The Ethical and Legal Implications of Black Box Artificial Intelligence

By Sharon D. Nelson, Esq. and John W. Simek

What is black box AI?

Put simply, black box artificial intelligence works according to rules that no one understands – it is designed to be impenetrable. If it is trained using biased data, it is going to produce biased results. Many organizations focused on artificial intelligence have bluntly said that black box AI is unethical.

Apple's credit card business was charged with having sexist lending models in November 2019 and an investigation by regulators is ongoing. Amazon retired an AI hiring tool after discovering (it took Amazon three years to make this discovery) that it discriminated against women.

COMPAS

The term "black box" has not been part of common parlance for a long time. We first heard the term widely used in 2016. In that year, the Correctional Offender Management Profiling for Alternative Sanctions (COMPAS) software used by some courts in predicting the likelihood of recidivism in criminal defendants was demonstrably shown by ProPublica to be biased against African Americans.

No one knew how it worked – it was "proprietary" so the company didn't want to be transparent about its programming – this is when many media stories began to talk about "black box" AI, in which no one can or will explain how the AI generates its output based on the input.

COMPAS is manufactured by the private company Northpointe which says that its algorithms are trade secrets. But should algorithms be used to arbitrate fairness? It's complicated.

Machine-learning algorithms are trained on "data produced through histories of exclusion and discrimination," writes Ruha Benjamin, an associate professor at Princeton University, in her book *Race After Technology*. Risk assessment tools like COMPAS are no different. Some people believe that they reduce inequities while others believe they make them worse. Hard to judge when they have no transparency.

In a controversial decision (*Loomis v. Wisconsin*, 2016), the Wisconsin Supreme Court decided that the recommendation from the COMPAS algorithm was not the sole grounds for refusing his request to be released on parole and hence the decision of the lower court did not violate Loomis's due process rights. Confirming the constitutionality of the recommendation risk assessment algorithm, the Court was widely seen as neglecting the strength of the 'automation bias'. Once a high-tech tool makes a recommendation, it is difficult for a human decision-maker to reject the recommendation.

Facial Recognition

Facial recognition using AI is a controversial topic. Is it accurate? What are the privacy concerns? Many people have fretted over facial recognition, but studies have proven conclusively that facial recognition technologies just aren't very accurate. An MIT study testing systems from IBM, Microsoft and the Chinese company Face++ showed that facial recognition was more accurate with respect to men than women – and, notably, far more accurate with white people than those with darker skin.

Researchers at MIT and elsewhere are working on an algorithm that will automatically 'de-bias' data sets by locating hidden biases and then re-sampling the data. A step in the right direction, but it's not going to be in place soon.

Who Trains the AI? Using What Data? Who Should You Sue?

In a 2020 post, Forbes discussed a case in which a major healthcare AI vendor's internal documents were leaked. The documents showed that the algorithms had made erroneous and unsafe cancer treatment recommendations in a number of cases. What happened? An internal investigation revealed that the software engineers had trained the AI using hypothetical data rather than real-world cases.

Mind-boggling but true. So who is liable for medical damage stemming from a black box algorithm? The manufacturer? The doctor or hospital that implemented it? Of course, the likelihood is that lawyers will sue them all until there is more clarity as to who is liable.

In the European Union, the General Data Protection Regulation (GDPR) includes an "explainability requirement" that applies to AI – which would seem to point at the manufacturer in imposing liability. There is no comparable law in the U.S.

The Dark Side of AI

There has been a lot written about “The Dark Side of AI” in the last two years. One article posted in ZDNet in August 2020 had the following headline: “Evil AI: These Are the 20 Most Dangerous Crimes That Artificial Intelligence Will Create.”

The most serious threat, as identified by scientists from the University College London (UCL), was deepfakes, particularly those used to undermine democracy. Clearly, manufacturers of those kind of deepfakes have no incentive whatever to make their algorithms public or transparent.

With the advent of autonomous cars, driverless vehicles could now become a weapon, as a delivery mechanism for explosives or even as a weapon in their own right (for instance, driving into crowds of peaceful protestors).

Cybercriminals are already using AI-infused attacks to penetrate government and industry networks – we do not imagine that they have any motivation to be transparent about their algorithms. Ditto for those cybercriminals engaged in various sorts of extortion attacks, easily enabled by AI’s ability to gather large amounts of personal data and information from social media and other resources.

We now routinely use AI systems for financial transactions, public utilities and public safety systems. Disrupting such systems could result in the failure of power grids, large scale damage to financial institutions – and just about any other chaos you can think of. This would all be via black box AI used by criminals or terrorists. Are we prepared to defend against these threats? Not very likely.

Dark AI Spawns XAI

If you didn’t know what dark AI was, you almost certainly don’t know what XAI is. The fears about dark AI have sparked interest in countermeasures, the chief one being Explainable Artificial Intelligence (XAI). There are many “major players” who agree on the need for XAI, including Microsoft and the Defense Advanced Research Projects Agency (DARPA) to name two organizations whose opinions don’t often coincide!

As DARPA notes, war presents a real problem if the AI utilized (and of course it is already being used by the military) is not explainable. Do we really want AI deciding who to kill or when to shoot down an airplane? DARPA is in favor of legally requiring humans to have the final say, in large part because it fears what dark AI may do. XAI, as DARPA points out, is the logical way to combat dark AI.

The Algorithmic Accountability Act of 2019

The proposed Algorithmic Accountability Act of 2019 was introduced in the House in April of 2019.

The bill, seeking to regulate bias in “high risk automated decision-making systems,” would require large companies (over $50 million in revenue or in possession of data of one million consumers) to audit their machine-learning systems for bias and discrimination in an “impact assessment.”

