AN INTRODUCTION TO ARTIFICIAL INTELLIGENCE FOR FEDERAL JUDGES

Federal Judicial Center
2023

An Introduction to
Artificial Intelligence
for Federal Judges

by

James E. Baker

Laurie N. Hobart

Matthew Mittelsteadt

This Federal Judicial Center publication was undertaken in furtherance of the Center's statutory mission to develop educational materials for the judicial branch. While the Center regards the content as responsible and valuable, this publication does not reflect policy or recommendations of the Board of the Federal Judicial Center.

This publication was produced and published at U.S. taxpayer expense.

Contents

Four Questions Every Judge Should Ask About AI

> "AI is not a single piece of hardware or software, but rather, a constellation of technologies that gives a computer system the ability to solve problems and to perform tasks that would otherwise require human intelligence."[1]

Artificial Intelligence—AI—is an ever more pervasive part of our lives. AI is embedded in shopping algorithms, navigational aids, and search engines, and, as we now know, it is used for public health contact tracing. Studies show that certain AI applications identify tumors with greater accuracy than medical personnel. Algorithms drive social media—and, increasingly, cars. It seems Generation Z has come of age knowing nothing *but* algorithms.[2]

Just as AI is transforming the economy, health care, and American society, it will also transform the practice of government and law. Law firms use AI platforms to conduct discovery. At least seventy-five countries use facial recognition for domestic security and law enforcement purposes.[3] AI is used to determine travel patterns, to link suspects with crime scenes, and to populate watch lists. Between 2011 and 2019, the FBI used its facial recognition algorithm to search federal and state databases, including some state driver's license databases, over 390,000 times.[4] The National Security Commission on Artificial Intelligence (NSCAI) has predicted that "[t]he development of AI will shape the future of power."[5]

[1] National Security Commission on Artificial Intelligence (NSCAI), Interim Report 8 (Nov. 2019), https://www.nscai.gov/wp-content/uploads/2021/01/NSCAI-Interim-Report-for-Congress_201911.pdf.

[2] Algorithms, in this context, are mathematical formulas that guide software. Merriam-Webster defines an algorithm more broadly as "a step-by-step procedure for solving a problem or accomplishing some end." https://www.merriam-webster.com/dictionary/algorithm (last visited May 20, 2021). A familiar example is a recipe, which details the steps needed to prepare a dish. In a computer, an algorithm is implemented in computer code and details the discrete steps and calculations a computer needs to implement to complete a task. An algorithm is the "engine" an AI uses to "think" and make predictions.

[3] Stephen Feldstein, The Global Expansion of AI Surveillance, 1 (Sep. 2019), https://carnegieendowment.org/2019/09/17/global-expansion-of-aisurveillance-pub-79847; NSCAI, *supra* note 1, at 12.

[4] U.S. Gov't Accountability Off., GAO-19-579T, Face Recognition Technology: DOJ and FBI Have Taken Some Actions in Response to GAO Recommendations to Ensure Privacy and Accuracy, But Additional Work Remains (June 4, 2019).

[5] NSCAI, *supra* note 1, at 9.

Judges must understand how AI works, its applications, its implications for the fact-finding process, and its risks. They should be able to answer the following four questions in context:

1. How is AI being used in court or to inform judicial decisions?
2. Does the fact finder understand the AI's strengths, limitations, and risks, such as bias?
3. Is the AI application authentic, relevant, reliable, and material to the issue at hand, and is its use or admission consistent with the Constitution, statutes, and the Rules of Evidence?
4. Has an AI algorithm, a human, or some combination of the two made "the judicial decision," and, in all cases, has that decision been documented in an appropriate and transparent manner allowing for judicial review and appeal?

This guide addresses these questions by providing some technical background and highlighting some potential legal issues. We do not provide legal judgments about the use of different AI applications. In discussing how AI is used today and may be used in the future, we do not endorse that use in any particular context or application. Rather, we identify core concepts and issues, so that when judges decide whether to admit AI applications into evidence or to use AI in a judicial determination, they decide wisely and fairly. Making these decisions requires judges and litigators to know enough about AI to ask the right questions, at the right moment, in the right depth. It is up to the trial fact finders to determine the facts in each context and to judges to determine the appropriate application of law. We hope this guide helps.

