ARTIFICIAL INTELLIGENCE

XAI—Explainable artificial intelligence

David Gunning[1]*[†], Mark Stefik[2], Jaesik Choi[3], Timothy Miller[4], Simone Stumpf[5], Guang-Zhong Yang[6][†]

Explainability is essential for users to effectively understand, trust, and manage powerful artificial intelligence applications.

Recent successes in machine learning (ML) have led to a new wave of artificial intelligence (AI) applications that offer extensive benefits to a diverse range of fields. However, many of these systems are not able to explain their autonomous decisions and actions to human users. Explanations may not be essential for certain AI applications, and some AI researchers argue that the emphasis on explanation is misplaced, too difficult to achieve, and perhaps unnecessary. However, for many critical applications in defense, medicine, finance, and law, explanations are essential for users to understand, trust, and effectively manage these new, artificially intelligent partners [see recent reviews (1–3)].

Recent AI successes are largely attributed to new ML techniques that construct models in their internal representations. These include support vector machines (SVMs), random forests, probabilistic graphical models, reinforcement learning (RL), and deep learning (DL) neural networks. Although these models exhibit high performance, they are opaque in terms of explainability. There may be inherent conflict between ML performance (e.g., predictive accuracy) and explainability. Often, the highest performing methods (e.g., DL) are the least explainable, and the most explainable (e.g., decision trees) are the least accurate. Figure 1 illustrates this with a notional graph of the performance-explainability tradeoff for some of the ML techniques.
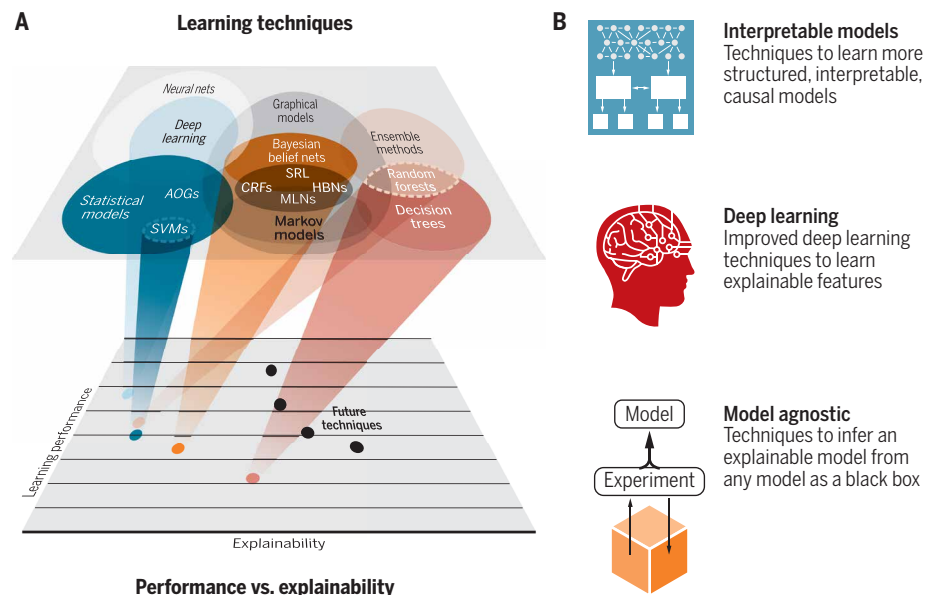
WHAT IS XAI?

The purpose of an explainable AI (XAI) system is to make its behavior more intelligible to humans by providing explanations. There are some general principles to help create effective, more human-understandable AI systems: The XAI system should be able to explain its capabilities and understandings; explain what it has done, what it is doing now, and what will happen next; and disclose the salient information that it is acting on (4).

