A GUIDE TO EXPLAINABLE AI

UCL
SCHOOL OF MANAGEMENT

AUTHORS
Jasmin Booth (Capgemini)
Paul Walton (Capgemini)
Iain Cooper (The UCL MBA)
Wendy Kent (The UCL MBA)

Capgemini

ABSTRACT

AS THE TITLE OF THE HBR ARTICLE MAKES CLEAR, "THE SECRET TO AI IS PEOPLE". SINCE AI AND PEOPLE WILL BE WORKING TOGETHER, THEY NEED TO BE ABLE TO UNDERSTAND EACH OTHER. THE AI NEEDS TO BE ABLE TO EXPLAIN TO PEOPLE WHAT IT'S DOING AND WHY.

Explainability is also a key element of the new regulatory environment emerging for AI.

But AI and machine learning challenge many deeply embedded habits and assumptions about technology and its uses. Implementing AI without understanding these habits will cause unexpected consequences.

This article was developed by Jasmin Booth, Paul Walton (both from Capgemini), Iain Cooper and Wendy Kent (both UCL MBA students) as part of an Analytics Lab project to consider these questions.

WHAT IS EXPLAINABILITY?

We have always had to provide documentation for conventional technology. This is a new problem because AI exposes some new issues. Explanations cover a wide range of possibilities. A Royal Society policy briefing describes the following types of explainability:

- "interpretable, implying some sense of understanding how the technology works;
- explainable, implying that a wider range of users can understand why or how a conclusion was reached;
- transparent, implying some level of accessibility to the data or algorithm;
- justifiable, implying there is an understanding of the case in support of a particular outcome;
- contestable, implying users have the information they need to argue against a decision or classification."

As well as these different uses, explanations need to have an appropriate level. Explaining how a machine learning model has derived a particular conclusion may not be helpful to someone unfamiliar with the model. The model may be based on features in the data that don't relate to the way in which people think. (Think about superficial explanations from call centre agents that leave customers very frustrated; or when train delays are ascribed to "operational problems".) Explainability may need to reach further into organisational knowledge to express the rationale much more clearly.

The relationship with organisational knowledge applies both ways. Machine learning may also contribute to organisational knowledge by finding patterns in the data that experts were not aware of. This may leave some people in the uncomfortable position of having their expertise challenged. Establishing trust in cases like this is especially important in the implementation of AI. Explainability should enhance human expertise not threaten it.

Explanations may be needed to support different business activities including, for example: regulation and compliance, customer service, the transfer of work between AI and people, the analysis of service effectiveness, and audit.

WHY DOES AI CAUSE AN ISSUE WITH EXPLAINABILITY?

We have always had to provide documentation for conventional technology. This is a new problem because AI exposes some new issues.

1. There are difficulties with black box machine learning. Unlike traditional system development, the rules that govern the decisions are not defined and documented by people as part of implementing the technology. So, there is no clarity about what the technology does and why.

2. There is a deeper level of explanation required. As more intelligent technology is built into business processes, the cognitive depth of the relationship between people and technology will increase. This will make the interface between them (including explanations) more difficult to manage.

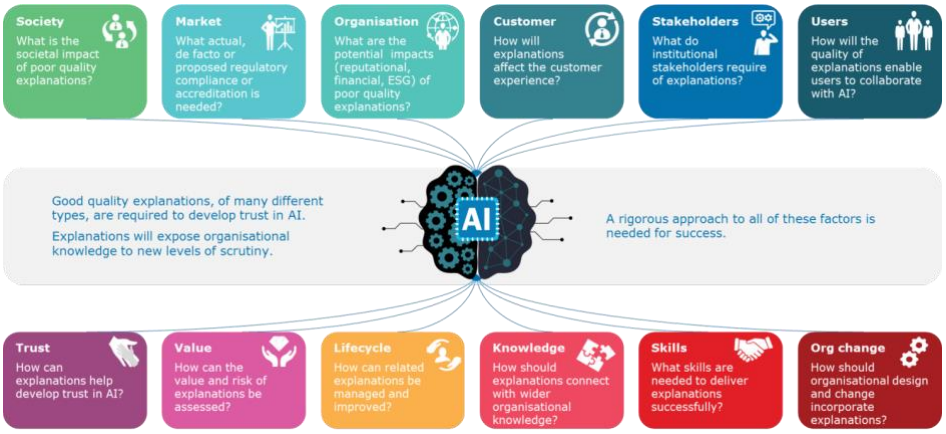
3. Explanations are part of a general trend to make organisational knowledge more readily available (familiar examples include chatbots and digital twins). Therefore, explanations need to be considered in that wider context. Explainability makes knowledge explicit and may reduce transaction costs. But exposing organisational knowledge more widely may reduce competitive advantage.

