CREST
CENTRE FOR RESEARCH AND
EVIDENCE ON SECURITY THREATS

5-MINUTE READ

GENERAL

AI AND SENIOR MANAGERS

This guide provides a brief overview of artificial intelligence systems for senior managers and answers questions relevant to their role.

> **We need to be careful about what we expect from AI and how much faith we place in its outputs.**

INTRODUCTION

Artificial intelligence (AI) systems can process huge quantities of data and discover patterns in the data that can benefit analysis. This could appear to be an attractive option when confronted with the problem of finding needles in haystacks where AI systems can be applied to filter specific instances from huge volumes of data. But we need to be careful about what we expect from AI and how much faith we place in its outputs.

It can be difficult to distinguish between AI and machine learning (ML) because many AI algorithms are built on ML models, but the main difference is the ability of algorithms to 'learn' – independently of any rules that have been programmed into them.

In many ML applications, we apply statistical models to data so that well-defined algorithms perform predictable manipulations of these data. In AI systems, the data manipulations are not predictable. This means that, as AI systems gain experience (from exposure to more data), they can generate their own 'policy'. For example, minimising the cost of performing an action and maximising the reward of that action.

While we might set values for 'cost' and 'reward', the AI could generate solutions that define an excellent policy, but which have no practical value for us. For example, an AI system that learned how to 'move from A to B' could decide not to develop legs (because controlling two or more legs can have a high cost) but instead learn to grow as tall as possible and then fall over (which would maximise the reward of making the movement as fast as possible). Clearly, this represents a solution that is optimal (in terms of the policy) but counter to common sense.

ANALYSIS THAT RELIES ON AI SYSTEMS IS DIFFICULT TO AUDIT

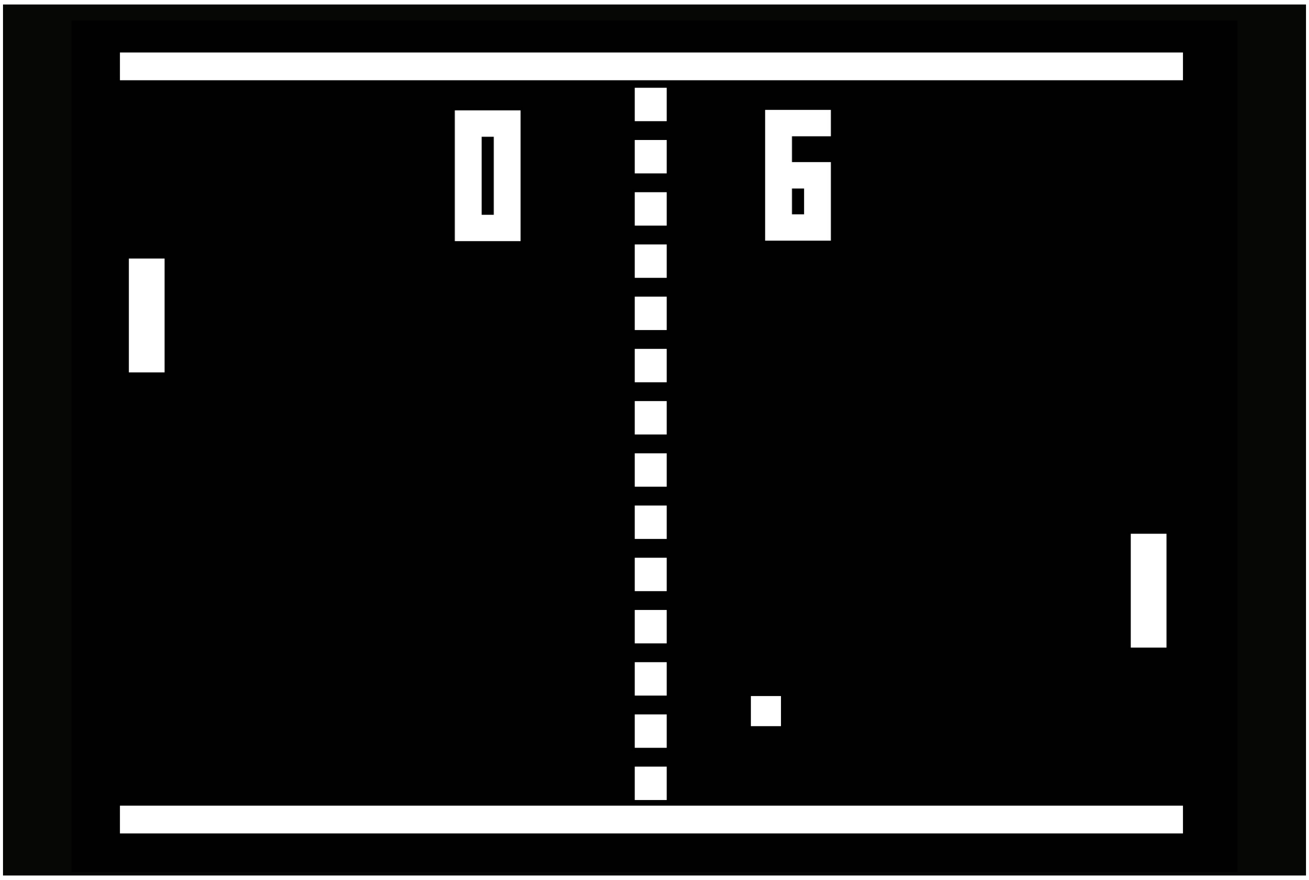
Deep AI can be a ‘black box’ that cannot be seen into. This means that the inner workings of AI systems are difficult to evaluate – we only have the outputs to interpret.

