warontherocks.com

I, Black Box: Explainable Artificial Intelligence and the Limits of Human Deliberative Processes

Molly Kovite

13-16 minutes

"AI has an 'explainability' problem. Your algorithm did XYZ, and everyone wants to know why, but because of the way that machine learning works, even its programmers often can't know why an algorithm reached the outcome that it did. It's a black box. Now, when you enter the realm of autonomous weapons, and ask, 'Why did you kill that person,' the complete lack of an answer simply will not do — morally, legally, or practically."

– *Liz O'Sullivan,* "I Quit My Job to Protest My Company's Work on Building Killer Robots"

Much has been made about the importance of understanding the inner workings of machines when it comes to the ethics of using artificial intelligence (AI) on the battlefield. Delegates at the Group of Government Expert meetings on lethal autonomous weapons continue to raise the issue. Concerns expressed by legal and scientific scholars abound. One commentator sums it up: "for human decision makers to be able to retain agency over the morally relevant decisions made with AI they would need a clear insight into the AI black box, to understand the data, its provenance

and the logic of its algorithms."

The underlying premise of such arguments is that if humans are making decisions on the ground, then other humans farther up the food chain in battlefield decision-making — commanders, political leadership, analysts, and so forth — will be able to find out why they made those decisions and respond accordingly. If algorithms are making these decisions, the thinking goes, we'll have no such insight, and we'll lose meaningful human control. But psychology research shows that we humans are not nearly as explainable as we give ourselves credit for, so we might be overstating the meaningfulness of the human control we thought we had in the first place.

Enter "explainable artificial intelligence," sometimes called XAI. With algorithms that can explain their decision-making processes — in a way that humans often can't — technology could increase, rather than decrease, the likelihood that those decision-makers who are not on the ground will get an accurate answer as to why a given decision was made.

The Introspection Illusion

There's considerable research that demonstrates humans' lack of insight into their own deliberative processes. People are generally willing to concede that they don't know where their own creative breakthroughs come from. But when it comes to more ordinary achievements of the mind, our mental processes seem more understandable. Say you're asked to name a brand of gum. You respond with "Wrigley's Doublemint." Why? "Because that jingle from their early '90s ad campaign is still stuck in my head," you might offer. Or if asked why you chose to buy a particular pair of

socks, you'll reason, "they were the best combination of high quality and low price."

Similarly, on the battlefield, warfighters can sometimes explain why they made a particular decision, while in other cases, they may acknowledge an inexplicable sort of divine inspiration. Mike Jaco, former Navy SEAL and author of *The Intuitive Warrior*, voices the latter when he explains, "by fine-tuning my intuition as a Navy SEAL, I was able to predict and avoid attacks to protect myself and my fellow soldiers."

On the other hand, Soviet Air Defense Forces Lt. Col. Stanislov Petrov was able to clearly articulate why he decided not to escalate the warning when the Soviet nuclear launch detection system malfunctioned on Sept. 26, 1983. The system was designed to notify Soviet leadership if the United States launched a nuclear attack so that they could retaliate before being destroyed. Petrov explained his reasons for his decision: it seemed illogical that the United States would only launch five missiles, the launch detection system was relatively new so he didn't trust it yet, the message passed through verification measures suspiciously quickly, and there was a lack of corroboration from ground radars.

Yet Petrov is likely as much in the dark as Jaco when it comes to understanding his own mental processes, as is the person who chooses Wrigley's Doublemint or the high-quality, low-price socks. Each of us has two "systems" of thought: the first is fast, instinctive, and emotional; the second, slow, deliberative, and conscious. But these aren't independent modes that we can voluntarily switch between. That second mode of thinking serves essentially as a press secretary for the first. We make decisions first, and then justify them — and even this slow, calculated process is opaque to

us.

One psychology experiment had students memorize a list of word pairs, some of which were intended to cue associations. Subjects who were given the word pair "ocean-moon" were more likely to name "Tide" when asked to name a laundry detergent — at about double the rate of the others. Yet when asked why they chose Tide, they offered many explanations, none of which mentioned the word pair (for example, "I like the Tide box" or "My mom uses Tide").

In another experiment, subjects were shown pictures of two women and asked which one they found more attractive. The experimenter then purported to remove the picture of the woman whom the subject did not select and asked subjects to explain why they found the remaining one more attractive. But some of the time the researcher had in fact removed the incorrect picture. When this happened, subjects noticed only 25 percent of the time. Those who didn't had no problem providing explanations as to why they chose the woman *they hadn't chosen* (such as "I like earrings" or "She seems approachable").

The extremes to which we delude ourselves about our reasoning processes are even more apparent in studies that examine patients who have had surgery to cut their corpus callosum, the bridge between the brain's left and right sides. This surgery prevents the two hemispheres from talking to each other. The left side of the brain, which is largely responsible for logic and language, finds itself justifying actions taken by the right side.

