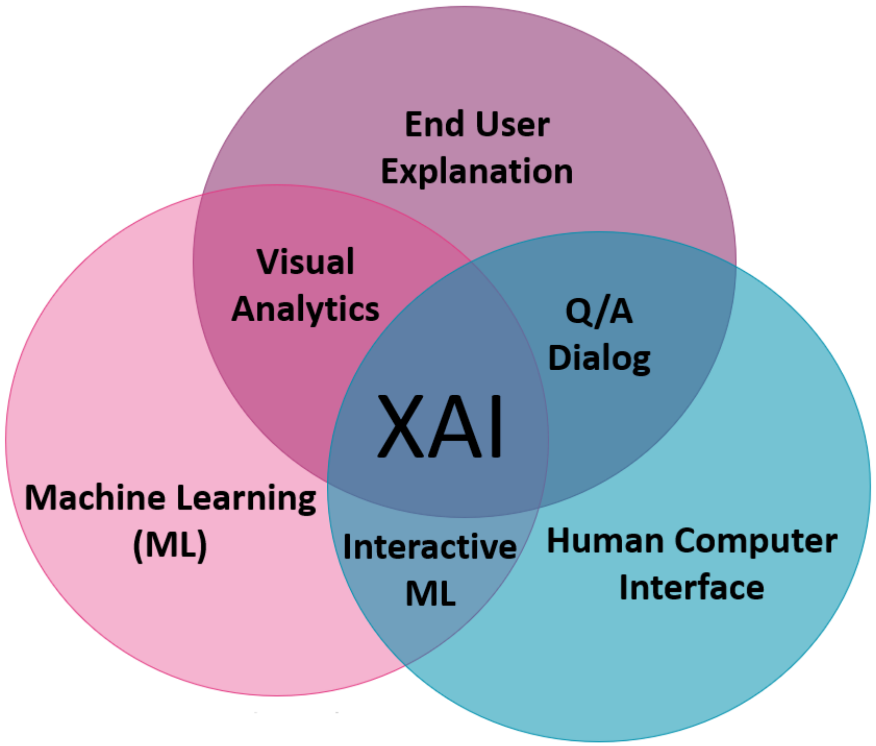
Miruna-Adriana Clinciu and Helen F. Hastie
Edinburgh Centre for Robotics
Heriot-Watt University, Edinburgh, UK

FEAR OF THE UNKNOWN…

EXPLAINABLE AI...

"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 behavior" [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].