For example, an AI system that masters video games or board games can display specific actions. From this, you can infer what the AI system might have been doing. But, even with simple games such as Pong, the strategy that the AI system uses might be different to that used by a human. In Pong, we might watch the ball as it leaves a bat and moves across the screen: the AI system calculates an endpoint based on the angle of incidence and impact force and ignores the path the ball takes as it flies across the field.

While this is a simple example, it illustrates that how we might imagine an action being performed is likely to differ from how the AI system performs it.

An alternative approach is to produce a surrogate model, which works on a reduced version of the problem space and which can highlight the important features that the AI uses. However, this surrogate model works on a local instance and there is no guarantee it will generalise to other instances.

In both cases, we can be certain of the output, but rarely can we be certain of how this output was achieved. This makes it difficult to defend the process by which the output was produced, especially under scrutiny or challenge.



Pong is a table tennis–themed arcade sports video game, featuring simple two-dimensional graphics, manufactured by Atari and originally released in 1972.

QUESTIONS FROM SENIOR MANAGERS ANSWERED

1. HOW WILL THE EXPLANATION OF THE AI SYSTEM'S OUTPUT CHANGE IF ANY OF THE INPUTS ARE CHANGED?

Analysis based on AI systems is not neutral. While the AI system will work with the data provided to it and produce accurate results from these data, this process is not neutral. Rather, the AI system contributes to an analysis process in which its inputs are defined by a data collection process and its outputs are defined by analysis objectives. The accountability for these processes lies with the managers of the process (rather than the users or developers of the AI).

2. HOW WILL THE AI SYSTEM CHANGE THE PROCESSES FOR WHICH I AM ACCOUNTABLE?

A lack of explanation from AI systems creates an explanatory void that people will fill with their own beliefs, hunches, and expectations. Even with more simple versions of AI systems (which make use of basic ML algorithms), the output will be based on the structure of the data and the statistical model applied to these data. This means that the algorithm has no belief as to why the results arise (so cannot generalise to account for missing or different data).

> “A model built on data collected at one point in time does not guarantee that the outputs will be relevant later”

But humans will apply interpretations as to why the outcome was produced. This can mean that people will read into outcomes in ways that are not justified by the model (but which can feel plausible). Seeing data presented in a simple graph could encourage people to infer causal relations that are not actually present.

3. HOW DO WE MAKE SURE THAT WE CAN CHECK AND CHALLENGE NOT ONLY THE OUTPUTS OF THE AI SYSTEM BUT ALSO THE INTERPRETATIONS THAT WE ARE APPLYING TO THESE?

AI system output need to be interpreted by expert analysts who are familiar with, and knowledgeable about, the situation. Even when AI produces a plausible output, the definition of plausibility should not simply be in terms of the algorithm's performance. Various metrics define how well an AI system performs (in terms of consistency, reliability, etc.) but these metrics are internal to the AI system.

There is widespread recognition that metrics external to the system are more important in analysis. For example, in medicine, an AI might propose a 'correct' recommendation based on the data available to it, but clinicians might reject this because it implies an inappropriate disease model or would lead to a dangerous treatment regime. This does not mean that the AI system was wrong, just that additional knowledge was being applied by humans.

4. HOW DO WE MAKE SURE THAT THE EXPERTISE OF OUR ANALYSTS IS USED APPROPRIATELY ALONGSIDE THE AI SYSTEM?

To make the best use of AI, the data that it uses needs to be collected, prepared, and curated. A model built on data collected at one point in time does not guarantee that the outputs will be relevant later (especially when the situation that produces the data is constantly changing).

The collection and preparation of data rely on assumptions about how these data are to be used, but if these assumptions do not match the underlying statistical models that are being applied by ML or AI, or if they do not match the reward functions applied by AI, there could be a discrepancy in the collecting, coding, or management of these data (particularly if these processes are performed by people other than the users of the AI).

5. HOW DO WE ENSURE THAT ALL PEOPLE AFFECTED BY USING THE AI SYSTEM UNDERSTAND AND APPRECIATE THEIR CONTRIBUTION TO THE PROCESS IT REQUIRES?

Outputs from AI systems need to comply with the principles of *FATML* (fairness, accountability, and transparency). The outputs from the AI system, both in terms of recommendations and the consequences arising from actions based on those recommendations, should accord with the principles of FATML.

About the authors

- Chris Baber, University of Birmingham
- Emily McCormick, University of Birmingham
- Ian Apperley, University of Birmingham

READ MORE

This guide comes from the Full Report: Understanding The Problem Of Explanation When Using AI In Intelligence Analysis.

You can find this and other outputs from the project ‘Human Engagement Through Artificial / Augmented Intelligence’ at https://crestresearch.ac.uk/projects/human-engagement-through-ai/

CREST Centre for Research and Evidence on Security Threats

Understanding The Problem Of Explanation When Using Artificial Intelligence In Intelligence Analysis

FULL REPORT
JUNE 2021

Chris Baber
Emily McCormick
Ian Apperley

IMAGE CREDITS
Copyright ©2021 R. Stevens / CREST (CC BY-SA 4.0)