🌻 Trust the algorithm, not the AI

26 May 2025

I often hear concerns about algorithms and AI, in everyday life as well as in evaluation, taking over our lives or making us submit to decisions made by machines.

The worry about losing control to machines is real, but we need to distinguish between different cases, and in particular between **using algorithms to make decisions** and **using AI to make decisions**, especially **evaluative decisions**. This is particularly relevant in the field of evaluation.

An algorithm is simply a set of explicit steps to make a decision or produce an output, usually expressed in code or clear language. Organizations have used such rule-based systems for decades.

Some different ways to make decisions

No algorithm: trust the human

The alternative (precursor) to algorithms is trusting humans to make decisions. This can be great if humans consider context and individual circumstances, what Scott calls "mētis," or local, practical, tacit knowledge, (Scott, 2020) but it can also lead to bias and corruption.

We can see **rubrics in evaluation** (King et al., 2013) as a kind of soft algorithm. We usually welcome rubrics because they make evaluation criteria more explicit, transparent, and less subject to the whims and unreliability of individuals.

Algorithms based on explicit criteria

Algorithms can help decide things like student admissions or loan approvals using clear steps (e.g., check age, if under 18 go to step 12, otherwise continue with step 5 ….). When implemented wisely, algorithms can improve fairness and consistency compared to human judgment alone.

Using statistical models

Some algorithms use statistical models to predict outcomes, like creditworthiness, by combining data such as age or location. A statistical model uses parameters like age or location each of which has shown to be associated with the outcome, which makes it somewhat transparent.

Both explicit and statistical algorithms can be criticized for bias, but at least they can be transparent if their rules are published. Problems arise when rules are hidden or people are discriminated against because of the groups they belong to.

In a more advanced statistical model we might find it increasingly hard to understand where the different parts of the formula come from: it might combine parameters in ways which for us seem meaningless and hard to justify but which are supposed to be associated with the outcome of interest. Opaque models can become what data scientist Cathy O'Neil calls 'Weapons of Math Destruction' (O'Neil, 2017).

Machine learning

Machine learning is a subset of artificial intelligence where systems learn from data to identify patterns and make decisions or predictions, from "is this a picture of a cat" to "should we approve this person's application" often without being explicitly programmed with step-by-step rules. Instead of following a predefined algorithm, ML models develop their own 'rules' (which are often opaque to humans) based on the data they are trained on. Unlike generative AI, you can't chat with a machine learning model, you give it input in a fixed format (say, a picture) and get a fixed output, e.g. yes/no.

Sandra Seitamaa https://unsplash.com/photos/a-dog-and-a-cat-sitting-on-a-couch-Y45fzr5p3ug

In the extreme case we might have an algorithm based on machine learning (a form of AI, but not generative AI), where perhaps a neural network has been trained to distinguish desirable from undesirable candidates in just the same way you can train it to recognise a cat or distinguish a cat from a dog. Machine learning can be used to make decisions without clear formulas or rules. The process

becomes a “black box,” where we input data and trust the output without understanding how the decision was made.

Generative AI

Generative AI is a type of artificial intelligence that can create new and original content, such as text, images, audio, or code, after having learning patterns and structures from large datasets. These models don't just classify or predict, but generate novel outputs based on the input they receive, for example, continuing a conversation or answering a question.

The most extreme case is using generative AI for evaluative decisions without clear criteria (using it as a big black box): simply asking the AI, for example:

- is this program component effective?
- should this client get a loan?

Conclusion: make good use of algorithms

People often misunderstand algorithms, which can provide explicit and transparent decision-making. The real concern is not so much the use of algorithms but the shift toward the use of machine learning and generative AI, where the decision-making process becomes less and less transparent.

> Using AI in decision-making can be worrying not because it uses algorithms but because it *doesn't*.

Related

- chapter intro