What makes a good explanation? Cognitive dimensions of explaining intelligent machines

Roberto Confalonieri, Tarek R. Besold
(name.surname@telefonica.com)
Alpha Health AI Lab
Telefónica Innovación Alpha

Tillman Weyde
(t.e.weyde@city.ac.uk)
Dept. of Computer Science
City, University of London

Kathleen Creel
(k.creel@pitt.edu)
Dept. of History and Philosophy of Science
University of Pittsburgh

Tania Lombrozo
(lombrozo@princeton.edu)
Dept. of Psychology
Princeton University

Shane Mueller
(shanem@mtu.edu)
Cognitive and Learning Sciences
Michigan Technological University

Patrick Shafto
(patrick.shafto@gmail.com)
Dept. of Math. and Computer Science
Rutgers University

Keywords: Explainability; Artificial Intelligence; Philosophy of Artificial Intelligence; Psychology; Cognitive Science

Explainability is assumed to be a key factor for the adoption of Artificial Intelligence systems in a wide range of contexts (Hoffman, Mueller, & Klein, 2017; Hoffman, Mueller, Klein, & Litman, 2018; Doran, Schulz, & Besold, 2017; Lipton, 2018; Miller, 2017; Lombrozo, 2016). The use of AI components in self-driving cars, medical diagnosis, or insurance and financial services has shown that when decisions are taken or suggested by automated systems it is essential for practical, social, and increasingly legal reasons that an explanation can be provided to users, developers or regulators.[1]Moreover, the reasons for equipping intelligent systems with explanation capabilities are not limited to user rights and acceptance. Explainability is also needed for designers and developers to enhance system robustness and enable diagnostics to prevent bias, unfairness and discrimination, as well as to increase trust by all users in *why* and *how* decisions are made. Against that background, increased efforts are directed towards studying and provisioning explainable intelligent systems, both in industry and academia, sparked by initiatives like the DARPA Explainable Artificial Intelligence Program (DARPA, 2016). In parallel, scientific conferences and workshops dedicated to explainability are now regularly organised, such as the 'ACM Conference on Fairness, Accountability, and Transparency (ACM FAT)' (Friedler & Wilson, n.d.) or the 'Workshop on Explainability in AI' at the 2017 and 2018 editions of the International Joint Conference on Artificial Intelligence.

However, one important question remains hitherto unanswered: *What are the criteria for a good explanation?*

Explainable Artificial Intelligence

While Explainable Artificial Intelligence (XAI) has recently received significant attention, its origins stem from several decades ago when AI systems were mainly developed as knowledge-based or expert systems, such as in MYCIN (Buchanan & Shortliffe, 1984) and NEOMYCIN (Hasling, Clancey, & Rennels, 1984). In these systems, explanations were conceived mainly as reasoning traces of the system — at first resulting in a very technical notion of what an explanation is, with only limited regard to cognitive aspects on the user's side. Still, in the context of REX (Wick & Thompson, 1992), there was already a discussion of how to adapt explanations to different user groups and the trade-offs involved. While interest in XAI subsided after the mid-1990s, recent successes in machine learning technology have brought explainability back into the focus. This has led to a plethora of new approaches for both autonomous and humans-in-the-loop systems, aiming to achieve explainability, as defined by respective system creators, without sacrificing system performance.

Many systems focus on interpretable *post-hoc* approximations of black-box models (Guidotti et al., 2018), using symbolic representations such as decision trees (Craven, 1996; Sarkar et al., 2016) or decision rules (Ribeiro, Singh, & Guestrin, 2018), feature importance (Lou, Caruana, & Gehrke, 2012), saliency maps (Selvaraju et al., 2017), or local regression models (Ribeiro, Singh, & Guestrin, 2016). On the other hand, there are efforts to design intelligent systems to be interpretable by design, e.g., in recommender systems (Zhang & Chen, 2018), or in a recently started project developing the concept of *perspicuous computing*.[2]

In these heterogeneous origins and developments of XAI, a discussion is still to be had on what precisely the roles of explanations are and, in particular, what makes an explanation a good explanation. To this end, we will bring together several experts of different aspects of the phenomenon "explanation" in this symposium, to analyze the notion of explanation in the context of artificial intelligence from different cognition-related perspectives.

What Makes a Good Explanation?

Starting out from the cognition of explanations, this symposium will foster scientific discourse about what

[1]As a case in point, the European Union's General Data Protection Regulation (GDPR) stipulates a right to "*meaningful information about the logic involved*"— commonly interpreted as a 'right to an explanation'— for consumers affected by an automatic decision (Parliament and Council of the European Union, 2016).

