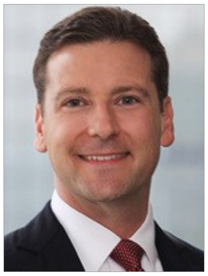
Expert Analysis

Explainability: Where AI And Liability Meet

By Danny Tobey



Danny Tobey

Law360 (February 21, 2019, 2:35 PM EST) – There is a commercial imperative to illuminate the mystery of artificial intelligence. Doctors are hard-pressed to rely on machine recommendations that contradict their own medical judgment without a clear explanation. Banks cannot rely on lending algorithms that may be discriminating against classes of borrowers through hidden variables. An air traffic controller whose software recommends a harrowing flight path will need more than a reason from the software — it must be a reason that humans can understand.

This is no small feat. DARPA recently announced a $2 billion investment toward the next generation of AI technology with “explainability and common sense reasoning.” If AI is going to augment rather than replace human decision-making, as many hope, then explainability is key. How else can a pilot, surgeon or soldier assess a machine’s recommendations — and override them when needed?

But therein lies the rub: The best AI will not just be faster or cheaper than human decision-makers, but reach better conclusions by seeing things people cannot. Sometimes, those counterintuitive recommendations may be explainable in hindsight. But in many cases, the more effective the AI, the harder it will be to explain its decisions in terms humans can understand. Indeed, many AI experts posit an inherent technical tradeoff between accuracy and explanation.

Imagine a chess app on your phone. You ask the game for a hint on what to do next. If it gives you a hint that sees two moves ahead, you might “get” the rationale, even if you didn’t see the move yourself. If the game gives you a hint based on 20 moves ahead — and all the mindboggling combinations that hint represents — you may never see the rationale, but which hint would you prefer? (Of course, the latest chess engines go beyond brute force, using deep learning to arrive at strategies even more inscrutable and effective.)

How much of the chess game do you want to understand, and how much understanding are you willing to trade away to get to checkmate? Now imagine not a chess game, but an airplane flight, or a bomb-defusing robot.

So makers and users of AI face a new and interesting problem. What is the acceptable tradeoff between explanation and accuracy? From a commercial perspective, the acceptable tradeoff varies by industry. It varies by type of device, level of autonomy and degree of risk inherent in getting a decision right or wrong.

It also varies by whether there are professional gatekeepers and regulators in the mix. A consumer may readily adopt an electronic toothbrush that follows an inscrutable algorithm, but a hospital will be reluctant to adopt devices that go against accepted wisdom, unless the U.S. Food and Drug Administration or professional boards embrace the technology first.

Enter the lawyers — because explainability is not only a commercial issue of consumer adoption, but also the close cousin of the lawyers’ art of finger-pointing: What happened, when did it happen and, of course, who got it wrong? Or, from a regulatory viewpoint, explainability is the investigator’s tool for the post-incident analysis: More like the black boxes of airplanes than AI, explainability can reveal what happened.

The law has long regulated the causal thicket between people and things. But what separates AI is twofold: The things are increasingly complex (a difference in degree), and they increasingly have agency of their own (a difference in kind). We are not yet at the point where AI has personhood enough to be subject to its own crime-and-punishment regime, whatever that might look like. And yet there is a widening disconnect between maker and machine that is not just expected but intentional — it is what makes AI work.

Causation and fault are increasingly opaque too, because of the number of human and machine interactions along the spectrum of AI usage. Here is a real-life example: a judge’s reliance on a bail-setting algorithm. The algorithm suggested the risk of releasing the defendant back into society was low. The judge followed the machine’s recommendation, only for the defendant to commit murder upon release.

So where did things go wrong? Should the judge have overruled the machine's recommendation? Perhaps not — if, according to valid statistics, the machine has a consistently higher predictive accuracy than humans, even if it got this one instance wrong. Or is human judgment more reliable in some cases than others — and if so, was this one such case, in which experience and instinct should have won out over the veneer of science?