1. An Overview of AI

From Turing to today

Popular and scientific literature identifies several benchmark events in AI development. In 1950, the English computer scientist and Bletchley Park code breaker Alan Turing wrote an article, "Computing Machinery and Intelligence." He asked, can machines think, and can they learn from experience as a child does? "The Turing Test" was Turing's name for an experiment testing the capacity of a computer to think and act like a human. A computer would pass the test when it could communicate with a person in an adjacent room without the person realizing they were communicating with a computer.

In 1956, Dartmouth College hosted the first conference to study AI.[6] The host, Professor John McCarthy, is credited by many with coining the term "Artificial Intelligence."[7] The funding proposal submitted to the Rockefeller Foundation stated:

> We propose that a 2 month, 10 man study of artificial intelligence be carried out…. We think that a significant advance can be made in one or more of these problems if a carefully selected group of scientists work on it together.[8]

Notwithstanding this optimistic start, progress in the field was neither linear nor exponential. It occurred in fits and starts. As a result, AI development went through a series of "AI winters," periods of low funding and low results. In the past twenty years, however, AI has emerged as one of the transformative technologies of the twenty-first century.

What changed? Experts point to several factors working synergistically—specifically, the development of complex algorithms, strides in computational speed, the invention of new sensors, an explosion in data, and the advent of cloud computing and machine learning.

Complex algorithms. Algorithms are the math equations embodied in software code that find, sort, and look for meaning in data. For a sense of scale and process, an algorithm can be as simple as the command "Insert." The Google search algorithm, in contrast, is said to consist of over two billion lines

[6] *Artificial Intelligence (AI) Coined at Dartmouth*, Dartmouth Coll., https://250.dartmouth.edu/highlights/artificial-intelligence-ai-coined-dartmouth (last visited Mar. 21, 2021).

[7] *Id.*

[8] Nick Bostrum, Superintelligence: Paths, Dangers, Strategies 6 (Oxford University Press 2014).

of code.[9] This code is dynamic, iteratively changed by its human engineers and, notably, decisions by the algorithm itself, which rewrites its code based on the accuracy of its prior predictions. There is no one, final Google search algorithm; the algorithm one uses today will be different from the algorithm one uses tomorrow.

Computational speed. The silicon transistors in computer chips, which drive computers, communicate in the form of electric pulses expressed in the form of zeros and ones. A one reflects a positive pulse of energy, a zero does not. Stringing those zeros and ones together creates computer code. The miniaturization of transistor circuitry allows an increasing volume of data to be processed in smaller and smaller spaces and thus with greater speed. For example, a 2016 iPhone 7 had the computational capacity of a 1985 Cray Supercomputer.[10] One of the defining characteristics of AI is its capacity to perform tasks at machine speed; computational capacity is at the heart of this feature.

Sensors. The development of sensor technology, such as that found in driverless cars, cell phones, and home devices, has resulted in more data and more applications for using that data to inform and influence human behavior. Personal assistants like Siri, Watson, and Alexa all use sensors to collect data.

Data. Data drives the AI revolution. As a general matter, the more data one has, the easier it is to train a computer system to perform a task or solve a problem, and likely the more accurate the result will be. (As we will see, the metrics selected in designing algorithms will also affect accuracy and the degree to which different forms of bias will affect accuracy.)[11]

Cloud computing. The advent of cloud computing allows more data to be stored on a permanent, retrievable basis. As the Supreme Court encountered in *Carpenter v. United States*,[12] unless purposefully deleted as a matter of law or policy, most data persist for years—likely forever.

Machine learning. It all comes together with a process known as machine learning, which refers to different methodologies to program software-driven machines to learn on their own and thus improve and optimize their functions. Much of machine learning research is predicated on trying to mimic the human brain (literally, in the case of efforts to replicate the brain using 3D printers) or with neurological metaphors like "artificial neural networks."

[9] Rachel Potvin, *Why Google Stores Billions of Lines of Code in a Single Repository*, YouTube (Sept. 14, 2015), https://www.youtube.com/watch?v=W71BTkUbdqE.

[10] Processing Power Compared, https://insightaas.com/infographic-processing-power-compared-1956-2015-experts-exchange.

[11] *See* remarks of Nisheeth Vishnoi at Yale Cyber Leadership Forum, "Session #1: Big Data, Data Privacy, and AI Governance," Feb. 18, 2022.

[12] 138 S. Ct. 2206 (2018).

