Must AI Accuracy Come at the Cost of Comprehensibility?

By François Candelon, Boston Consulting Group; Theodoros Evgeniou, INSEAD; and David Martens, University of Antwerp

Companies looking to integrate AI in their operations should think twice before turning their backs on simpler, more explainable AI algorithms in favour of complex ones.

Artificial intelligence is constantly pushing boundaries and making complex decisions better and faster in ever more diverse aspects of our lives, from credit approvals, online product recommendations to recruitment. Companies are jumping onto the AI bandwagon and investing in automated tools to keep up with the times (and technology) – even if they are not always able to explain to customers how their algorithms arrive at decisions.

In 2019, Apple’s credit card business was accused of sexism when it rejected a woman’s request for credit increase while her husband was offered 20 times her credit limit. When she complained, Apple representatives reportedly told her, “I don’t know why, but I swear we’re not discriminating. It’s just the algorithm.”
There is a real risk when organisations have little or no insight into how their AI tools are making decisions. Research has shown that a lack of explainability is one of executives’ most common concerns related to AI. It also has a substantial impact on users’ trust in and willingness to use AI products. But many organisations continue to invest in AI tools with unexplainable algorithms on the assumption that they are intrinsically superior to simpler, explainable ones. This perception is known as the accuracy-explainability trade-off.

**Does the trade-off between accuracy and explainability really exist?**

To understand the dilemma, it is important to distinguish between so-called black box and white box AI models: White box models typically include a few simple rules, possibly in the form of a decision tree or a simple linear model with limited parameters. The small number of rules or parameters makes the processes behind these algorithms more easily understood by humans.

On the other hand, black box models use hundreds or even thousands of decision trees (known as “random forests”), with potentially billions of parameters (as deep learning models do). But humans can only comprehend models with up to about seven rules or nodes, according to cognitive load theory, making it practically impossible for observers to explain the decisions made by black box systems.

Contrary to common belief that less explainable black box models tend to be more accurate, our study shows that there is often no trade-off between accuracy and explainability. In a study with Sofie Goethals from the University of Antwerp, we conducted a rigorous, large-scale analysis of how black and white box models performed on nearly 100 representative datasets, or what is known as benchmark classification datasets. For almost 70 percent of the datasets across domains such as pricing, medical diagnosis, bankruptcy prediction and purchasing behaviour, we found that a more explainable white box model could be used without sacrificing accuracy. This is consistent with other emerging research exploring the potential of explainable AI models.

In earlier studies, a research team created a simple model to predict the likelihood of loan default, which was just less than 1 percent less accurate than an equivalent black box model and simple enough for the average banking customer to understand. Another high-profile example relates to the COMPAS tool that is widely used in the United States justice system for
predicting the likelihood of future arrests. The complex black box tool has been proven to be no more accurate than a simple predictive model that considers only age and criminal history.

**Understand the data you are working with**

While there are some cases in which black box models are ideal, our research suggests that companies should first consider simpler options. White box solutions could serve as benchmarks to assess whether black box ones in fact perform better. If the difference is insignificant, the white box option should be used. However, there are also certain conditions which will either influence or limit the choice.

One of the selection considerations is the nature and quality of the data. When data is noisy (with erroneous or meaningless information), relatively simple white box methods tend to be effective. Analysts at Morgan Stanley found that simple trading rules worked well on highly noisy financial datasets. These rules could be as simple as “buy stock if company is undervalued, underperformed recently, and is not too large”.

The type of data is another important consideration. Black box models may be superior in applications that involve multimedia data replete with images, audio and video, such as image-based air cargo security risk prediction. In other complex applications such as face detection for cameras, vision systems in autonomous vehicles, facial recognition, image-based medical diagnostics, illegal/toxic content detection and, most recently, generative AI tools like ChatGPT and DALL-E, a black box approach may sometimes be the only feasible option.

**The need for transparency and explainability**

Transparency is an important ingredient to build and maintain trust, especially when fairness in decision-making, or when some form of procedural justice is important. Some organisations learnt this the hard way: A Dutch AI welfare fraud detection tool was shut down in 2018 after critics called it a “large and non-transparent black hole”. Using simple, rule-based, white box AI systems in sensitive decisions such as hiring, allocation of transplant organ and legal decisions will reduce risks to both the organisation and its users.
In fact, in certain jurisdictions where organisations are required by law to be able to explain the decisions made by their AI models, white box models are the only option. In the US, the Equal Credit Opportunity Act requires financial institutions to be able to explain why credit has been denied to a loan applicant. In Europe, according to the General Data Protection Regulation (GDPR), employers must be able to explain how candidates’ data has been used to inform hiring decisions and candidates have the right to question the decision. In these situations, explainability is not just a nice-to-have feature.

**Is your organisation AI-ready?**

In organisations that are less digitally developed, employees tend to have less understanding, and correspondingly, less trust in AI. Therefore, it would be advisable to ease employees into using AI tools by starting with simpler and explainable white box models and progressing to more complex ones only when teams become accustomed to these tools.

Even if an organisation chooses to implement an opaque AI model, it can mitigate the trust and safety risks due to the lack of explainability. One way is to develop an explainable white box proxy to explain, in approximate terms, how a black box model arrives at a decision. Increasing understanding of the model can build trust, reduce biases and increase AI adoption among users and help developers improve it. In cases where organisations have very limited insight into how a model makes decisions and developing white box proxies are not feasible, managers can prioritise transparency in talking about the model both internally and externally, acknowledging the risks and being open to address them.

Our research demonstrates that simple, interpretable AI models perform just as well as black box alternatives in the majority of cases and companies should first consider white box models before considering more complex solutions. But most importantly, managers can make more informed and conscious choices only when they have a sound understanding of the data, users, context and legal jurisdiction of their use case.

*This is an adaptation of an article published in Harvard Business Review.*

**Find article at**

https://knowledge.insead.edu/strategy/must-ai-accuracy-come-cost-comprehensibility
About the author(s)

François Candelon  is a Managing Director and Senior Partner at the Boston Consulting Group. He is also the Global Director of the BCG Henderson Institute.

Theodoros Evgeniou  is a Professor of Decision Sciences and Technology Management at INSEAD. He has been working on machine learning and AI for over 25 years.

David Martens  is Professor of Data Science at the Department of Engineering Management, University of Antwerp, Belgium, where he teaches data science (ethics) and researches actively on the topic of Explainable AI.

About the research


About the series

AI: Disruption and Adaptation

Delve deeper into developments in artificial intelligence, especially the disruptions across value chains. This series examines AI’s impact on a range of sectors, including business consulting, education and the media. It also sizes up the regulatory and ethical questions tied to this game-changing technology.