CORA Cork Open Research Archive
Cartlann Taighde Oscailte Chorcaí

Title	There is no such thing as an ethical black box
Authors	O'Sullivan, James
Publication date	2025
Original Citation	O'Sullivan, J. (2025) 'There is no such thing as an ethical black box', Education after the algorithm: Co-designing critical and creative futures, Dublin, Ireland, 20-21 February.
Type of publication	Conference item
Rights	© 2025, the Author. - (http://creativecommons.org/licenses/by/4.0/)
Download date	2026-10-13 21:39:58
Item downloaded from	https://hdl.handle.net/10468/17118

UCC
University College Cork, Ireland
Coláiste na hOllscoile Corcaigh

There is no such thing as an ethical black box

James O'Sullivan
Higher Education Authority

Originally presented at *Education after the algorithm: Co-designing critical and creative futures*,
Dublin City University, February 21st, 2025

Over the last few months I've been engaging with colleagues across Ireland's institutes of higher education on the subject of generative AI, and some of those conversations have prompted my brief position paper here today.

One of the things I've noted is that there is a belief across many of our institutions that tools like ChatGPT and Copilot can be implemented in ways which align with our values as educators. And I'm inherently sceptical of that view.

What I hope to argue — and I offer this argument as a collegial provocation more than anything else — is that there is a fundamental incompatibility between most generative AI systems and the core tenets of higher education.

I think most of us would agree that higher education should be built on openness and trust. And of course generative AI poses a real threat to these values, because tools like ChatGPT obscure their processes from scrutiny. They are designed, very intentionally, as the antithesis to trust and openness.

As many will be aware the term black box originates in engineering and computer science, referring to systems where we can observe inputs and outputs but cannot see, or fully comprehend, the processes happening in between. In other words, we know what goes in, we see what comes out, but what happens inside is largely hidden from view.

This is particularly true of generative AI, particularly proprietary deep learning models, which are often either hidden from users or described as 'unexplainable' (and I'm speaking here in terms of explainability in terms of access to underlying parameters, not in the technical, neural network sense).

The reality is that many AI models, particularly those developed by commercial tech companies, are proprietary. Their architectures, datasets, parameters, compute and training methods are not publicly disclosed.

And this raises a fundamental question: If we do not know how an AI system works, how can we trust it? If it cannot be scrutinised, how can it be held accountable? And, most importantly for our discussion today, can such a system ever be considered ethical in an educational setting? And it's important to remember that these are not abstract concerns. AI models are already being used in ways that shape the learning experience, so if we don't demand transparency now, we risk embedding black box systems into higher education in ways that truly would be irreversible.

The problem here is that higher education is not just about acquiring information; it's about developing an understanding in students about how knowledge is created, it's about questioning

assumptions, developing the ability to critically assess sources. However, when we integrate black box AI systems into our educational infrastructure, we actively work against these principles.

And this creates three major problems:

First, it erodes trust. Students and staff are being asked to accept AI-generated content without fully understanding how it was produced. And if we do not know where the knowledge comes from, how can we assess its validity? How can we expect students to engage critically with texts that have no discernible authorship or accountability?

Second, it undermines digital literacy. One of the fundamental aims of education is to equip learners with the skills to navigate an increasingly complex information landscape, but black box AI discourages critical engagement with sources.

And third, it reinforces inequality. The companies that own and control generative AI systems have access to proprietary knowledge and high performance computing resources that are not available to the wider academic community unless you're willing to pay to play. This creates a divide—those with the financial resources and institutional access to these tools have an advantage, while others are left in the dark. And I think we all would agree that higher education should be a leveller, not a mechanism for further stratifying knowledge access.

I wrote about some of these issues on my Substack this week in the context of what I call the physicality of learning. I wrote:

"Higher education learning used to be as much about the process as it was any answer. Learning was physical, tangible, and immersive … it cultivated a particular kind of intellectual intimacy with the material that is increasingly under threat in an era where algorithms, and now, generative AI, deliver answers on demand."