However, every explanation is set within a context that depends on the task, abilities, and expectations of the user of the AI system. The definitions of interpretability and explainability are, thus, domain dependent and may not be defined independently from a domain. Explanations can be full or partial. Models that are fully interpretable give full and completely transparent explanations. Models that are partially interpretable reveal important pieces of their reasoning process. Interpretable models obey "interpretability constraints" that are defined according to the domain (e.g., monotonicity with respect to certain variables and correlated variables obey particular relationships), whereas black box or unconstrained models do not necessarily obey these constraints. Partial explanations may include variable importance measures, local models that approximate global models at specific points and saliency maps.

EXPECTATION FROM USERS

XAI assumes that an explanation is provided to an end user who depends on the decisions, recommendations, or actions produced by an AI system, yet there could be many different

[1]Defense Advanced Research Projects Agency (DARPA), 675 North Randolph Street, Arlington, VA 22201, USA. [2]Palo Alto Research Center, 3333 Coyote Hill Road, Palo Alto, CA 94304, USA. [3]Graduate School of Artificial Intelligence, Korea Advanced Institute of Science and Technology, 291 Daehak-ro, Yuseong-gu, Daejeon 34141, Republic of Korea. [4]School of Computing and Information Systems, University of Melbourne, Victoria 3010, Australia. [5]Centre for HCI Design, School of Mathematics, Computer Science and Engineering, City, University of London, London EC1V 0HB, UK. [6]Institute of Medical Robotics, Shanghai Jiao Tong University, Shanghai, China.
*Present address: Facebook AI Research, 770 Broadway, New York, NY 10003, USA.
[†]Corresponding author. Email: dgunning@fb.com (D.G.); gzyang@sjtu.edu.cn (G.-Z.Y.)

Fig. 1. Performance versus explainability tradeoff for ML techniques. (**A**) Learning techniques and explainability. Concept adapted from (9). (**B**) Interpretable models: ML techniques that learn more structured, interpretable, or causal models. Early examples included Bayesian rule lists, Bayesian program learning, learning models of causal relationships, and using stochastic grammars to learn more interpretable structure. Deep learning: Several design choices might produce more explainable representations (e.g., training data selection, architectural layers, loss functions, regularization, optimization techniques, and training sequences). Model agnostic: Techniques that experiment with any given ML model, as a black box, to infer an approximate explainable model.

kinds of users, often at different time points in the development and use of the system (*5*). For example, a type of user might be an intelligence analyst, a judge, or an operator. However, other users who demand an explanation of the system may be developers or test operators who need to understand where there might be areas of improvement. Yet, another user might be policy makers, who are trying to assess the fairness of the system. Each user group may have a preferred explanation type that is able to communicate information in the most effective way. An effective explanation will take the target user group of the system into account, who might vary in their background knowledge and needs for what should be explained.

EXPLAINABILITY—EVALUATION AND MEASUREMENT

A number of ways of evaluating and measuring the effectiveness of an explanation have been proposed; however, there is currently no common means of measuring whether an XAI system is more intelligible to a user than a non-XAI system. Some of these measures are subjective measures from the user's point of view, such as user satisfaction, which can be measured through a subjective rating of the clarity and utility of an explanation. More objective measures for an explanation's effectiveness might be task performance; i.e., does the explanation improve the user's decision-making? Reliable and consistent measurement of the effects of explanations is still an open research question. Evaluation and measurement for XAI systems include evaluation frameworks, common ground [different thinking and mutual understanding (*6*)], common sense, and argumentation [why (*7*)].

XAI—ISSUES AND CHALLENGES

There remain many active issues and challenges at the intersection of ML and explanation.

1) Starting from computers versus starting from people (*8*). Should XAI systems tailor explanations to particular users? Should they consider the knowledge that users lack? How can we exploit explanations to aid interactive and human-in-the-loop learning, including enabling users to interact with explanations to provide feedback and steer learning?

2) Accuracy versus interpretability. A major thread of XAI research on explanation explores techniques and limitations of interpretability. Interpretability needs to consider tradeoffs involving accuracy and fidelity and to strike a balance between accuracy, interpretability, and tractability.