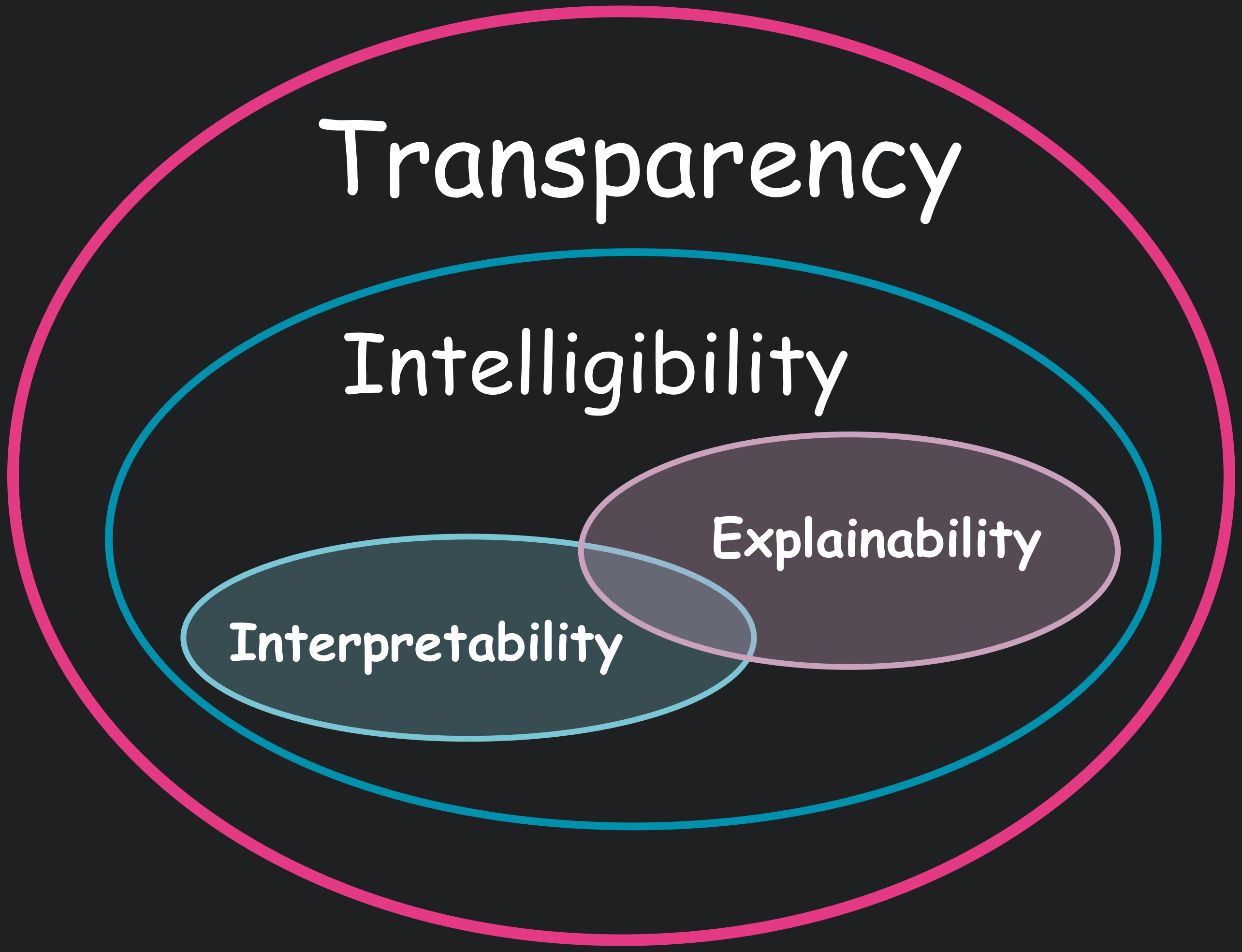


Michael van Lent, William Fisher, and Michael Mancuso. 2004. An explainable artificial intelligence system for small-unit tactical behavior. In Proceedings of the 16th conference on Innovative applications of artifical intelligence (IAAI'04), Randall Hill (Ed.). AAAI Press 900-907.

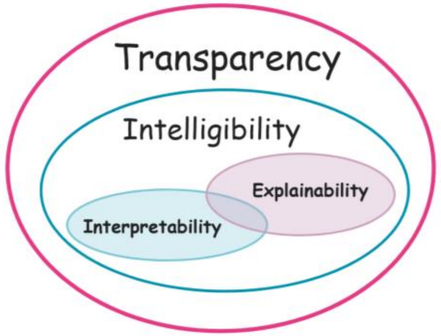
Amina Adadi and Mohammed Berrada. 2018. Peeking Inside the Black-Box: A Survey on Explainable Artificial Intelligence (XAI). IEEE Access, 6:52138–52160.


Decision-Making
Probability
Black-box
Using
Systems
Intelligibility
Different
Interpretability
Learning
Transparency
Process
Model
Life
Data
Will
Interpretability
Defined
Models
Box
Reasoning
Terms
Figure
Paper
Neural
Networks
Possible
Two
Impact
Explanation
Models
XAI
System
Black
Can
Deep
Black
Deep
Used
Learning
E.G
Artificial
Systems
System
Research
Proposed
Transparency
Neural
Results
Decisions
Quality
Box
Network
Explainable
Explainability
Research
Trust
Model
Users
Explanations
Transparency
Networks

EDINBURGH CENTRE FOR ROBOTICS


Transparency
Intelligibility
Explainability
Interpretability

TRANSPARENCY



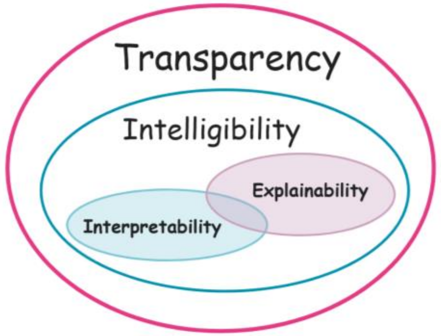
Dictionary definitions:

“clear and easy to understand” (Cambridge Dictionary)

“easily seen through, recognized, understood, detected; manifest, evident, obvious, clear” (Oxford English Dictionary)

“language or information that is transparent is clear and easy to understand” (The Longman Dictionary of Contemporary English)

TRANSPARENCY



2007

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.

2008

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.

2012

"**Transparency** clearly describing the model structure, equations, parameter values, and assumptions to enable interested parties to understand the model" (Briggs et al., 2012)

2016

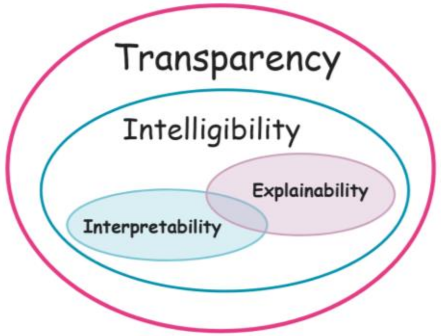
"Informally, **transparency** is the opposite of opacity or blackbox-ness. It connotes some sense of understanding the mechanism by which the model works. We consider transparency at the level of the model (simulatability), at the level of individual components (e.g.parameters) (decomposability), and at the level of the training algorithm (algorithmic transparency)" (Lipton,2016).