These issues arise because AI challenges many assumptions about how organisations work. It has taken a long time to come to terms with the ways in which people and organisations interact with conventional technology. So, we can expect that it will take time to come to terms with AI as it becomes used more widely.

On the flip side, if you expect to be able to use existing processes for implementing explainability then you may come unstuck. Instead, you need to look more deeply at the different ways in which explanations will affect what you do.

WHAT DO YOU NEED TO THINK ABOUT?

With this complex context, we need to zoom out and consider the context in which explainability sits.



There are five important themes to consider:

Trust.
Developing trust is critical to successful AI. A lack of trust will impact the brand and will damage relationships with regulators, stakeholders and users with a consequent impact on reputation and financial results.

But trust also has to be matched with sufficient scepticism—it is important to understand when explanations are not right and to build in feedback loops.

Value.
Business cases need to consider risks very carefully. Because explanations are a way of explaining organisational knowledge to the outside world as well as internal users, incorrect explanations may have a large impact. The advantages of machine learning in increasing the level of automation need to be balanced against the risks of poor-quality explanations.

Lifecycle and knowledge.
Explanations have a lifecycle. A single interaction may need to be explained in potentially different ways, to customers, customer service agents, managers and regulators. This lifecycle also needs to be integrated with the management of organisational knowledge, as the provision of explanations may expose inadequacies in organisational knowledge and its management.

Providing explanations exposes organisational knowledge, so how can you protect any related competitive advantage? Or, in the future, will competitive advantage be derived from the knowledge itself or from the ability to manage, change, and deploy it as required in a changing environment?

Skills.
Achieving trust will depend on getting the small details right. This means that rigour will be needed throughout the implementation lifecycle including in such diverse disciplines as user research, product management, data science, testing, training and support.

This presents a problem. Many of these skills are in short supply, especially in relation to AI. So, any implementation of AI and machine learning may need to prioritise, partly at least, on the basis of skills availability.

Organisational change.
AI and machine learning are changing the dynamics of organisational change. As well as the different nature of the technology, some expertise will migrate from people to technology. This means that:

- Traditional expertise may be challenged as people may no longer be the sole source of organisational knowledge
- As people work with AI, organisational decision-making will shift more to the front
- People and technology will increasingly collaborate on more complex tasks requiring a higher level of mutual explanation and understanding.

These factors change fundamental aspects of organisational change. How should processes in which people and AI collaborate be designed to take advantage of the unique characteristics of each? Just because technology is available, should it be used? How will learning on the job change if some of the job is being done by AI? How is it possible to reduce the impact of the resulting "shadow learning" that places unexpected stress on people and teams? As the management of organisational knowledge becomes more important, what needs to change to make it a primary consideration in organisational change?

And, since the secret to AI is people, how can organisational change develop and maintain trust in AI?

CONCLUSION

AI and machine learning challenge many deeply embedded habits and assumptions about technology and its use.

So, to avoid the pitfalls of explainable AI, you need to:

- Ensure that AI developments are trusted
- Understand the value of, and the risks associated with, AI explanations
- Connect explanations throughout their lifecycle with organisational knowledge management
- Recognise and build the skills you need
- Build a rigorous approach to explainability into organisational change.

ABOUT US

CAPGEMINI

Capgemini is a global leader in consulting, digital transformation, technology and engineering services. The Group is at the forefront of innovation to address the entire breadth of clients' opportunities in the evolving world of cloud, digital and platforms.

THE UCL SCHOOL OF MANAGEMENT

The **UCL School of Management** is the business school of University College London, one of the world's leading universities, consistently ranked in the global top 20 for its academic excellence and research. The School offers innovative undergraduate, postgraduate, PhD and executive programmes in Management, Entrepreneurship, Business Analytics, Business Information Systems, and Finance, designed to prepare students for leadership roles in the next generation of innovation-intensive organisations.

THE ANALYTICS LAB

The Analytics Lab is an enrichment module for business students where they are able to explore topical questions in the domain of business analytics and digital economy via hands-on experience. Students are offered the opportunity to conduct research and work on projects with leading technology service and consulting companies.

It aspires to help UCL business students and alumni to be in the heart of fundamental changes and digital transformations in the business environment. Students enhance their practical abilities to manage and operate business activities effectively in view of rapidly developing digital and technological advancements in data analytics.