CHRIS BABER

WHY AI SYSTEMS NEED TO EXPLAIN THEMSELVES

Chris Baber and his team's work for CREST explores the question of 'explanation' in human interaction with Artificial Intelligence (AI) systems.

WHY ARE YOU TELLING ME THAT?

AI systems provide information based on complex algorithms and often massive collections of data. While explanations to help guide understanding of AI systems and the decisions they reach are necessary, explanation should not be solely about the algorithms and data that AI systems use. The point of explanation is not only *how* the decision was reached, but *why* the decision was reached, and what impact these decisions have on our beliefs and actions. Explanation should account for the consequences of the decision. As we suggest below, explanation as it relates to the *why* and the *consequence*, is too complex to be left to the developers of AI systems and instead should be achieved through supporting conversation between users and the AI system to negotiate what would make a useful answer to the question 'why are you telling me that?'

EXPLAINABLE AI

Our concept of explanation combines three elements:

1. Perception of the situation
2. Background knowledge
3. Definition of relevance (of a decision to the situation).

The perception that people and AI systems have of their immediate situation should not only relate to the data that are available but also the environment in which the analysis occurs or activity that occurs within the environment. From this, one can see that a human analyst would most likely 'know' more than the AI system in terms of wider, less tangible perceptions, just as the AI system would clearly 'know' more than the human in terms of the wealth of data available to it. For example, in medical applications, AI systems will outperform humans in the ability to scan millions of cases and discern patterns and associations — far more than a human physician (even a specialist in a particular branch of medicine) is likely to see over the course of their career. This is because contemporary AI systems continue to prove remarkably robust at solving well-defined problems, often achieving levels of performance that spectacularly outperform human counterparts, particularly in areas like board games or image classification. The definition of 'performance' here favours the AI system. However, in the medical arena, outcome is arguably more important, and here, comparison of the accuracy of diagnosis tends to show the human experts perform as well as AI systems.

