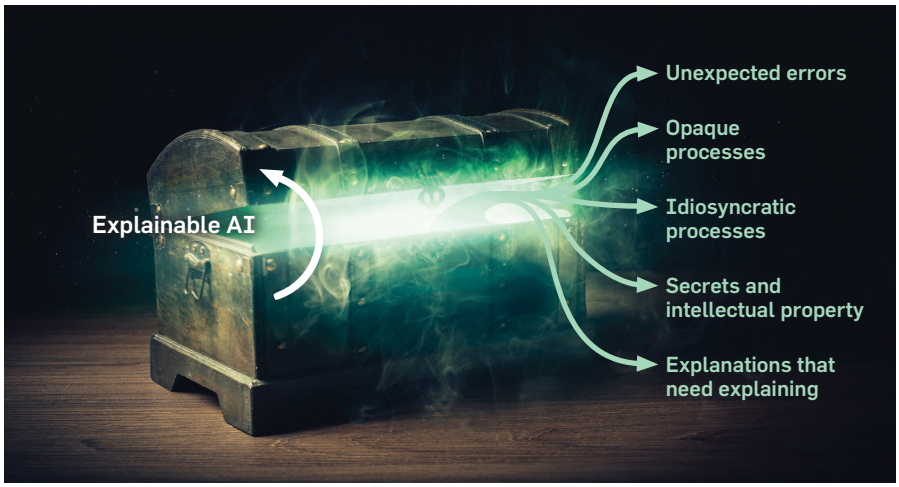
DOI:10.1145/3490699

Veda C. Storey, Roman Lukyanenko, Wolfgang Maass, and Jeffrey Parsons

Viewpoint
Explainable AI

Opening the black box or Pandora's Box?

Advances in AI, especially based on machine learning, have provided a powerful way to extract useful patterns from large, heterogeneous data sources. The rise in massive amounts of data, coupled with powerful computing capabilities, makes it possible to tackle previously intractable real-world problems. Medicine, business, government, and science are rapidly automating decisions and processes using machine learning. Unlike traditional AI approaches based on explicit rules expressing domain knowledge, machine learning often lacks explicit human-understandable specification of the rules producing model outputs. With growing reliance on automated decisions, an overriding concern is understanding the process by which "black box" AI techniques make decisions. This is known as the problem of explainable AI.[2] However, opening the black box may lead to unexpected consequences, as when opening Pandora's Box.



Pandora's Box of explainable AI.

Black Box of Machine Learning

Advanced machine learning algorithms, such as deep learning neural networks or support vector machines, are not easily understood by humans. Their power and success stems from the ability to generate highly complex decision models built upon hundreds of iterations over training data.[5] The performance of these models is dependent on many factors, including the availability and quality of training data and skills and domain expertise of data scientists. The complexity of machine learning models may be so great that even data scientists struggle to understand the underlying algorithms. For example, deep learning was used in the program that famously beat the reigning Go world champion,[6] yet the data scientists responsible could not always understand how or why the algorithms performed as they did.

Opening the black box involves providing human-understandable explanations for why a model reaches a decision and how it works.

Opening the black box involves providing human-understandable explanations for why a model reaches a decision and how it works. The motivation is to ensure decision making is justified, fair, and ethical, and to treat the "right to explanation" as a basic human right.[7] Notably, the European Union's General Data Protection Regulation requires companies to provide "meaningful information" about the logic in their programs (Article 13.2(f)). The goal is to ensure the rules, data, assumptions, and development processes underlying a machine learning model are understandable, transparent, and accessible to as many people as possible or necessary, including managers, users, customers, auditors, and citizens.