While AI may mimic, and in some cases outperform, human intelligence, it is not actual human intelligence. It is machine capacity and optimization, hence the preferable term: Human-Level Machine Intelligence (HLMI). (Machine learning is explained in more detail in chapter 2.)

In sum, "AI is not a single piece of hardware or software, but rather, a constellation of technologies that gives a computer system the ability to solve problems and to perform tasks that would otherwise require human intelligence."[13] This means that each of the components comprising a particular AI application will be subject to potential legal challenge and validation. It also means that many AI components are iterative, evolving on an ongoing basis in ways that will make fixed case law precedent less useful.

AI now

Specialists refer to two types of AI: Narrow AI, which is where we are today, and Artificial General Intelligence, or Strong AI, which is where we are headed in this century. For our purposes, narrow AI can be defined as the "ability of computational machines to perform singular tasks at optimal levels, or near optimal levels, and usually better than, although sometimes just in different ways, than humans."[14] Under this umbrella come many single-purpose technologies, such as facial recognition, driverless vehicles, and drones, among others. These technologies are "intelligent" in only one domain, a limitation on their ability to be used for multiple purposes or deal with certain complex situations. All AI currently in use falls in this category. Today's AI is particularly good at correlating, connecting, and classifying data; recognizing patterns; and weighting probabilities, which is why it is good, and getting better, at tasks like facial recognition, image compression and identification, and voice recognition.

At present, narrow AI can be "brittle," by which engineers mean incapable of adapting to new circumstances for which it is not trained on its own and thus lacking in situational awareness. To illustrate the point, AI philosophers like to point to the thought experiment known as the Trolley Problem, now more commonly conveyed as a crosswalk problem. In the problem, a driverless car loses its brakes at just the moment it is coming up to a crosswalk at speed. There are various pedestrians in the crosswalk of different ages and of different perceived virtues—for example, a pregnant woman and an armed robber carrying a bag of stolen money. The car's computer must make a choice: swerve left, swerve right, brake, or drive ahead. An alert human driver, after first

[13] NSCAI, *supra* note 1, at 8.

[14] James E. Baker, The Centaur's Dilemma: National Security Law for the Coming AI Revolution 34 (Brookings Institution Press, 2020).

trying to brake, would make a values-based, ethical choice about where to aim the car, likely at the bank robber. The AI-driven car, on the other hand, unless it has been specifically trained to identify a "bank robber" or a "pregnant woman" and to adjust its decisional weights to favor one over the other in a choice scenario, will likely perceive the persons in the crosswalk as "persons in the crosswalk," no more. Chances are the software code will select the path of least numerical damage. This weakness in current AI is especially important where an AI application is likely to encounter changing or novel circumstances, like driving, or where there is an incentive for external actors to spoof or fool the AI application, as might be the case with military, intelligence, and law enforcement surveillance tools.

Recognizing AI's current level of sophistication, judges and lawyers must ask three questions: (1) Is the AI in question brittle? (2) What is the variance rate, i.e., the rate at which the AI reaches the wrong (false positive or false negative) result? and (3) Does the AI perform in the same manner and with the same accuracy in "real-world" operation as it does in lab conditions and testing?

Many narrow AI applications are known to consumers who rely on it daily. If you shop on Amazon, you are using AI algorithms. Amazon back-propagates training data from all purchases made on Amazon as well as data from individual consumers. Algorithms then identify patterns in the data and weight those patterns, allowing the algorithm to suggest (predict) additional purchases to the shopper. The algorithm adjusts as it goes based on the responses (or lack of responses) from recipients. This is an example of predictive big-data analytics. It is also an example of a push, predictive, or recommendation algorithm.

Why do companies use AI? Former Secretary of the Navy Richard J. Danzig explains:

> [M]achines can record, analyze and accordingly anticipate our preferences, evaluate our opportunities, perform our work, etc. better than we do. With ten Facebook "likes" as inputs, an algorithm predicts a subject's other preferences better than the average work colleague, with 70 likes better than a friend, with 150 likes better than a family member and with 300 likes better than a spouse.[15]

Narrow AI is also embedded in mapping applications, which sort through route alternatives with constant, near-instantaneous calculations factoring speed, distance, and traffic to determine the optimum route from A to B. Then the application uses AI to convert numbered code into natural language telling the driver to turn left or right.