3) Using abstractions to simplify explanations. High-level patterns are the basis for describing big plans in big steps. Automating the discovery of abstractions has long been a challenge, and understanding the discovery and sharing of abstractions in learning and explanation are at the frontier of XAI research today.

4) Explaining competencies versus explaining decisions. A sign of mastery by highly qualified experts is that they can reflect on new situations. It is necessary to help end users to understand the competencies of the AI systems in terms of what competencies a particular AI system has, how the competencies should be measured, and whether an AI system has blind spots; that is, are there classes of solutions it can never find?

From a human-centered research perspective, research on competencies and knowledge could take XAI beyond the role of explaining a particular XAI system and helping its users to determine appropriate trust. In the future, XAIs may eventually have substantial social roles. These roles could include not only learning and explaining to individuals but also coordinating with other agents to connect knowledge, developing cross-disciplinary insights and common ground, partnering in teaching people and other agents, and drawing on previously discovered knowledge to accelerate the further discovery and application of knowledge. From such a social perspective of knowledge understanding and generation, the future of XAI is just beginning.

REFERENCES AND NOTES

1. W. Samek, G. Montavon, A. Vedaldi, L. K. Hansen, K. R. Muller, *Explainable AI: Interpreting, Explaining and Visualizing Deep Learning* (Springer Nature, 2019).
2. H. J. Escalante, S. Escalera, I. Guyon, X. Baró, Y. Güçlütürk, U. Güçlü, M. van Gerven, *Explainable and Interpretable Models in Computer Vision and Machine Learning* (Springer, 2018).
3. O. Biran, C. Cotton, Explanation and justification in machine learning: A survey, paper presented at the IJCAI-17 Workshop on Explainable AI (XAI), Melbourne, Australia, 20 August 2017.
4. V. Bellotti, K. Edwards, Intelligibility and accountability: Human considerations in context-aware systems. *Hum. Comput. Interact.* **16**, 193–212 (2009).
5. T. Kulesza, M. Burnett, W. Wong, S. Stumpf, Principles of explanatory debugging to personalize interactive machine learning, in *Proceedings of the 20th International Conference on Intelligent User Interfaces* (ACM, 2015), pp. 126–137.
6. H. H. Clark, S. E. Brennan, Grounding in communication, in *Perspectives on Socially Shared Cognition*, L. B. Resnick, J. M. Levine, S. D. Teasley, Eds. (American Psychological Association, 1991), pp. 127–149.
7. D. Wang, Q. Yang, A. Abdul, B. Y. Lim, Designing theory-driven user-centric explainable AI, in *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (ACM, 2019), paper no. 601.
8. T. Miller, Explanation in artificial intelligence: Insights from the social sciences. *Artif. Intell.* **267**, 1–38 (2018).
9. D. Gunning, Explainable artificial intelligence (XAI), DARPA/I2O; www.cc.gatech.edu/~alanwags/DLAI2016/(Gunning)%20IJCAI-16%20DLAI%20WS.pdf.

Citation: D. Gunning, M. Stefik, J. Choi, T. Miller, S. Stumpf, G.-Z. Yang, XAI—Explainable artificial intelligence. *Sci. Robot.* **4**, eaay7120 (2019).

Science Robotics

XAI—Explainable artificial intelligence

David Gunning, Mark Stefik, Jaesik Choi, Timothy Miller, Simone Stumpf and Guang-Zhong Yang

ARTICLE TOOLS	http://robotics.sciencemag.org/content/4/37/eaay7120
RELATED CONTENT	http://robotics.sciencemag.org/content/robotics/4/37/eaaz8586.full http://robotics.sciencemag.org/content/robotics/4/37/eaay6276.full http://robotics.sciencemag.org/content/robotics/4/37/eaay4663.full
REFERENCES	This article cites 2 articles, 0 of which you can access for free http://robotics.sciencemag.org/content/4/37/eaay7120#BIBL
PERMISSIONS	http://www.sciencemag.org/help/reprints-and-permissions

Use of this article is subject to the Terms of Service