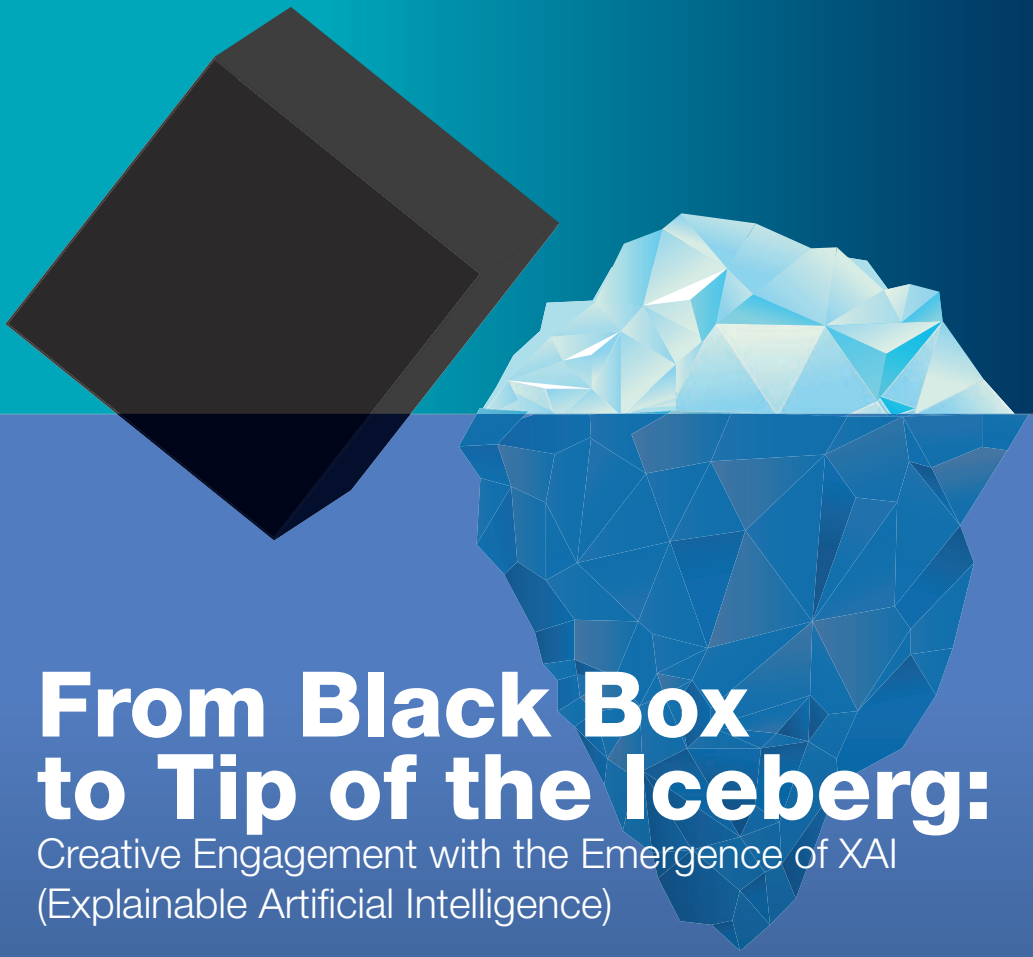
DEPARTMENT OF INFORMATION STUDIES &
UCL INTERACTION CENTRE

UCL



From Black Box to Tip of the Iceberg:

Creative Engagement with the Emergence of XAI (Explainable Artificial Intelligence)

UCL Grand Challenges

UCL Grand Challenge of Transformative Technology

https://blogs.ucl.ac.uk/hexai/

What is Explainable Artificial Intelligence (XAI)?

There are a number of definitions for Explainable Artificial Intelligence in use, including for example, this one; “Explainable AI (or ‘XAI’) is a machine learning application that is interpretable enough that it affords humans a degree of qualitative, functional understanding, or what has been called ‘human style interpretations’”(PwC, 2018).

Those at the workshop did not seek to define XAI. Instead they sought to engage with it as both an area of general societal concern and an emerging cross-disciplinary research field. This field is itself ill-defined, although attempts are now being made to map it out more systematically (Abdul, Vermeulen, et al., 2018). Terms and ideas which those involved in XAI are seeking to pin down include; interpretability, explanation, fairness, transparency, and accountability.

Explainable Artificial Intelligence or XAI has gained currency and significance as an area of societal concern, because AI is now being deployed in systems of decision-making that can have a real impact on people's lives. As a result AI, its workings and reasoning, have come to much wider attention and a need is being felt for it to be better explained to more people. There is also a sense that it may need to be better regulated through the creation of new governance frameworks (Panel for the Future of Science and Technology, 2019).