[15] Richard Danzig, *An Irresistible Force Meets a Moveable Object: The Technology Tsunami and the Liberal World Order*, Lawfare Research Paper Series 1, 5 (Aug. 28, 2017).

AI computations and algorithms are also used to spot finite changes in stock pricing and generate automatic sales and purchases of stock as well as spot anomalies that generate automatic sales and purchases. All of this is based on algorithms created and initiated by humans but programmed to act autonomously and automatically because the calculations are too large, the margins too small, and the speed too fast for humans to keep pace and make decisions in real time. Of course, as one trader's algorithm gets faster, the next trader must change either his algorithm's design, its speed, or both to achieve advantage, reducing the window of opportunity for real-time human control even further. AI machine learning and pattern recognition are also used for translation, logistics planning, and spam detection, among many, many more commercial applications. In 2017, the former Chief Scientist for Baidu, Andrew Ng declared AI "the new electricity."[16]

Perhaps the most prominent illustration of next-generation AI is the driverless vehicle. AI empowers driverless cars by performing myriad data input and output tasks simultaneously, as a driver does, but in a different way. Human drivers rely on intuition, instinct, experience, and rules to drive—seemingly all at once—using the neural networks of the brain. In driverless cars, sensors instantaneously feed computers data based on speed, conditions, and images of the sort ordinarily processed by the driver's eyes and brain. The car's software processes the data to determine the best outcome based on probabilities and based on what it has been programmed to understand and decide. This requires constant algorithmic calculations that a human actor could not make in real time. Luckily, humans do not rely on math to drive cars. They exercise their judgment and intuition, which is why (if they're alert) they generally handle situational change better than AI applications do. On the other hand, AI does not fall asleep at the wheel, text while driving, or drive drunk.

Perhaps the most successful application of AI to date is found in the field of medical diagnostics. Here, narrow AI's capacity to detect and match patterns and find anomalies has led to breakthroughs in the detection of tumors as well as the onset of diabetic retinopathy. In places like India, where there is a shortage of ophthalmologists, the use of such screening diagnostics can help prioritize access to doctors and treatment by identifying at-risk patients.[17] Studies indicate that AI is generally more accurate than humans in detecting cancerous tumors. However, that is not the same as saying that humans are prepared to rely on AI alone, or wish to receive medical diagnoses from machines rather than doctors.

[16] *Why AI Is the 'New Electricity'*, Knowledge at Wharton (Nov. 7, 2017), https://knowledge.wharton.upenn.edu/article/ai-new-electricity/.

[17] Cade Metz, *India Fights Diabetic Blindness with Help from AI*, N.Y. Times (Mar.10, 2019), https://www.nytimes.com/2019/03/10/technology/artificial-intelligence-eye-hospital-india.html.

Law enforcement authorities use AI for predictive policing and surveillance. As NSCAI noted, "at least seventy-four … countries are engaging in AI powered surveillance, including many liberal democracies."[18] According to a 2019 Government Accountability Office report, the FBI has logged hundreds of thousands of searches of its facial recognition system, which has access to 641 million face photos.[19] The FBI reported its system has proven 86% accurate at finding the right person, if a search was able to generate a list of fifty possible matches.[20]

For each potential application, humans must decide whether it is wise and fair to use AI. Where it is employed, AI will generally be best used to augment rather than supplant human judgment, much as judges corroborate confessions rather than rely on confessions alone to determine guilt. One issue AI policymakers, designers, and ethicists must resolve in context is how to structure human-machine teaming to allocate responsibility and accountability. Judges in turn will have to determine whether as a matter of law, or a matter of law and fact, the humans made the correct decisions. Judges will also have to consider to what extent, if any, they should rely on AI applications to inform their decisions. Judges might wish to consider the next time they shop online or use a search engine the extent to which they would rely solely, if at all, on the recommendation of a shopping platform or the accuracy of the search algorithm to establish legal facts or determine legal outcomes.

Where AI is headed

Beyond narrow AI, computer engineers contemplate the emergence of Artificial General Intelligence, or AGI. AGI is an AI multitasking capacity that can serve multiple purposes. Much like a human, AGI can understand and perform multiple tasks and shift from task to task as needed. Most analysts foresee AGI arriving, if it arrives at all, as a stage in development, like the advent of flight, not necessarily a moment in time, like the Soviet launch of the Sputnik satellite in 1957. AGI will present more complex legal questions

[18] NSCAI, *supra* note 1, at 12. The number has increased since NSCAI published its report, which is why the number varies in this publication.