What if the algorithm was trained on poor data, or even trained on excellent data that just happened to mismatch the local patterns in the judge's jurisdiction? When tragedies like this occur, how can we learn what happened? And how costly would it be, in terms of both money and human life, to find out?

And yet we also should not forget Occam's Razor. In this example, the answer turned out to be none of the above, but something quite basic. A human data-entry error led to a falsely low risk score for the defendant: just a simple blip at one of many human/machine interfaces, and someone died. And this was a basic example — a run-of-the-mill risk calculator. Imagine when the issue involves the functionality and safety of massive systems — such as continuous monitoring and management of trains, power grids or nuclear plants, all with orders of magnitude more variables, operators and flux.

Is explainability the solution? Many people think so. But, to sound like a lawyer, it depends on what the definition of explainability is. Some AI models, such as decision trees, are "interpretable." A human operator can look under the hood, trace the path of decisions a computer made and see how the AI arrived at its decision — maybe not fast enough for real-time review, but perhaps with the luxury of hindsight.

But interpretability does not necessarily provide anything in the way of meaning or causation from a human point of view. Those decision trees — however traceable — are still often based on statistical correlations, not mechanisms of action or causal steps that a human could stack into a logical flow of reasoning.

Other models are not even interpretable. These so-called black box algorithms represent layers and layers of compounding, arbitrary, non-linear combinations of vast variables — nothing a human could trace ("interpretability"), much less understand ("explainability").

Consider this telling example from Carnegie Mellon's Robotics Institute and its collaborators. They asked an AI to explain how it played Ms. Pac-Man. Its answer was simple: Just look for certain patterns of random pixels (patterns that to the human mind would not be obvious as patterns at all) and act accordingly. That's right: AI can beat you at Ms. Pac-Man without having any idea what a ghost is, or a pellet, or a video game.

The team then taught the machine to explain what it was thinking in terms people might understand: When it notices certain spectral pixels around the machine's point of interest, call it a "ghost." Google Brain is working to add such features on demand, like tools to ask an image processor how much the presence of "stripes" affects its detection of a zebra in a photograph, even if a concept of "stripes" had nothing to do with its detection process.

But what happens in a far less synthetic or contained environment than Atari or wildlife photography? Take, for instance, the ripples through a sea of data that might foretell a medical crisis or weather emergency — involving millions of equally-weighted variables that cannot be framed in human terms. Do you order the medical intervention that has its own stable of possible harms? Do you order the evacuation of an entire city? How do you judge if the warning is a clever improvement over existing predictions or a statistical misfire that should be overridden by human hands?

There is no one-size-fits-all solution for manufacturers and users of AI when it comes to explainability, safety and liability. In some cases, the risk profile of the endeavor will demand the highest level of accuracy, regardless of explainability. In others, the need for explainability, whether practical or moral — like the demand to have humans in the loop for lethal weapons — will require trading some accuracy for explainability.

Sometimes, there may be very little tradeoff between efficacy and explainability. In other cases, the gulf may be vast. Moreover, the desirable threshold between these tradeoffs may vary, based on as-yet-unknown legal and regulatory structures for AI. In strict-liability regimes, manufacturers may want maximum accuracy since they could be held responsible regardless of fault. In fault-based regimes, explainability could help manufacturers or operators who took due care to clear their names.

From a consumer perspective, explainability is again a double-edged sword — it can illuminate fault but also cause it; it can enhance safety and also reduce it. Like machines themselves, these issues can be case-sensitive. And for professional gatekeepers (doctors, lawyers, generals) who stand between a machine's recommendations and the end user (patients, clients, enemy armies), explainability is essential to judging a counterintuitive recommendation.

So how to untangle this thicket? A few concepts may help consumers and manufacturers, regulators and jurists begin to navigate this world. First, we must ask what explainability means as a standard: Is it the ability of a human to understand the machine's justification for a recommendation, or is it a human's ability to understand any justification for a machine's recommendation, however the machine got there?

