Cracking the black box: The quest to understand the machines that run our lives

This contribution was published by The Times of Malta.

Suggested citation: Camilleri, M.A. (2026) Cracking the black box: the quest to understand the machines that run our lives, The Times of Malta, 30th May: https://timesofmalta.com/article/cracking-black-box-quest-understand-machines-run-lives.1129214

Artificial intelligence (AI) is now part of everyday life. It recommends what we watch online, helps banks approve loans, assists doctors in hospitals and even acts as a digital gatekeeper for who gets hired. Many people enjoy the convenience of these systems, yet few truly understand how they work. That is where the "Explainable AI" notion comes in. Essentially, it is a growing movement that is aimed at increasing AI transparency, to earn user trust.

For years, AI systems were treated like mysterious "black boxes". You feed information into them and they produce an answer. However, at times, it proves hard to clearly explain how they have reached their conclusions. Even the engineers who have built these systems sometimes struggle to fully understand the internal reasoning behind complex AI models.

This becomes worrying when AI is used in areas such as healthcare, education, banking, policing or public services. Imagine applying for a loan and being rejected by an AI system without any explanation. Alternatively, consider a hospital using AI to help doctors diagnose patients without anyone being able to explain why the system recommended a particular treatment. In such situations, people may naturally ask: Why did the machine decide this? Explainable AI (XAI) tries to answer that question.

The basic idea is simple. AI systems should not only give answers; they should also explain how they arrived at them. Users deserve understandable reasons behind decisions that affect their lives. Transparency builds trust. Without it, people may fear that AI is unfair, biased or unreliable.

These concerns are not imaginary. There have already been cases where AI systems produced erroneous results because they learned from flawed or incomplete data. Some systems have treated people unfairly because of gender, race, age or where they resided. If the data used to train an AI system contains bias, the machine will usually amplify it.

Notwithstanding, the world around us never stands still. Economies shift, behaviours evolve and social conditions change. As a result, the AI models that are trained on old data may become inaccurate over time. Hence, an AI system that worked well two years ago may suddenly start making poor or unfair decisions today. Experts call this "data drift" or "concept drift".

This is why explainability matters so much. When AI systems can be examined and understood, it is much easier to detect and correct their errors and biases.

Researchers and technology companies are already developing tools to make AI more understandable. Some of these diagnostic tools have unusual names such as SHAP and LIME. Despite the technical labels, their purpose is quite straightforward: These interpretability frameworks can help identify which factors have influenced an AI decision the most.

For example, if an AI system denies someone a bank loan, these tools can show whether income, employment history or debt level has played the biggest role in the decision. This allows humans to review whether the outcome was fair and reasonable.

In this day and age, explainable AI has moved beyond the lab; it is now a concern for everyone, not just for tech experts. Regulators, governments and businesses are demanding for more transparency. In Europe, the new AI Act and existing privacy laws such as GDPR are pushing organisations to become more accountable for how AI systems operate.

There is also growing recognition that humans must remain involved in important decisions. Experts often refer to this as the “human-in-the-loop” approach. In simple terms, AI is meant to support human judgement. It should not replace it. A doctor, teacher, judge or manager should still be able to question and override an AI recommendation when necessary.

This balance is essential because AI systems are powerful, but they are not infallible. They can make mistakes, misunderstand situations, hallucinate or fail to recognise unique human circumstances.

We simply cannot afford to trust algorithms blindly. This is where explainable AI steps in. It helps ordinary users feel more confident about the technology they use every day. When people understand how a system works, they are far more likely to accept it. Thus, transparency will replace fear and confusion.

However, the challenge is that there is often a trade-off between power and simplicity. The most advanced AI systems, including modern generative AI tools, are often the hardest to explain. Simpler systems are easier to understand but may not perform as well. Therefore, researchers are striving in their endeavours to find the right balance between accuracy and transparency.

Arguably, the future of AI hinges on trust. Society is unlikely to fully embrace technologies that appear secretive or uncontrollable. Businesses and governments must therefore ensure that AI systems are fair, explainable and aligned with human values.

If these systems remain opaque, as their modus operandi are hidden, blurry or impossible to see through, their users may lose confidence in them. On the other hand, if we make AI understandable, we can finally harness its full power to build a fairer, more beneficial future for all.

Debatably, explainable AI is more than a technical upgrade. It is a moral safeguard. It ensures that humans don't get sidelined as machines become smarter.

In this new information era, explainable AI isn’t just a technological upgrade; it’s a moral boundary. It ensures that as machines get smarter, humans don't get left in the dark. In a world shaped by intelligent machines, we must hold on to one simple rule: if an algorithm makes a choice that changes your life, you have every right to know how it reached its conclusion, why that decision was made, where the data came from and when the logic was applied.

Professor Mark Anthony Camilleri has extensively researched and published on the intersection of Explainable AI (XAI), sustainable business practices and the future of digital transformation.