The explainable AI challenge usually focuses on how to open the black box of AI; for example, by considering

FIGURE USES IMAGE BY FER GREGORY

how various features contribute to the output of a model or by using counterfactual explanations that measure the extent to which a model output would change if a feature were missing.[7] We pose a seldom-asked, but vital, question: *Once a mechanism is in place to open the black box, how do we, as a society, prepare to deal with the consequences of exposing the reasoning that generates the output from AI models?*

Pandora's Box of Explainable AI

In Greek mythology, Pandora's Box refers to a container of evils that are unleashed and cannot be contained once the box is opened. We employ this analogy because, although opening the black box of AI may shed transparency on the machine learning model, it does not mean the processes underlying the model are free of problems. As in Pandora's Box, these problems are revealed once we move from a black box to a white box of AI. Machine learning explainability is a worthy goal; however, we must prepare for the outcome. Opening the black box can metaphorically open a Pandora's Box, as shown in the accompanying figure.

Pandora's Box of explainable AI is relevant to organizations, customers, governments, and citizens concerned with explainability, as well as to machine learning development teams. The problems may be grounded in flawed data or model designs, opaque or ill-defined organizational processes, secrets or sensitive information, and uncomfortable organizational truths.

Flawed data or model designs. Machine learning models are only as good as their training data. The sheer size of the training data used may prevent data scientists from fully assessing its quality. As organizations increasingly seek to integrate internal data with data from extra-organizational sources such as social media, evaluating the quality of such data becomes even more challenging. In addition, models are contingent on many decisions of data scientists, which may be flawed if developed by inexperienced teams or teams that lack deep domain knowledge. Then, errors can creep into the models and be unexpectedly revealed when subjected to public scrutiny.[4]

These issues plague even industry leaders in AI. For example, MD Anderson tested IBM Watson for its mission to eradicate cancer.[3] IBM's role was to enable clinicians to "uncover valuable insights" from rich patient and research databases. However, IBM engineers trained the software on hypothetical cancer patients rather than real ones. Medical specialists identified unsafe and incorrect treatment recommendations, including one to give a cancer patient with severe bleeding a drug that could worsen it.

By opening the black box of AI, organizations must prepare to take responsibility for the consequences of errors in their machine learning models. A troubling possibility is that customers or auditors might be first to discover flaws. Revealing such flaws can be embarrassing, undermine an organization's reputation, or even provoke sanctions and litigation. Increased attention to data management and the quality of machine learning model development is imperative.

Opaque or ill-defined organizational processes. Even perfect data would not solve the Pandora's Box problem. Data is only as good as the processes from which it was generated. Accurate and complete data might capture organizational reality, but the process by which decisions are made might be problematic. Some organizational processes are well-specified and managed; others might be based on tacit norms and deviant or improvised employee behavior. Organizations might not be fully aware of the exact practices that generate the training data upon which machine learning models are built.

AI explanations might reveal decisions are influenced by factors that do not align with explicit organizational policies. Amazon cancelled a plan to use AI to identify the best job candi-

Explainable AI, which aims to open the black box of machine learning, might also be a Pandora's Box.

dates for technology positions upon discovering the models were biased against women because the training data consisted predominantly of males, reflecting historic hiring practices.[a] If the process by which decisions are made remains a mystery, caution is needed when automating it. Explainable AI could reveal an uncomfortable or unacceptable reality. The general challenge is whether machine learning is an appropriate solution considering the kinds of organizational practices to be automated. Some tasks remain difficult for modern machine learning, such as automating unusual or exceptional cases. This especially affects smaller organizations, although larger, data-rich companies with stable routines are not completely immune, as the Amazon hiring case demonstrates.

Organizations must not consider machine learning a solution to all problems. In the age of white box (or transparent) AI, any misalignment between explicitly stated procedures and those actually implemented by machine learning are subject to public scrutiny. Organizations must understand and prepare for such possibilities, which might very well be the most important consequence of opening the Pandora's Box of AI.

Secrets or sensitive information. The drive to open the black box of machine learning models should be tempered by the risks associated with sensitive information spillage. Explainable AI could expose intellectual property or proprietary knowledge or breach privacy and confidentiality. Transparency could also lead to greater exposure of security vulnerabilities, requiring organizations to increase oversight of machine learning processes and continuously reflect on which parts of their operations can be automated.