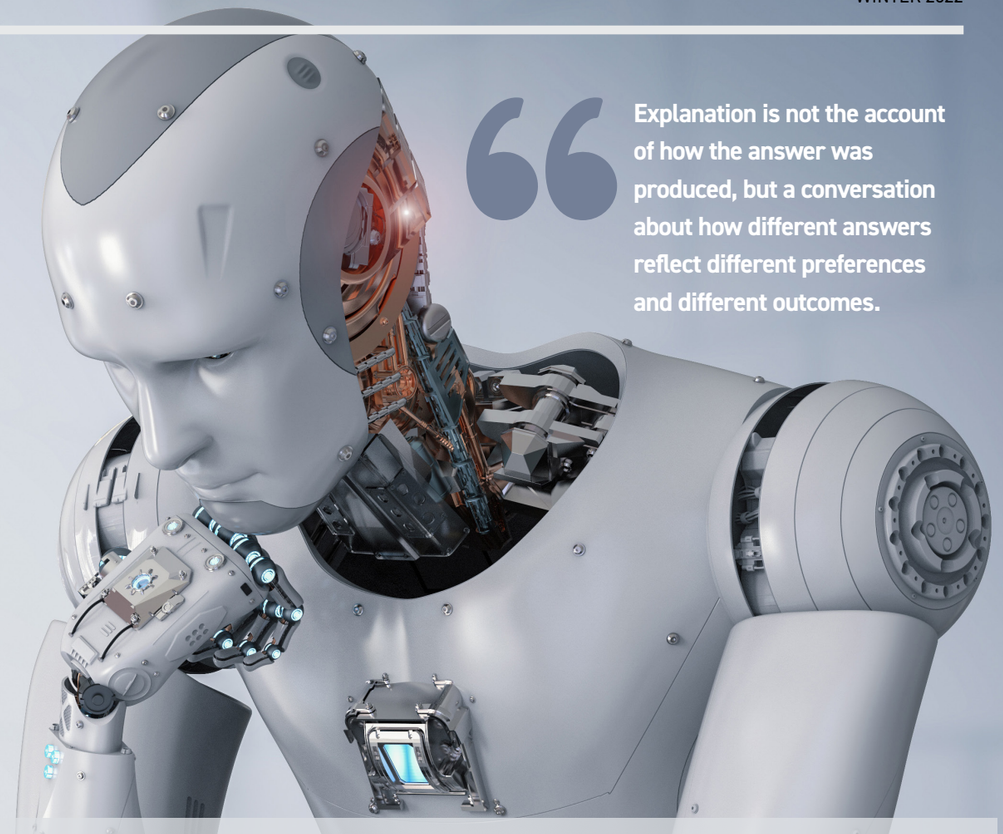
THE IMPORTANCE OF HYPOTHESIS TESTING

Where there are differences, these are not because the human is unable to produce a 'correct' (i.e., plausible for that situation) response, but because the AI system is often not able to juggle competing or ambiguous solutions. The experienced human physician can weigh up competing hypotheses, which lead to questions they ask the patient to seek other information. That is, the process of diagnosis involves the forming and testing of hypotheses through evidence collection informed by prior experience and expertise. We used AI tools (i.e., reinforcement learning) to model the use of information in human decision making and proposed that, in the absence of other sources, the optimal decision should accept the recommendation of an AI system only when its confidence exceeds 94%.

WHEN HUMANS INTERACT WITH AI

For human interaction with AI systems, differences in perception of the situation and background knowledge create different ways in which the conversation can be managed.

For example, recommender systems (which can suggest films, books, recipes, gifts, potential dates, etc.) assume that you and the AI system share the same interpretation of the situation (i.e., the criteria that define movies, such as genre) and the same definition of relevance (i.e., matching criteria to a list of recommendations, such as labelling the same movies as action-adventure). Any differences between what the AI system recommends can be easily handled by editing the criteria or rejecting the suggestions until you find one that you like. In this way, the conversation is not about agreeing with the answers but about agreeing on how best to define your taste.



> **Explanation is not the account of how the answer was produced, but a conversation about how different answers reflect different preferences and different outcomes.**

If, for example, the AI system recommends you watch *Highlander 2*, then it (probably) has a definition of relevance that differs from yours. In this case, there are two broad options. The first is to adapt the AI system's definition of relevance to better match yours. However, the other is to 'nudge' you into adapting to the one that the AI system has decided is optimal. For the latter option, let's assume that the AI system is providing 'health' advice and decides that the choices you make (for food, alcohol, tobacco, or exercise) are not optimal. It might introduce goals, reminders, or instructions to encourage changes in behaviour.

For this to be successful, the AI system needs to have a correct model of an optimal outcome, and you, the user, need to accept that the solution is optimal. In all cases, the outcomes for you (i.e., an enjoyable film night or healthier lifestyle) are the more important explanation points, as opposed to the algorithms that got you there.

Stuart Russell, in his 2021 Reith Lecture on *Living with AI,* defined 'traditional AI' as seeking to optimise a decision in terms of given data and criteria, but posited that 'future AI' ought to be designed to appreciate that humans might not know the exact criteria for a 'correct' decision or their true objectives.

To shift from finding patterns in data to determining questions to ask, an AI system would need to change, so that the AI system is able to reason about its own reasoning and decision-making. Rather than blandly presenting an 'answer', AI systems ought to be able to discuss options available to their human users, with the AI system predicting the likely consequences of different options.

In this way, explanation is not the account of how the answer was produced, but a conversation about how different answers reflect different preferences and different outcomes. But the differences between how people and AI systems reach their decisions need not be as far removed as might be imagined.

Our work has shown that, for decisions which involve the selection and judgement of information, the strategy that a person uses can be modelled using AI algorithms and this suggests that it might be possible to find a common language through which AI systems and people are able to review and negotiate their decisions.

You can read more about this project at: crestresearch.ac.uk/projects/human-engagement-through-ai

Professor Chris Baber is Chair of Pervasive and Ubiquitous Computing at the University of Birmingham.

Image credit: *phonlamaiphoto* | stock.adobe.com