What Does Explainable AI Really Mean? A New Conceptualization of Perspectives

Derek Doran[1], Sarah Schulz[2], and Tarek R. Besold[3]

[1] Dept. of Computer Science & Engineering, Kno.e.sis Research Center
Wright State University, Dayton, Ohio, USA
derek.doran@wright.edu
[2] Institute for Natural Language Processing
University of Stuttgart, Stuttgart, Germany
sarah.schulz@ims.uni-stuttgart.de
[3] Department of Computer Science
City, University of London, London, UK
tarek-r.besold@city.ac.uk

Abstract. We characterize three notions of explainable AI that cut across research fields: *opaque systems* that offer no insight into its algorithmic mechanisms; *interpretable systems* where users can mathematically analyze its algorithmic mechanisms; and *comprehensible systems* that emit symbols enabling user-driven explanations of how a conclusion is reached. The paper is motivated by a corpus analysis of NIPS, ACL, COGSCI, and ICCV/ECCV paper titles showing differences in how work on explainable AI is positioned in various fields. We close by introducing a fourth notion: truly *explainable systems*, where automated reasoning is central to output crafted explanations without requiring human post processing as final step of the generative process.

1 Introduction

If you were held accountable for the decision of a machine in contexts that have financial, safety, security, or personal ramifications to an individual, would you blindly trust its decision? How can we hold accountable Artificial Intelligence (AI) systems that make decisions on possibly unethical grounds, e.g. when they predict a person's weight and health by their social media images [8] or the world region they are from [7] as part of a downstream determination about their future, like when they will quit their job [12], commit a crime [4], or could be radicalized into terrorism [1]? It is hard to imagine a person who would feel comfortable in blindly agreeing with a system's decision in such highly consequential and ethical situations without a deep understanding of the decision making rationale of the system. To achieve complete trustworthiness and an evaluation of the ethical and moral standards of a machine [15], detailed "explanations" of AI decisions seem necessary. Such explanations should provide insight into the rationale the AI uses to draw a conclusion. Yet many analysts indeed blindly 'accept' the outcome of an AI, whether by necessity or by choice.

To overcome this dangerous practice, it is prudent for an AI to provide not only an output, but also a human understandable explanation that expresses the rationale of the machine. Analysts can turn to such explanations to evaluate if a decision is reached by rational arguments and does not incorporate reasoning steps conflicting with ethical or legal norms.

But what constitutes an explanation? The Oxford English dictionary has no entry for the term 'explainable', but has one for *explanation*: *A statement or account that makes something clear; a* **reason** *or justification given for an action or belief.* Do present systems that claim to make 'explainable' decisions really provide explanations? Those who argue yes may point to Machine Learning (ML) algorithms that produce rules about data features to establish a classification decision, such as those learned by decision trees [14]. Others suggest that rich visualizations or text supplied along with a decision, as is often done in deep learning for computer vision [16,5,6], offer sufficient information to draw an explanation of why a particular decision was reached. Yet "rules" merely shed light into *how*, not *why*, decisions are made, and supplementary artifacts of learning systems (e.g. annotations and visualizations) require human-driven post processing under their own line of reasoning. The variety of ways "explanations" are currently handled is well summarized by Lipton [9] when he states that *"the term interpretability holds no agreed upon meaning, and yet machine learning conferences frequently publish papers which wield the term in a quasi-mathematical way"*. He goes on to call for engagement in the formulation of problems and their definitions to organize and advance explainable AI research. In this position paper, we respond to Lipton's call by proposing various "types" of explainable AI that cut across many fields of AI.

2 Existing Perspectives in Explainable AI

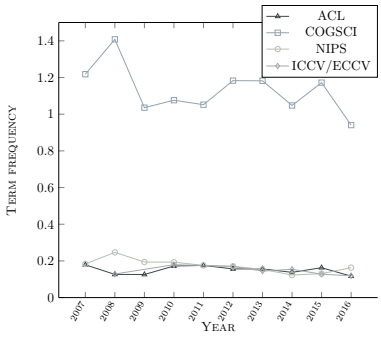

Fig. 1: Normalized corpus frequency of "explain" or "explanation".