The issue is far from academic. The FDA, at the cutting edge of trying to regulate AI in medical devices, has intermixed the two meanings of explainability in its draft guidance currently under review. In some passages, the FDA would require regulation where machine recommendations were based on "information whose meaning could not be expected to be independently understood by the intended health care professional user" — that is, the doctor must understand the machine's basis for the decision. In other sections, the FDA would require regulation unless the user could "reach the same recommendation on his or her own without relying primarily on the software function" — that is, where the doctor can justify the machine recommendation independently of the machine's basis.

Which is right? Should the user have to understand how a machine actually reached its recommendation, or only be able to reach a similar recommendation on other grounds? No less than which AI devices could be subject to FDA regulation turns on this question.

One can make an argument for both standards. But the first step is tease them apart, and recognize the difference between them. We might call the two definitions direct explainability and indirect explainability. Direct explainability would require AI to make its basis for a recommendation understandable to people — recall the translation of pixels to ghosts in the Ms. Pac-Man example. Indirect explainability would require only that a person can provide an explanation justifying the machine's recommendation, regardless of how the machine got there. These are different criteria, with different pros and cons.

Second, we might quantify the tradeoff between explainability and accuracy for a given model. Imagine a machine recommendation to eject a pilot based on 1,000 variables. How much predictive accuracy would be lost going from 1,000 variables to, say, the top 20? The top 3? We can imagine variables that explain the tradeoff between the full model and its explanation, or between explainable and unexplainable models: How close is the machine's approach to one people can understand? And how much accuracy or power is lost between the two?

Armed with that information, human observers, whether regulators, professional gatekeepers or others, could then better assess when human oversight is beneficial, or if explanations are doing more harm than good in a particular setting. One could even imagine the ability to tune between explanation and accuracy for a given question, finding the ideal tradeoff for that scenario.

Third, we might consider a different meaning of explainability — one suited to an age where the substance of what AI is doing is, ultimately, unexplainable. Because there are multiple ways to "explain" AI. One is to understand the substance of its reasoning. The other is to understand its process. As machine recommendations become more complex and inscrutable, procedural explainability, as opposed to substantive explainability, may be the next best alternative.

If a doctor cannot explain how scores of variables interact minutely to cause an output, she can understand the process the machine used: its data sources, training, validation, and sensitivity and specificity. This will allow only a more limited review of machine recommendations, but as machine reliability grows, this may increasingly become a norm.

And we should recognize the possibility, despite all the billions of dollars going toward explainability, that at some point, it may not matter at all. For now, the centuries-old assumption of human-monitored technology as safer may still hold, and "human-in-the-loop" AI can still be superior to machine-only outcomes in some contexts. But as that changes, so may our expectations for explainability and liability. A black box that consistently outperforms humans may supplant humans — regardless of how it works.

Manufacturers should anticipate this possibility as well, because it, too, shifts the legal analysis. Legal doctrines like the "informed intermediary" and "learned professional" have long broken the chain of causation leading back to manufacturers where skilled professionals stood between machine and end user. But as machines become the better-informed intermediary, those doctrines will come under pressure.

And courts that have long been reluctant to impose professional malpractice duties on technology companies may find themselves looking for the last human standing when technology supplants professionals. Indeed, the makers of IDx-

DR — in the FDA’s words, the first device authorized to make screening decisions “without the need for a clinician” — recently told the FTC that “IDx maintains medical malpractice insurance in case issues of liability arise.”

It is nearly impossible to predict the evolution of technology. AI may one day be able to explain its deepest, farthest thinking with perfect clarity and fidelity. But until then, the law will have to grapple with tradeoffs and ambiguity. That is an inevitable part of the legal response for clearer, safer AI.

Danny Tobey, M.D., is a partner at DLA Piper LLP.

The opinions expressed are those of the author(s) and do not necessarily reflect the views of the firm, its clients, or Portfolio Media Inc., or any of its or their respective affiliates. This article is for general information purposes and is not intended to be and should not be taken as legal advice.