[2]https://www.perspicuous-computing.science

functions an explanation needs to fulfill and the criteria that define its quality. Some of the aspects to be addressed are:

- Objective and subjective value of explanations
- Dimensions of explanations: complete vs compact, abstract vs concrete, reduced vs simplified, ...
- Anchoring to known concepts
- Counter-factual explanations and actionability
- Personalisation
- Legal requirements
- Grounding in personal and social experience and intuition

A panel of recognised scholars and researchers will bring insights and expertise from different points of view, including psychology, cognitive science, computer science, and philosophy, and will foster knowledge exchange and discussion of the multiple facets of explanation:

- Kathleen Creel will talk about 'Understanding Machine Science: XAI and Scientific Explanations', drawing on the literature on scientific explanation in philosophy and cognitive science, and arguing that for scientific researchers, good explanations require more access to the functional structure of the intelligent system than is needed by other human users.
- Tania Lombrozo will talk about 'Explanatory Virtue & Vices', considering the multiple functions and malfunctions of human explanatory cognition with implications for XAI. In particular, she will suggest that we need to differentiate between different possible goals for explainability, and that doing so it highlights why human explanatory cognition should be a crucial constraint on design.
- Shane Mueller will talk about 'Ten fallacies of Explainable Artificial Intelligence', reviewing some of the assumptions made until now about what properties lead to good explanations, and describing how each constitutes a fallacy that might backfire if used for developing XAI systems. He will then describe a framework developed for the DARPA XAI Program for measuring the impact of explanations that incorporates cognitive science theory related to mental models, sensemaking, context, trust, and self-explanation that can provide a principled approach for developing explainable systems.
- Patrick Shafto will talk about 'XAI via Bayesian Teaching', raising questions about the use of modern machine learning algorithms in societally important processes, and theoretical questions about whether and how the opaqueness of these algorithms can be ameliorated, in the framework of Bayesian teaching.
- Roberto Confalonieri and Tillman Weyde will talk about 'An Ontology-based Approach to Explaining Artificial Neural Networks', addressing the challenges of extracting symbolic representations from neural networks, exploiting domain knowledge, and measuring understandability of decision trees with users both objectively and subjectively.

References

Buchanan, B. G., & Shortliffe, E. H. (1984). *The Mycin Experiments of the Stanford Heuristic Programming Project*. Addison-Wesley Longman Publishing Co., Inc.

Craven, M. W. (1996). *Extracting comprehensible models from trained neural networks*. (Ph.D. Thesis)

DARPA. (2016). *Explainable AI - program.*

Doran, D., Schulz, S., & Besold, T. R. (2017). What does explainable AI really mean? A new conceptualization of perspectives. In *CEUR Workshop Proc.* (Vol. 2071).

Friedler, S. A., & Wilson, C. (Eds.). (n.d.). *Proceedings of machine learning research* (Vol. 81). PMLR.

Guidotti, R., Monreale, A., Ruggieri, S., Turini, F., Giannotti, F., & Pedreschi, D. (2018). A survey of methods for explaining black box models. *Comp. Surv.*, *51*(5), 1–42.

Hasling, D. W., Clancey, W. J., & Rennels, G. (1984). Strategic explanations for a diagnostic consultation system. *Int. Journal of Man-Machine Studies*, *20*(1), 3 - 19.

Hoffman, R. R., Mueller, S. T., & Klein, G. (2017). Explaining explanation, part 2: Empirical foundations. *IEEE Intelligent Systems*, *32*(4), 78-86.

Hoffman, R. R., Mueller, S. T., Klein, G., & Litman, J. (2018). Metrics for explainable AI: challenges and prospects. *CoRR*, *abs/1812.04608*.

Lipton, Z. C. (2018, June). The mythos of model interpretability. *Queue*, *16*(3), 30:31–30:57.

Lombrozo, T. (2016). Explanatory preferences shape learning and inference. *Trends in CogSci*, *20*(10), 748-759.

Lou, Y., Caruana, R., & Gehrke, J. (2012). Intelligible models for classification and regression. In *Proc. of the 18th ACM KDD* (pp. 150–158). ACM.

Miller, T. (2017). Explanation in artificial intelligence: Insights from the social sciences. *CoRR*, *abs/1706.07269*.

Parliament and Council of the European Union. (2016). *General Data Protection Regulation.*

Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). "Why Should I Trust You?": Explaining the Predictions of Any Classifier. In *Proc. of the 22nd Int. Conf. on Knowledge Discovery and Data Mining* (pp. 1135–1144). ACM.

Ribeiro, M. T., Singh, S., & Guestrin, C. (2018). Anchors: High-precision model-agnostic explanations. In *AAAI* (pp. 1527–1535). AAAI Press.

Sarkar, S., Weyde, T., Garcez, A., Slabaugh, G. G., Dragicevic, S., & Percy, C. (2016). Accuracy and interpretability trade-offs in machine learning applied to safer gambling. In *CEUR Workshop Proc.* (Vol. 1773).

Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., & Batra, D. (2017). Grad-cam: Visual explanations from deep networks via gradient-based localization. In *ICCV* (pp. 618–626).

Wick, M. R., & Thompson, W. B. (1992, March). Reconstructive expert system explanation. *Artificial Intelligence*, *54*(1-2), 33–70.

Zhang, Y., & Chen, X. (2018). Explainable recommendation: A survey and new perspectives. *CoRR*, *abs/1804.11192*.