As stated by Lipton, terms like *interpretability* are used in research papers, despite the lack of a clear and widely shared definition. In order to quantify this observation, we suggest a corpus-based analysis of relevant terms across research communities which strongly rely on ML-driven methodologies. The goal of this analysis is to gain insights into the relevance of explainability across AI-related research communities and to detect how each fields defines notions of explainability. We carried out an experiment over corpora of papers from the computer vision, NLP, connectionist, and symbolic reasoning communities. We base our analysis on corpus statistics compiled from the proceedings of conferences where researchers employ, inter alia, ML techniques to approach their research objectives: the Annual Meeting of the Association for Computational Linguistics (ACL), the Annual Conference

on Neural Information Processing Systems (NIPS), the International/European Conference on Computer Vision (ICCV/ECCV), and the Annual Conference of the Cognitive Science Society (COGSCI). The corpora include proceedings from 2007 to 2016. This allows us to observe trends regarding the use of words related to various concepts and the scope these concepts take. We perform a shallow linguistic search for what we henceforth call "explanation terms". We apply simple substring matches such as "explain" or "explanation" for explainability, "interpret" for interpretability and "compreh" for comprehensibility. "Explanation terms" serve as an approximation to aspects of explainability. Frequencies are normalized, making them comparable between years and conferences.

The normalized frequencies for explanation terms are plotted in Figure 1. We omit frequency plots for interpretation and comprehension terms because they exhibit a similar pattern. The frequency level of explainability concepts for COGSCI is significantly above those of the other corpora. This could be due to the fact that Cognitive Science explicitly aims to explain the mind and its processes, in many cases leading to qualitatively different research questions, corresponding terminology, and actively used vocabularies. The NIPS corpus hints at an emphasis of explainability in 2008 and a slight increase in interest in this concept in 2016 in the connectionist community. To better understand how consistent topics and ideas around explainability are across fields, we also analyze the context of its mentions. Word clouds shown in Figure 2 are a simple method to gain an intuition about the composition of a concept and its semantic contents by highlighting the important words related to it. Important words are defined as words that appear within a 20 words window of a mention of an explanation term with a frequency highly above average[4].

All communities focus on the explainability of a *model* but there is a difference between the nature of models in Cognitive Science and the other fields. The COGSCI corpus mentions a *participant*, a *task* and an *effect* whereas the other communities focus on a *model* and what constitutes *data* in their fields. It is not surprising that the neural modeling and NLP communities show a large overlap in their usage of explainability since there is an overlap in the research communities as well. We note further differences across the three ML communities compared to COGSCI. In the ACL corpus, explainability is often paired with words like features, examples, and words, which could suggest an emphasis on using examples to demonstrate the decision making of NLP decisions and the relevance of particular features. In the NIPS corpus, explainability is more closely tied to methods, algorithms, and results suggesting a desire to establish explanations about how neural systems translate inputs to outputs. The ICCV/ECCV falls between the ACL and NIPS corpus in the sense that it pairs explainability with data (images) and features (objects) like ACL, but may also tie the notion to how algorithms use (using) images to generate outputs.

The corpus analysis establishes some differences in how various AI communities approach the concept of explainability. In particular, we note that the term

[4] Word clouds are generated with the word-cloud package (`http://amueller.github.io/word_cloud/index.html`).

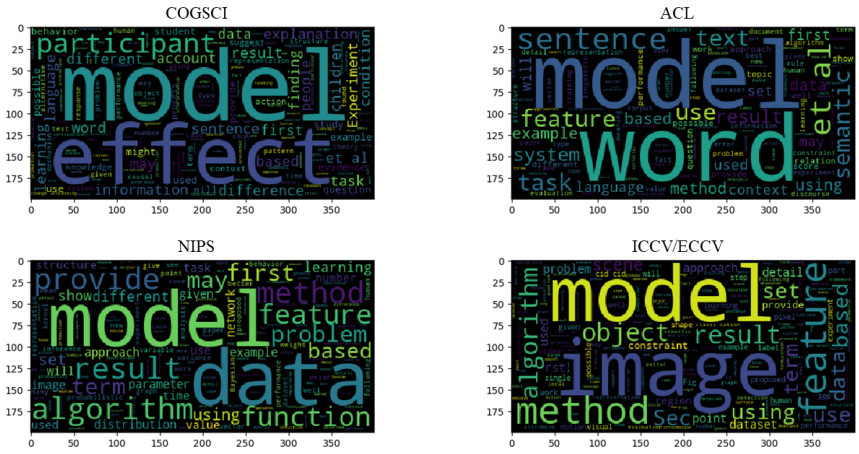

Fig. 2: Word clouds of the context of explanation terms in the different proceedings corpora.