For example, researchers flashed a picture of a chicken foot in a split-brained patient's right field of vision, targeting his brain's left hemisphere, and a picture of a snow shovel in his left field of vision,

targeting his right hemisphere. When they asked him to point to which picture he had seen, he pointed to the chicken foot with his right hand (controlled by the left hemisphere) and the snow shovel with his left (controlled by the right hemisphere). When asked why he pointed to both, his left hemisphere (the side responsible for verbal justification) didn't know what the right side had seen — but that didn't stop it from offering an explanation. Instead of saying, "I don't know," the patient offered, "the chicken foot goes with the chicken, and you need a shovel to clean out the chicken shed."

Under certain circumstances, people do seem quite aware of the factors that influenced their decision-making. Specifically, when the stimulus that influenced the decision was noticeable, and the person found it plausible that the stimulus would in fact have an influence, people were able to link the stimulus to their decision. Perhaps unsurprisingly, there's also ample evidence that, when explaining our reasoning, we tend to come up with theories that make us look good to ourselves and others.

Petrov explained his reasoning but also noted, "I had a funny feeling in my gut." It was probably that feeling more than any of his articulated reasons that drove his decision. As the research suggests, he — like most humans — couldn't fully explain his brain's inner workings.

Explainability and Correctability

Most of us have ignored our navigation systems in favor of a "better" route only to end up in a traffic jam, thereby learning that overriding the machine can often be the wrong decision. We override the machine initially because we assume that it has misunderstood our desires or has incorrect information.

The problem is that we don't know whether our navigation system is being a dumb machine or we're being dumb humans. If the machine could explain itself to us, we'd be much more likely to make the right decision when using it. To that end, DARPA has undertaken a huge initiative to research the creation of AI that can explain itself to humans, which "will be essential if future warfighters are to understand, appropriately trust, and effectively manage an emerging generation of artificially intelligent machine partners." This field of research is called explainable AI, or XAI.

One paper to come out of this undertaking warns of the dangers of human operators not understanding the machine's methods:

> We believe that it is fundamentally unethical to present a simplified description of a complex system in order to increase trust if the limitations of the simplified description cannot be understood by users, and worse if the explanation is optimized to hide undesirable attributes of the system. Such explanations are inherently misleading, and may result in the user justifiably making dangerous and unfounded conclusions.

Ironically, what the authors of the paper have just described is the shortcomings of *human* efforts to explain their own behavior — shortcomings that, if done correctly, would be absent in XAI. Our minds are bafflingly complex, and our explanatory systems are highly limited. We fail to recognize these limitations, and instead believe explanations that are in fact optimized to hide undesirable attributes. And indeed, as the authors predict, this results in humans making dangerous and unfounded conclusions — conclusions that can have horrific results on the battlefield.

As a 2018 CNA report describes, in 2011 U.S. soldiers in

Afghanistan saw a group of men digging and moving rocks, building what appeared to be a defensive fighting position. At least one of them had a weapon slung across his back. The commander called the position into air support and authorized an attack in self-defense. It turned out they had attacked a group of girls who were carrying metal sickles to cut grass and using their head scarves to carry the grass.

A number of cognitive biases likely played a part in this tragedy. The soldiers had been attacked in that area every day the previous week. Due to the availability heuristic, that short-term, emotionally charged data would feel very important to a human, but an algorithm would make a more accurate assessment of danger based on longer-term data. Further, because the humans were on edge, confirmation bias would make them more likely to find evidence of danger, which may have contributed to their assessment that the girls were men and the sickles weapons. This kind of confirmation bias is unlikely to take root in a well-designed algorithm; instead, such an algorithm could have picked up on other indicators that would have led to the correct assessment. If we humans could observe our own cognitive processes, it would be easier for us to pick up on and correct for these flaws in reasoning.

In another example of battlefield tragedy, U.S. forces struck a Doctors Without Borders hospital in Kunduz, Afghanistan, in 2015, killing approximately 40 civilians. The U.S. soldiers were supposed to strike another, similarly shaped building nearby that was occupied by Taliban fighters. A myriad of minor errors led to this outcome. One such error occurred before the strike, when one of the air crew on the gunship conducting the attack questioned whether they were on the right target, and asked whether the

people milling around outside the building might be civilians. The person on the ground communicating with the crew could have taken this opportunity to reassess whether he had successfully communicated the target description. Instead he provided tautological reasoning to justify what he'd already decided: "compound is currently under control of [Taliban], so those nine [personnel] are hostile." A machine is unlikely to make the mistake of tautological reasoning or to fail to fully run a reassessment of a target when asked to do so. Due to the lack of insight we have into our own reasoning processes, we fail to even notice we're taking these shortcuts.

There are many reasons to be concerned about incorporating AI into lethal decision-making: the risks of improperly defining the machine's objective and the vulnerability to data manipulation and hacking are difficult to overstate. But if DARPA's undertaking succeeds, the decision-making of explainable AI may be significantly more transparent — and therefore more correctable — than human cognitive processes. This has the potential to improve targeting, reduce civilian casualties and friendly fire, and ultimately diminish unnecessary suffering in war. We shouldn't let our all too human cognitive bias towards the status quo stand in the way of that outcome.

Molly Kovite is the senior legal advisor for the International Humanitarian Law Department of the American Red Cross. She is also a Major in the U.S. Army JAG Corps Reserves, and has taught international humanitarian law at the University of Washington. The views expressed here are her own and do not necessarily reflect those of any of her employers.

Image: École polytechnique photo by J.Barande