[19] Drew Harwell, *FBI, ICE Find State Driver's License Photos Area a Gold Mine for Facial-Recognition Searches*, Wash. Post (July 7, 2019), https://www.washingtonpost.com/technology/2019/07/07/fbi-ice-find-state-drivers-license-photos-are-gold-mine-facial-recognition-searches/. *See* U.S. Gov't Accountability Off., GAO-16-267, Face Recognition Technology: The FBI Should Better Ensure Privacy and Accuracy (May 16, 2016); U.S. Gov't Accountability Off., GAO-19-579T, Face Recognition Technology: DOJ and FBI Have Taken Some Actions in Response to GAO Recommendations to Ensure Privacy and Accuracy, But Additional Work Remains (June 4, 2019).

[20] Harwell, *supra* note 19.

than narrow AI. A system that can write and rewrite its own code as well as shift from task to task will be harder to regulate, requiring courts and legislators to wrestle with questions of accountability and responsibility for actions the AI takes or information it provides.

Experts have described three waves of AI development. The first wave of AI machine learning consisted of if-then linear learning, a process that relies on the brute-force computational power of modern computers. With linear learning, a computer is in essence "trained" that if something occurs, then it should take a countervailing or corresponding step. This is how the IBM computer Deep Blue beat Gary Kasparov in chess in 1997, a significant AI milestone. The computer was optimizing its computational capacity to sort through and weigh every possible move in response to each of Kasparov's actual and potential moves through to the end of the game. It did so with the knowledge of all of Kasparov's prior games, while on the clock in real time. Deep Blue was an impressive demonstration of computational force, an endless and near instantaneous series of if-this-then-that calculations.

We are in the second wave of machine learning now, the benchmark of which is AlphaGo, the Google computer that beat the world's best Go player in 2016. The AlphaGo victory was a milestone not just because Go is a more complex, multidimensional game than chess but because AlphaGo won using reinforcement learning: it got better at the game by playing it. AlphaGo improved with experience, adjusting its own decisional weights internally—in the so-called "black box" of internal machine calculation—without training data or other if-this-then-that learning.[21] Surpassing brute force computational power, this was a machine optimizing its capacity. Was it "thinking"? No. But was it learning? Yes.

Indications of a potential third wave of AI machine learning were already being discussed in 2016, the year AlphaGo won. As the *National Artificial Intelligence Research and Development Strategic Plan* reported in October of that year,

> [t]he AI field is now in the beginning stages of a possible third wave, which focuses on explanatory and general AI technologies . . . . If successful, engineers could create systems that construct explanatory models for classes of real world phenomena, engage in natural communication with people, learn and reason as they encounter new tasks and situations, and solve novel problems by generalizing from past experience."[22]

[21] David Silver et al. *Mastering the Game of Go Without Human Knowledge,* 550 Nature 354 (Oct. 19, 2017).

[22] National Science and Technology Council, Networking and Information Technology Research and Development Subcommittee, The National Artificial Intelligence Re-

Imagine a computer linked to the internet, the cloud, and the Internet of Things (IoT). Next, imagine that the computer is programmed not to play chess or Go, a single task and limitation, but to solve problems or answer questions generally. It moves fluidly from one task to the next. Now consider that if a computer could do that, it could not only write code, which computers can do now, but could rewrite, improve, and change its own code to optimize the task it was originally programmed to perform, or even find new, unanticipated ways to execute the task.

From 2015 to 2016, a group of scholars associated with the Oxford Future of Humanity Institute, AI Impacts, and Yale University surveyed "all researchers who published at the 2015 NIPS and ICML [Workshop on Neural Information Processing Systems and International Conference on Machine Learning] conferences (two of the premier venues for peer-reviewed research in machine learning)."[23] The survey asked respondents to estimate when HLMI would arrive. The study did not define AGI but stipulated that "Human-Level Machine Learning is achieved when unaided machines can accomplish every task better and more cheaply than human workers." Three-hundred and fifty-two researchers responded, a return rate of 21%. The results ranged across the board from never to beyond 100 years. What is noteworthy is that the "aggregate forecast gave a 50% chance of HLMI occurring within 45 years and a 10% chance of it occurring within 9 years." The two countries with the most survey respondents were China and the United States. The median response for the Americans was 76; for the Chinese, 28.[24]

As the survey indicates, experts do not agree on whether or when we will get to AGI. What we do know is that AI is already a transformative twenty-first-century technology. Moreover, AI tools and methods will continue to change at exponential rates. The courts, like other elements of society, must adjust to AI, just as they previously adjusted to computers and electronic filing. Preparation starts with an understanding of what AI is and is not, and the confidence that, explained in plain language, AI can be accessible to judges, litigants, and jurors.