At the same time, governments and policymakers should consider whether all, or only specific parts of, machine learning models can be subject to transparency and scrutiny. This requires research on how to ensure machine learning models remain effective and transparent while dealing with sensitive information. In the world of white box AI, organizations must consider transparency from the point of view of the receiver. Approaches such as differential privacy, which shares general insights but introduces noise to hide details on individuals, can be useful.[1]

Organizations must not consider machine learning a solution to all problems.

Uncomfortable organizational truths. Even when a model correctly captures organizational reality, and is based on accurate data, the unredacted presentation of the internal organizational logic might not align well with the expectations and needs of those affected by the decisions. For example, explainable AI might provide a technically correct explanation that could cause serious psychological harm. From machine learning output, a consultation with a very sick person might explain: *You will die within 30 days. I conclude this by analyzing Harvard's research library, EBSCO search engine, and patent files submitted by the pharma industry. This conclusion has 98% confidence.*

Intermediaries might be needed to translate and convey a message so it is more accepted by humans and adapted to individual needs. Organizations should consider the psychology, needs and values of potential recipients and reconcile technical explanations with explanations most appropriate for a situation or context. This might require adding relevant domain experts (such as doctors, psychologists) to a machine learning development team.

Implications

Explainable AI, which aims to open the black box of machine learning, might also be a Pandora's Box. Opening the black box might undermine trust in an organization and its decision-making processes by revealing truths about how processes are actually run, the limitations of an organization's data, or model defects. Organizations should prepare for explainable AI. They must foster good data management and machine learning practices to ensure high-quality data and models. They should carefully review their internal processes and ensure they are well understood and managed. Organizations must be prepared to change legal and communications strategies and respond to unexpected and unforeseen disclosure of operational practices.

Although fully opening the black box of AI may be many years away, it is prudent to prepare for the potential challenges. As the world continues to face large, complex problems, computer solutions will continue to be an effective means of tackling them. Thus, significant effort is required to understand the need for explainable AI, how to properly conduct it, and how to avoid its Pandora's Box effects. C

a "Amazon Scraps Secret AI Recruiting Tool That Showed Bias Against Women." Reuters (Oct. 9, 2018); https://reut.rs/3I3tkzy

References

1. Abadi, M. Deep learning with differential privacy. In *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security* (2016), 308–318.
2. Castelvecchi, D. Can we open the black box of AI? *Nature News 538*, 7623 (2016), 1–20.
3. Davenport, T.H. and Ronanki, R. Artificial intelligence for the real world. *Harvard Business Review 96*, 1 (2018), 108–116.
4. McCradden, M.D. et al. Ethical limitations of algorithmic fairness solutions in health care machine learning. *The Lancet Digital Health 2*, 5 (2020), e221–e223.
5. Mnih, V. et al. Human-level control through deep reinforcement learning. *Nature 518*, 7540 (2015), 529–533.
6. Silver, D. et al. Mastering the game of Go with deep neural networks and tree search. *Nature 529*, 7587 (2016), 484.
7. Wachter, S., Mittelstadt, B., and Russell, C. Counterfactual explanations without opening the black box: Automated decisions and the GDPR. *Harv. JL & Tech. 31* (2017), 841.

Veda C. Storey (vstorey@gsu.edu) is the Tull Professor of Computer Information Systems and Professor of Computer Science at Georgia State University, Georgia State University, Atlanta, GA, USA.

Roman Lukyanenko (roman.lukyanenko@hec.ca) is an associate professor in the Department of Information Technologies at HEC, Montreal, Quebec, Canada.

Wolfgang Maass (wolfgang.maass@iss.uni-saarland.de) is a professor at Saarland University, Germany, and Scientific Director at the German Research Center for Artificial Intelligence (DFKI), Saarbrücken, Germany.

Jeffrey Parsons (jeffreyp@mun.ca) is a University Research Professor at Memorial University of Newfoundland, Canada.

This research was supported by funding from the J. Mack Robinson College of Business, Georgia State University to Veda C. Storey and from the Natural Sciences and Engineering Research Council of Canada to Jeffrey Parsons.