The bill met with quite a bit of criticism and does not appear to have gained any traction.

The definition of a high-risk automated decision system is broad and includes systems that pose a “significant risk” to individual data privacy or security or that result in biased or unfair decision-making; those that make decisions that significantly impact consumers using data about “sensitive aspects,” such as work performance and health; those that involve personal data like race, political and religious beliefs, gender identity and sexual orientation, and genetic information; or those that monitor a large public space.

The impact assessments would evaluate how an automated system is designed and used, including the training data it relies on, the risks a system poses to privacy or security, and other factors. Companies would have to reasonably address concerns these assessments identify, but they would not be required to disclose these impact assessments. Failure to comply would be considered an unfair or deceptive act under the Federal Trade Commission Act and subject to regulatory action.

That sounds like a plan that simply will not work. What company is going to confess (without coercion) to the results of an impact assessment that would subject it to a fine? Perhaps it would be better to have the assessments publicly available with information about what the company did in the course of conducting the assessment, redacting any proprietary information?

And if the risk is high, why would the law be restricted to “Goliath-sized” companies? We don’t see this kind of legislation as likely to pass anytime soon.

NIST Proposes Four Principles to Make Artificial Intelligence Explainable

Since Congress is unlikely to pass much legislation for the moment, we were happy to see that the National Institute of Standards and Technology (NIST) published a 30-page August 2020 draft of "Four Principles of Explainable Artificial Intelligence." The principles are a part of NIST's research to build trust in AI systems by understanding theoretical capabilities and limitations of AI, and by improving accuracy, reliability, security, robustness, and explainability in the use of the technology.

"AI is becoming involved in high-stakes decisions, and no one wants machines to make them without an understanding of why," said NIST electronic engineer Jonathon Phillips, one of the report's authors. "But an explanation that would satisfy an engineer might not work for someone with a different background. So, we want to refine the draft with a diversity of perspective and opinions."

The four principles for explainable AI are:

1. AI systems should deliver accompanying evidence or reasons or their outputs

2. AI systems should provide meaningful and understandable explanations to individual users;

3. Explanations should correctly reflect the AI system's process for generating the output; and

4. The AI system "only operates under conditions for which it was designed or when the system reaches a sufficient confidence in its output."

"As we make advances in explainable AI, we may find that certain parts of AI systems are better able to meet societal expectations and goals than humans are," said Phillips. "Understanding the explainability of both the AI system and the human opens the door to pursue implementations that incorporate the strengths of each."

Similar guidance was offered in July 2020 by the U.K.'s Information Commissioner's Office which may be found at https://ico.org.uk/for-organisations/guide-to-data-protection/key-data-protection-themes/guidance-on-ai-and-data-protection/.

National Association of Insurance Commissioners (NAIC) Adopts AI Principles

Lots of people want a bite of this apple and we have tried to make sure we mention some of the most current developments. On August 14, the NAIC adopted principles for AI developed by the NAIC's Artificial Intelligence Working Group – and yes, virtually every entity has some kind of AI working group these days.

The principles require that insurers and others using AI:

1. Take proactive steps to avoid proxy discrimination against protected classes. (As NAIC President Ray Farmer noted, this is part of the NAIC's broader effort to address racial equality issues.)

2. Monitor the operation of its AI system and remediate harmful, unintended consequences.

3. Provide responsible disclosures and give consumers an opportunity to inquire about and seek review of AI-driven decisions.

4. Take a risk management approach to each phase of the AI system's life cycle.

While these principles are not laws, they do explain what regulators expect and these expectations will form the basis for how AI is regulated.

Deloitte's 2020 Global Marketing Trends Report

Though Deloitte's 2020 Global Marketing Trends Report was released in October 2019, it made some interesting observations about algorithmic bias and lack of transparency. As the reported noted, "Trust is hard to gain and easy to lose."

Those words carry a lot of meaning. Many people have lost trust in black box AI because it has proven faulty time and again.

The report indicates a strong desire for explainable AI technology. AI is now used in making medical diagnoses. As a result, health care companies are working on and implementing solutions that give each diagnosis a confidence score that explains the probability and contribution of each patient's symptoms (vital signs, information from medical reports, lifestyle information, etc.) to the diagnosis. This allows medical professionals to see exactly why the diagnosis was made – and, critically, the ability to change the diagnosis if needed.

Final thoughts

As we noted at the outset, the clamor for explainable AI went global in 2016. Black box AI is now regarded with suspicion. As AI that is not explainable has, in practice, failed to meet expectations, it has come under attack and a host of organizations and companies are putting out ethical rules about AI, virtually all of them including the need for explainability along with privacy protections.

The General Data Protection Regulation (GDPR) in the EU includes an "explainability requirement" that applies to AI. It requires companies building some forms of AI that make decisions about individuals be able to explain how the decision was made. We have no such laws yet, but they are sure to come in time.

As a Google AI expert commented, "The era of black box machine learning is behind us." Maybe not quite yet – but soon.

Sharon D. Nelson, Esq. is a practicing attorney and the president of Sensei Enterprises, Inc. She is a past president of the Virginia State Bar, the Fairfax Bar Association and the Fairfax Law Foundation. She a co-author of 18 books published by the ABA. snelson@senseient.com.

John W. Simek is vice president of Sensei Enterprises, Inc. He is a Certified Information Systems Security Professional, Certified Ethical Hacker, and a nationally known expert in the area of digital forensics. He and Sharon provide legal technology, cybersecurity and digital forensics services from their Fairfax, Virginia firm. jsimek@senseient.com.