Taking a human-centred approach, this workshop brought together a number of individuals (full list overleaf) from a range of disciplines to explore and engage with ideas around explainable artificial intelligence. The results can be found inside.

Comments on the day from participants...

"

Fantastic meeting of interdisciplinary minds

Nice not to have a packed schedule

Very provoking and made me creative

My next workshop will be modelled on this one

"

The focus of the workshop was very much on the process of engagement, both with the topic and with everyone participating. Nevertheless, this leaflet was created subsequently to provide a record of that process and to try to highlight the main points that arose during the day. These points are summarised here, in the hope that they will stimulate others to engage with and explore ideas around Explainable Artificial Intelligence.

1.

“AI is not just Machine Learning” – Explaining AI before Explainable AI

The participants ranged from experts to novices in AI and consequently their views of exactly what it consisted of varied widely. Those working within AI sometimes grew frustrated with all being lumped together under the same label – AI – and with that label often seeming to equate to only one particular approach – machine learning.

Research into AI has a very long history and those involved in researching it form many different sub-communities and trains of thought, utilising a wide variety of approaches and mechanisms. Making this history, the rich individual stories and interesting individuals within it more widely known, would it was felt, be a good first step towards humanising and explaining AI in human terms – that is to say in terms of the humans who have created and sustained this vibrant field.

2.

“We need to change the Metaphor” – From Black Box to Tip of the Iceberg

Within XAI there is a lot of talk about black boxes and this metaphor draws attention to what those boxes contain (AI) and whether or not we get to see inside. Some participants discussed their disliking for the way in which this metaphor framed the conversation, and experimented instead with alternatives.

One suggestion was that of the iceberg. With an iceberg, there is a lot going on below the surface, but we don’t always need to worry about that bit. Sometimes it may be fine to just look at the pretty tip, but sometimes seeing below the surface is absolutely vital. With this metaphor it was felt the attention shifted from the AI inside the box to our human need for explanation and insight at specific moments and in specific contexts. For this reason it was also suggested that we should shift from XAI to Why-‘Y’AI.

3.

“We are all designers of explanations” – Understanding Explaining

A strong feeling to emerge from the day was that we need to understand a lot more about explanation as a contextual human behaviour with a role in cementing social cohesion and trust. It was noted that the giving, receiving and accepting of explanations was an interactive and iterative process involving multiple parties.

It was also recognised that explanations were crafted artefacts and that, to design them effectively, answers were needed to questions such as; When do we need to offer an explanation? When do we want to receive one? How detailed do they need to be? Why are we asking for an explanation? The participants imagined a unified theory of explanation and guidelines for general explainability, identifying this as the crux of the matter for anyone (or any system) seeking to offer explanations in the real world. As one participant put it;

“There is not one single good explanation, there are many good explanations for each individual involved, for each algorithm used. Understanding these requires a concerted effort from us to study the people, algorithms and the environment they are in.”

To this end, one concrete suggestion to come out of the workshop was for an ethnographic study around those systems (e.g. credit scores) where it was already possible to request an explanation of a decision made by a system assisted by AI.

Present at the Workshop

Mark Bell, The National Archives
Maura Bellio, University College London
Jenny Bunn, University College London
Jake Hearn, University College London
Emre Kazim, University College London
Adriano Koshiyama, University College London
Jo Pugh, The National Archives
Yvonne Rogers, University College London
Aidan Slingsby, City, University of London
Leontien Talboom, University College London
David Tuckey, Imperial College London
Cagatay Turkay, City, University of London
Luca Viganò, King's College London

There in spirit

Ann Borda, University of Melbourne
Daniele Magazzeni, King's College London
Raghad Zenki, University of Northampton

References

Abdul, Ashraf, Vermeulen, Jo, Wang, Danding, Lim, Brian, and Kankanhalli, Mohan (2018). 'Trends and Trajectories for Explainable, Accountable and Intelligible Systems: An HCI Research Agenda.' *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, April 21-26, Montreal, Canada, Paper No. 582.

Panel for the Future of Science and Technology (2019). *A governance framework for algorithmic accountability and transparency.* European Parliamentary Research Service, Scientific Foresight Unit (STOA), PE 624.262. Available at http://www.europarl.europa.eu/thinktank/en/document.html?reference=EPRS_STU(2019)624262

PwC, *Explainable AI: Driving business value through greater understanding*, 2018. Available at https://www.pwc.co.uk/audit-assurance/assets/explainable-ai.pdf