is sometimes used to help probe the mechanisms of ML systems (e.g. we seek an *interpretation* of how the system works), and other times to relate explanations to particular inputs and examples (e.g. we want to *comprehend* how an input was mapped to an output). We use these observations to develop the following notions, also illustrated in Figure 3:

Opaque systems. A system where the mechanisms mapping inputs to outputs are invisible to the user. It can be seen as an oracle that makes predictions over an input, without indicating how and why predictions are made. Opaque systems emerge, for instance, when closed-source AI is licensed by an organization, where the licensor does not want to reveal the workings of its proprietary AI. Similarly, systems relying on genuine "black box" approaches, for which inspection of the algorithm or implementation does not give insight into the system's actual reasoning from inputs to corresponding outputs, are classified as opaque.

Interpretable systems. A system where a user cannot only see, but also study and understand how inputs are mathematically mapped to outputs. This implies model *transparency*, and requires a level of understanding of the technical details of the mapping. A regression model can be interpreted by comparing covariate weights to realize the relative importance of each feature to the mapping. SVMs and other linear classifiers are interpretable insofar as data classes are defined by their location relative to decision boundaries. But the action of deep neural networks, where input features may be automatically learned and transformed through non-linearities, is unlikely to be interpretable by most users.

Comprehensible systems. A comprehensible system emits symbols along with its output (echoing Michie's *strong* and *ultra-strong machine learning* [11]). These symbols (most often words, but also visualizations, etc.) allow the user

to relate properties of the inputs to their output. The user is responsible for compiling and comprehending the symbols, relying on her own implicit form of knowledge and reasoning about them. This makes comprehensibility a graded notion, with the degree of a system's comprehensibility corresponding to the relative ease or difficulty of the compilation and comprehension. The required implicit form of knowledge on the side of the user is often an implicit cognitive "intuition" about how the input, the symbols, and the output relate to each other. Taking the image in Figure 3 as example, it is intuitive to think that users will comprehend the symbols by noting that they represent objects observed in the image, and that the objects may be related to each other as items often seen in a factory. Different users may have different tolerances in their comprehension: some may be willing to draw arbitrary relationships between objects while others would only be satisfied under a highly constrained set of assumptions.

3 Defining Notions of Explainability

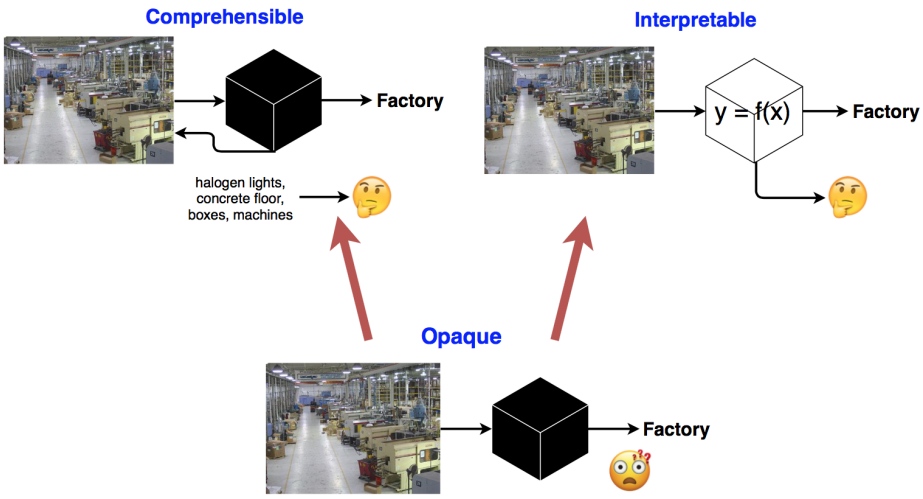

Fig. 3: Relation between opaque, comprehensible, and interpretable AI.

The arrows in Figure 3 suggest that comprehensible and interpretable systems each are improvements over opaque systems. The notions of comprehension and interpretation are separate: while interpretation requires transparency in the underlying mechanisms of a system, a comprehensible one can be opaque while emitting symbols a user can reason over. Regression models, support vector machines, decision trees, ANOVAs, and data clustering (assuming a kernel that is itself interpretable) are common examples of interpretable models. High dimensional data visualizations like t-SNE [10] and receptive field visualization on convolutional neural networks [2] are examples of comprehensible models.