This is the core of the problem. AI in higher education should not just be about efficiency or convenience, it must be about fostering deeper engagement with knowledge. And that cannot happen if the systems that we use to generate knowledge remain fundamentally opaque.
And there are ways in which educators across the country, and indeed, many of the people in this room, are attempting to respond to this challenge.

It seems to me that the most prevalent response has been to prioritise critical AI literacy. Students are allowed to use generative AI tools but are taught to interrogate their outputs, identifying instances of bias, misinformation, or superficial reasoning.

This is a start—but ultimately I think this response is insufficient in itself.

Fundamentally, this strategy of critiquing AI outputs relies on a passive, rather than active, model of learning.

Higher education is about more than just understanding knowledge—it is about creating knowledge. And critical AI literacy, as currently practised in most contexts, really only teaches students how to *react* to AI-generated content. It does not teach them how to be active participants in knowledge production.

So if AI remains a black box, students are placed in a passive position: they are left to evaluate something they did not create, something they cannot interrogate beyond surface-level scrutiny. And instead of shaping the future of AI, they are relegated to mere consumers of it.

At an even more fundamental level, it seems impossible that institutes of higher education can ethically justify the integration of tools like ChatGPT and Co-pilot because of second-degree plagiarism.

The fundamental problem is that large language models like ChatGPT are trained on sources that are neither visible nor acknowledged, but we do know, from a great many court cases, that this training data includes material that was obtained without permission, without compensation, and without citation.

Whatever response we take to generative AI in our classrooms, if the models we are using are trained in such a manner, we are wilfully engaging with tools that draw on a corpus of stolen intellectual property. And that seems impossible to justify in any educational context.

But it's not all grim, there are of course way in which we can move towards an ethical AI framework for higher education.

But this will require a radical shift in how we integrate AI into our universities, and institutions will essentially need to break from their reliance on the big edtech vendors.

Firstly, universities must demand transparent and explainable AI systems—models that allow students and educators to interrogate their decision-making processes.

This means encouraging the use of open-source AI models over proprietary, closed systems; promoting algorithmic literacy, ensuring students and educators understand how AI models function, and requiring explainability features in AI tools so that users can see and question the logic behind AI-generated evaluations.

Secondly, process-oriented assessment, which I know many people here have been speaking to. This means assessing intellectual process as opposed to product, effectively requiring students to show their reasoning by documenting their research path—what sources they consulted, how they weighed competing arguments, and how they arrived at their conclusions.

Third, implementing AI as a participatory tool. If students are using AI, they must also be engaging with it critically and shaping its role within their disciplines. This means encouraging students—across all disciplines—to experiment with AI development, including actually creating and modifying AI models in controlled environments.

Essentially, universities must resist proprietary AI. We cannot go down the edtech route we went with other aspects of our digital education systems and services. In an ideal world, we'd have a state-funded, open systems that provide transparent, ethical and equitable access to the affordances of gen AI.

And this isn't a pipe dream.

The OpenEuroLLM project, for example, is a consortium of 20 leading European research institutions, companies and EuroHPC centres currently building a family of open-source multilingual, large language foundation models for commercial, industrial and public services.

Similar projects will emerge: an Irish project of this scale could emerge.

When we talk about AI systems, one of the key measures we consider is performance. In simple terms, performance refers to how well an AI system is doing at the task we designed it for.

For example: in a speech recognition AI, performance might mean how accurately it transcribes spoken words. In a chatbot like ChatGPT, performance might be measured by how well it generates relevant and coherent responses, but it's OpenAI, its developers, who determine what constitutes relevant and coherent (which is the problem).

With open models, universities and educators would get to decide what good performance looks like. And this isn't some loose moral notion, we could literally tweak models to perform based on our values.

So multi-objective optimisation is a technique which allows us to balance multiple priorities in a computer system at the same time, so in the case of a large language model, we could optimise for: accuracy but also fairness.

While different institutions and disciplines will hold different interpretations of accuracy and fairness, the key message is that AI performance should not be something dictated to us—we should define it ourselves. And we can. It is less a technical challenge than it is a political and financial challenge, but it is possible and in my view it is the only way we can ethically integrate AI into Irish higher education.