2018

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".

EDINBURGH CENTRE FOR ROBOTICS

INTELLIGIBILITY



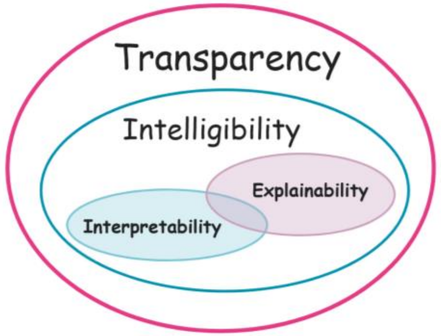
Dictionary definitions:

"clear enough to be understood" (Cambridge Dictionary)

"capable of being understood; comprehensible" (Oxford English Dictionary)

"easily understood" (The Longman Dictionary of Contemporary English)

INTELLIGIBILITY



2001

"**Intelligibility**; 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).

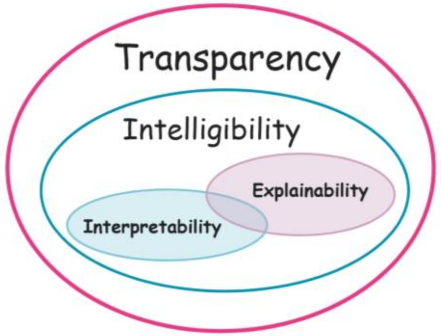
2009

"**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).

2018

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).

INTERPRETABILITY



Dictionary definitions:

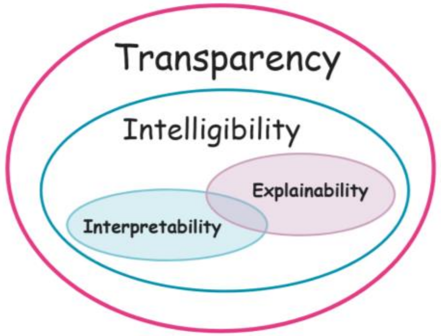
"to decide what the intended meaning of something is" (Cambridge Dictionary)

"clear or explicit; to elucidate; to explain" (Oxford English Dictionary)

"to explain the meaning of something " (The Longman Dictionary of Contemporary English)

EDINBURGH CENTRE FOR ROBOTICS

INTERPRETABILITY



2016

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)

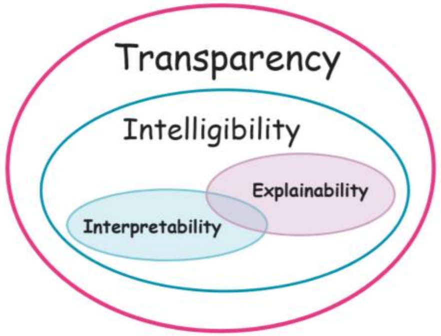
2018

“An **explanation** can be evaluated in two ways: according to its **interpretability**, and according to its completeness” (Gilpin et al., 2018).

2019

“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).

EXPLAINABILITY



Dictionary definitions:

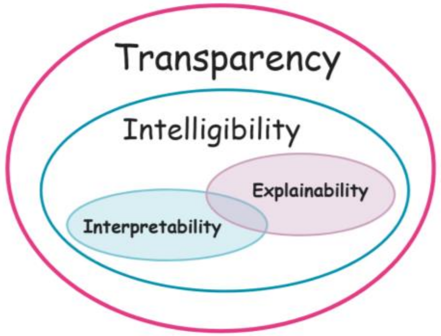
"to make something clear or easy to understand by describing or giving information about it" (Cambridge Dictionary)

"to provide an explanation for something. to make plain or intelligible" (Oxford English Dictionary)

"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)

EDINBURGH CENTRE FOR ROBOTICS

EXPLAINABILITY



2017

"Explanation is considered closely related to the concept of interpretability; Systems are interpretable if their operations can be understood by a human, either through introspection or through a produced explanation" (Biran and Cotton, 2017).

2018

"I equate interpretability with explainability" (Miller, 2018).

2018

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".

2019

"Transparent design: model is inherently interpretable (globally or locally)" (Lucic et al., 2019).

How can we learn from NLG ?

Natural Language Generation will be key to providing explanations, and rationalization.

XAI can learn how to structure and generate explanations from NLG (both rule-based and data-driven approaches).

A framework for evaluation of explanations is necessary, providing subjective and objective measures for transparency, interpretability etc., but also combined with traditional NLG metrics.

THANK YOU!!!

References

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), Randall Hill (Ed.). AAAI Press 900-907.

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

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.

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.

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 isporsmdm modeling good research practices task force-6. Value in Health, 15(6):835–842.

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.

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

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

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.

EDINBURGH CENTRE FOR ROBOTICS

References

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

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

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.

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.

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.

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

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.

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

EDINBURGH CENTRE FOR ROBOTICS