It is important that research in both interpretable and comprehensible systems continue forward. This is because, depending on the user's background and her purpose of employing an AI model, one type is preferable to another. As a real-life example of this, most people think of a doctor as a kind of black box that transforms symptoms and test results into a diagnosis. Without providing information about the way medical tests and evaluations work, doctors deliver a diagnosis to a patient by explaining high-level indicators revealed in the tests (i.e. system symbols). Thus, when facing a patient, the physician should be like a comprehensible model. When interacting with other doctors and medical staff, however, the doctor may be like an interpretable model: She can sketch a technical line of connecting patient symptoms and test results to a particular diagnosis. Other doctors and staff can interpret a diagnosis in the same way that an analyst can interpret an ML model, ensuring that the conclusions drawn are supported by reasonable evaluation functions and weight values for the evidence presented.

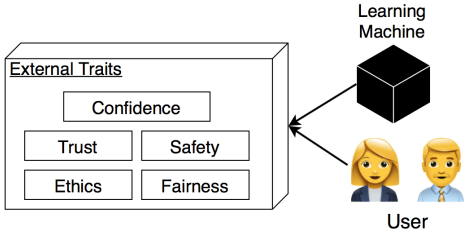

Fig. 4: External traits of a machine related to explainable AI. The traits depend not only on properties of the learning machine, but also the user. For example, confidence in an interpretable learning system is a function of the user's capability to understand the machine's input/output mapping behavior.

Explainable system traits. Often discussed alongside explainable AI are the external traits such systems should exhibit. These traits are seen as so important that some authors argue an AI system is not 'explainable' if it does not support them [9]. Figure 4 presents such traits and conveys their dependence on not only the learning model but also the user. For example, explainable AI should instill confidence and trust that the model operates accurately. Yet the perception of trust is moderated by a user's internal bias for or against AI systems, and their past experiences with their use. Safety, ethicality, and fairness are traits that can only be evaluated by a user's understanding of societal standards and by her ability to reason about emitted symbols or mathematical actions. Present day systems fortunately leave this reasoning to the user, keeping a person as a stopgap preventing unethical or unfair recommendations from being acted upon.

We also note that "completeness" is not an explicit trait, and might not even be desirable as such. Continuing with the doctor example from above, it may be desirable for a system to present a simplified (in the sense of incomplete, as opposed to abstracted) 'explanation' similar to a doctor using a patient's incomplete and possibly not entirely accurate preconceptions in explaining a complex diagnosis, or even sparing the patient especially worrisome details which might not be relevant for the subsequent treatment.

4 Truly Explainable AI Should Integrate Reasoning

Interpretable and comprehensible models encompass much of the present work in explainable AI. Yet we argue that both approaches are lacking in their ability to formulate, for the user, a line of *reasoning* that explains the decision making process of a model *using human-understandable features of the input data*. Reasoning is a critical step in formulating an explanation about why or how some event has occurred (see, e.g., Figure 5). Leaving explanation generation to human analysts can be dangerous since, depending on their background knowledge about the data and its domain, different explanations about why a model makes a decision may be deduced. Interpretive and comprehensible models thus *enable* explanations of decisions, but do not yield explanations themselves.

Efforts in neural-symbolic integration [3] aim to develop methods which might enable explicit automated reasoning over model properties and decision factors by extracting symbolic rules from connectionist models. Combining their results with work investigating factors influencing the *human comprehensibility* of representation formats and reasoning approaches [13] might pave the way towards systems effectively providing full explanations of their own to their users.

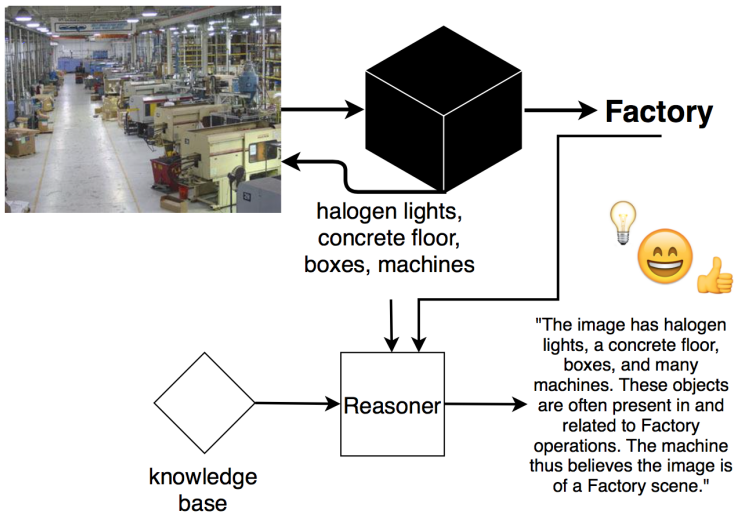

Fig. 5: Augmenting comprehensible models with a reasoning engine. This engine can combine symbols emitted by a comprehensible machine with a (domain specific) knowledge base encoding relationships between concepts represented by the symbols. The relationships between symbols in the knowledge based can yield a logical deduction about their relationship to the machine's decision.

