Embracing the unexplainable: AI's black box in the courtroom

Kelly Friedman

The Sedona Conference®

DRAFT: FOR DISCUSSION PURPOSES ONLY

Embracing the unexplainable: AI's black box in the courtroom

> By far, the greatest danger of Artificial Intelligence is that people conclude too early that they understand it.[1]

The black box metaphor is often used to describe the reasoning behind an AI system's predictions and decisions. AI systems are built on complex mathematical models and multi-dimensional data sets, and the inner workings are not readily understood by humans. As such, AI forces us to confront predictions or decisions being made in a black box into which we have no visibility. Today, researchers and practitioners are increasingly focused on the explainability of AI or, otherwise put, explaining what occurs inside the black box. There are very good reasons for wanting transparency. When AI models are used to make decisions which have significant impacts on individuals, we would like to know how these decisions were made. However, when it comes to proving a case in court involving the output of an AI system which is but one factor in the court's ultimate decision, explainability is a costly distraction.

As AI systems become increasingly complex, their functioning may simply not be reducible to explanations at all, let alone explanations that laypersons and the judiciary can understand. Any diligent attempt to provide an explanation would likely involve testimony from myriad experts with different areas of concentration, such as data scientists, software engineers, data engineers and ethicists, to name a few. Such experts will have to be properly qualified, testify in court and be subject to cross-examination, with all the costs and delays this entails, only to leave the trier-of-fact with minimal clarity on how the system generated the response it did. Moreover, not many parties will be able to afford to engage in this battle of the experts. We can avoid this predicament entirely if we accept that the functioning of the system is unknowable, or at least unknowable for the purposes of the courtroom context.

There are examples in society of "black boxes" that we have grown accustomed to. In the area of pharmacology, for example, scientists strive to understand the specific biochemical, physiological or molecular processes through which a drug produces its therapeutic effects. However, there are commonly used drugs for which the exact mechanism of action is unknown[2]. Lithium and aspirin fall into this category of drugs and have been used successfully for decades.[3] We do not eschew the use of these drugs because we don't understand the inner workings of the black box, in this case the human body. Instead, we confirm its safety and efficacy by rigorously analyzing the measurable effects on the body, such as through clinical trials, adverse event monitoring, and the like.

There are examples in the legal sphere which underscore a general acceptance that lack of insight into a black box does not necessarily hamper the court's ability to do justice. Consider the forensic examiner who opines on whether two fingerprints belong to the same individual. When forensic scientists testify in court, to be successful, they have to clearly and accurately explain their opinion to laypersons.

---

[1] Eliezer Yudkowsky, from https://www.brainyquote.com/quotes/eliezer_yudkowsky_596818

[2] https://en.wikipedia.org/wiki/Category:Drugs_with_unknown_mechanisms_of_action

[3] https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7479624/

However, forensic scientists who deal with pattern evidence cannot reveal exactly how they came to the conclusions they did, as the analysis occurs in the black box of their mind. While they might be able to point to some features that they considered, their conclusion is more than a sum of features. They might also be influenced by their experience as well as factors outside of the expert's own awareness, such as the amount of visual information required to arrive at a response and the many cognitive biases the expert has. Expert fingerprint examiners glean information from a person's prints that cannot be reduced down to a number of details. In fact, any listing of similarities actually risks misleading the court into believing that the fingerprint examination process is a simple matter of counting up matching features; if it was just a matching exercise, then arguably expert evidence would not even be necessary since laypeople can do the matching themselves, and the judge ought not to admit the "expert" evidence. Instead, we tacitly accept the subjective nature of the expert's determination and do not peer into their mind.[4]

In practice, the testimony of the expert is tested against the rules of evidence. In the United States, the preliminary question of whether expert scientific evidence is admissible is answered by applying the principles laid down in *Daubert v. Merrell Dow Pharmaceuticals, Inc*[5] (the "Daubert Principles") which ask the trier-of-fact to consider the following:

1. Whether the theory or technique can be and has been tested;

2. Whether the theory or technique has been subjected to peer review and publication;

3. The known or potential rate of error or the existence of standards; and

4. Whether the theory or technique used has been generally accepted.

In Canadian courtrooms, in order for expert opinion to be admissible,

1. the information must be reasonably necessary (considered as likely to be outside the experience and knowledge of a judge or jury) and relevant (logically probative of the fact for which it is tendered);
2. the expert must be properly qualified as having a special or peculiar knowledge through study or experience in respect to matters on which he undertakes to testify; and
3. the information must not offend an exclusionary rule, the most important of which is that evidence should be excluded if the potential for prejudice substantially outweighs the probative value.[6]

In addition to the foregoing criteria, Canadian courts also look to the Daubert principles for guidance.[7]

In most legal cases, when an AI system is used to generate a piece of evidence, the question of how the system arrived at the decision can be ignored, with a focus instead on the inputs and the outputs and a careful application of the laws of evidence to authenticate the AI evidence. With respect to the inputs

[4] Searston, Rachel A, and Chin, Jason M, The Legal and Scientific Challenge of Black Box Expertise, University of Queensland Law Journal, Vol 38(2):237-60
[5] 509 U.S. 579 (1993) at pp. 593-94
[6] Glancy DG and Bradford JMW: The Admissibility of Expert Evidence in Canada, J Am Acad Psychiatric Law 35:350-6, 2007
[7] *R v J (L-J)*, 2000 SCC 51 (CanLII) at para 33

into the black box, we need to ask: Has the methodology used to develop the evidence been published and subjected to review by others in the same field? Have standard methods and protocols for operation of the AI system generally accepted within the field been followed? With respect to the outputs of the black box, we need to ask: Has the validity and reliability of the AI evidence been verified through independent testing? Is the error rate associated with the AI system acceptable in the circumstances?[8]

If the inner workings of the AI system are directly in dispute, then an explanation of its decision-making will be pertinent. However, in the vast majority of cases, it is the Daubert-style questions that matter, not the precise inner workings of the AI model. Even in those cases where data scientists and engineers do have the ability to explain what is happening in it, I submit that we should not waste resources trying to get the trier-of-fact to understand this evidence. I am not saying that explanations do not matter. Ideally, logical explanations will be available, and our trust in the systems will increase as a result. I am simply saying that justice does not require such explainability. In the courtroom, I fear that preoccupation with explainability will be a distraction from a rigorous examination of the inputs and outputs and confuse the trier-of-fact as to whether the output of the model should be given weight.

Furthermore, what we need in the legal context is justification, not explanation. Are there legal norms which justify the decision? Mireille Hildebrandt, a leading professor at Vrije Universiteit in Brussels who studies artificial intelligence as it deals with law, particularly the criminal justice system, put it this way:

> In law … we want a justification, and an explanation is not a justification. So what if the system says, "due to your scoring on the following six variables, you've crossed this threshold and therefore we're going to make this decision." Well, isn't that fascinating? The point is whether there is a justification in the form of a legal norm that justifies making that decision. I always use the example, if I go to court and the judge tells me, "I'm going to sentence you to 15 years because I had an argument with my wife this morning, the dog did something nasty on the carpet. I was in a traffic jam before I came here, and I don't like your hair," then I will tell the judge, "I don't care. I don't care about all these explanations because you can only convict me for reasons provided by the law."[9]

Simply put, in the courtroom, if a decision cannot be justified with logical legal principles, it is unjust. An explanation of how a system or decision-maker came to the decision is not needed. Explainability of AI should be the focus of other disciplines, not the legal one. Judges should be sceptical of parties wanting to parade numerous expert witnesses to explain the inner workings of an AI system. The AI system needs to be tested against is the principles of evidence law. Evidence law does not require the system to be explainable, it is irrelevant to the proceedings. The legal profession should leave the black box exploration to the data and computer scientists, and embrace that, within the legal context, we can leave the black box intact.

---

[8] Grossman, Maura and Grimm, Paul and Brown, Dan and Xu, Molly, The GPTJudge: Justice in a Generative AI World (May 23, 2023). Duke Law & Technology Review, Vol. 23, No. 12023, Duke Law School Public Law & Legal Theory Series No. 2023-30, Available at SSRN: https://ssrn.com/abstract=4460184

[9] Grossman, Maura and Grimm, Paul and Hildebrandt, Mireille and Gless, Sabine, Artificial Justice: The Quandary of AI in the Courtroom, Judicature International, September 2022