Acknowledgement. The authors thank the Schloss Dagstuhl – Leibniz Center for Informatics and organizers and participants of Dagstuhl Seminar 17192 on Human-Like Neural-Symbolic Computing for providing the environment to develop the ideas in this paper. This work is partially supported by a Schloss Dagstuhl travel grant and by the Ohio Federal Research Network. Parts of the work have been carried out at the Digital Media Lab of the University of Bremen.

References

1. Al Hasan, M., Chaoji, V., Salem, S., Zaki, M.: Link prediction using supervised learning. In: Wkshp. on link analysis, counter-terrorism and security (2006)
2. Bau, D., Zhou, B., Khosla, A., Oliva, A., Torralba, A.: Network Dissection: Quantifying Interpretability of Deep Visual Representations. In: Proc. of Computer Vision and Pattern Recognition (2017)
3. Garcez, A.d., Besold, T.R., De Raedt, L., Földiak, P., Hitzler, P., Icard, T., Kühnberger, K.U., Lamb, L.C., Miikkulainen, R., Silver, D.L.: Neural-symbolic learning and reasoning: contributions and challenges. In: Proceedings of the AAAI Spring Symposium on Knowledge Representation and Reasoning: Integrating Symbolic and Neural Approaches, Stanford (2015)
4. Gerber, M.S.: Predicting crime using Twitter and kernel density estimation. Decision Support Systems 61, 115–125 (2014)
5. Johnson, J., Karpathy, A., Fei-Fei, L.: Densecap: Fully convolutional localization networks for dense captioning. In: Proc. of Computer Vision and Pattern Recognition. pp. 4565–4574 (2016)
6. Karpathy, A., Fei-Fei, L.: Deep visual-semantic alignments for generating image descriptions. In: Proc. of Conference on Computer Vision and Pattern Recognition. pp. 3128–3137 (2015)
7. Katti, H., Arun, S.: Can you tell where in India I am from? Comparing humans and computers on fine-grained race face classification. arXiv preprint arXiv:1703.07595 (2017)
8. Kocabey, E., Camurcu, M., Ofli, F., Aytar, Y., Marin, J., Torralba, A., Weber, I.: Face-to-bmi: Using computer vision to infer body mass index on social media. arXiv preprint arXiv:1703.03156 (2017)
9. Lipton, Z.C.: The mythos of model interpretability. Workshop on Human Interpretability in Machine Learning (2016)
10. Maaten, L.v.d., Hinton, G.: Visualizing data using t-SNE. Journal of Machine Learning Research 9, 2579–2605 (2008)
11. Michie, D.: Machine learning in the next five years. In: Proc. of the Third European Working Session on Learning. pp. 107–122. Pitman (1988)
12. Saradhi, V.V., Palshikar, G.K.: Employee churn prediction. Expert Systems with Applications 38(3), 1999–2006 (2011)
13. Schmid, U., Zeller, C., Besold, T.R., Tamaddoni-Nezhad, A., Muggleton, S.: How Does Predicate Invention Affect Human Comprehensibility? Inductive Logic Programming: ILP 2016 Revised Selected Papers pp. 52–67 (2017)
14. Shafiq, M.Z., Erman, J., Ji, L., Liu, A.X., Pang, J., Wang, J.: Understanding the impact of network dynamics on mobile video user engagement. In: ACM SIGMETRICS Performance Evaluation Review. pp. 367–379 (2014)
15. Skirpan, M., Yeh, T.: Designing a Moral Compass for the Future of Computer Vision using Speculative Analysis. In: Proc. of Computer Vision and Pattern Recognition (2017)
16. You, Q., Jin, H., Wang, Z., Fang, C., Luo, J.: Image captioning with semantic attention. In: Proc. of Computer Vision and Pattern Recognition. pp. 4651–4659 (2016)