Some philosophers and commentators contemplate the potential emergence of Super Intelligence (SI), a state of AI beyond general intelligence where computers are generally smarter than humans with unlimited sources of information and energy, care of the internet. When the Cambridge physicist Stephen Hawking remarked in 2017 that "AI may be the best thing to ever happen to humanity or the worst," he probably was contemplating something

search and Development Strategic Plan 14 (Oct. 2016).

[23] Katja Grace et al., *When Will AI Exceed Human Performance? Evidence from AI Experts,* 62 J. of Artificial Intelligence Res. 729–54 (July 31, 2018).

[24] *Id.*

akin to SI. AI philosophers like Nick Bostrum and Elon Musk have garnered headlines with apocalyptic predictions about SI.

Many computer engineers and government officials dismiss SI as science fiction and a distraction from the real and immediate challenges of today's narrow AI. Judges should be aware of the concept of super intelligence, however. This line of AI inquiry tends to dominate popular AI literature and movies. Nick Bostrom's book *Superintelligence* was a world-wide bestseller; presidential reports on AI are not. Therefore, the shadow of SI can be expected to impact the way jurors perceive AI evidence. Further, because expert testimony may allude to the concept, judges need to be able to place it in context in qualifying experts to testify and instructing jurors. Conscious that jurors will come to the subject of AI from many different angles, including SI and science fiction, judges need to be able to ask the right questions, understand the underlying technology, offer clear and careful jury instructions, and state on the record their evidentiary analysis.

Machine learning in a nutshell

Machine learning (ML) is one of those terms that defines itself. Machine learning (ML) is the process by which a machine learns to perform tasks and improve on the performance of those tasks. *How* machines go about "learning" and with what degree of autonomy, change, and accuracy is the complicated part. Specifically, learning, as applied to machines, refers to the mathematical means by which they identify, aggregate, and derive meaning from data. When designing machine learning algorithms, engineers use three kinds of ML data sets: training data, validation data, and testing data. Training data is intended for the AI, labeled and curated so that the AI can analyze it, learn from it, and adjust its coding to ultimately form better predictions. Validation data is intended for the developer, who uses this data to stress test the model's training and decide whether he/she must update its settings and sensitivities. Testing data provides a final round of analysis; this data is used on the trained, tuned model to evaluate the model's fit to the data and overall accuracy.

In general, there are three ways to teach machines to learn: supervised learning, unsupervised learning, and reinforcement learning. Any of these might entail "deep learning," which involves the use of artificial neural networks within the machine to break down data and make predictions about its meaning. But deep learning is not the only method by any stretch. In addition, there are multiple mathematical theories, equations, and methods by which machines learn. Whatever the method, ML is usually continuous, which means the use of AI should itself include an ongoing validation process.

Precisely because ML-driven AI continues to learn as it operates (and thus, presumably, gets better at performing the tasks and solving the problems for which it was programmed), the relevance and reliability of AI output can be moving targets—a fact judges and litigators need to understand. Judges further need to remember that opening discovery or testimony to AI/ML potentially opens the door to an array of subordinate questions involving methodologies, data, and testing. Judges may thus have to determine where it is essential for fact finders to understand the underlying methodologies or just the conclusions or results derived from those methodologies.

The following section outlines one form of machine learning: supervised learning using a deep learning neural network. It is an illustration of one AI/ML methodology, intended to demonstrate the myriad questions that may arise when courts consider AI generated evidence or when judges utilize AI. These questions and answers may vary depending on the AI methodology and use.

2. Machine Learning Illustrated: Supervised Learning

From images to numbers and back again

In computing, almost anything—pictures, film, text—can be expressed as a number. So can thoughts and descriptions, including socioeconomic data and descriptors. There is concern, however, about the degree to which computers can express qualities accurately, and without bias, in context.

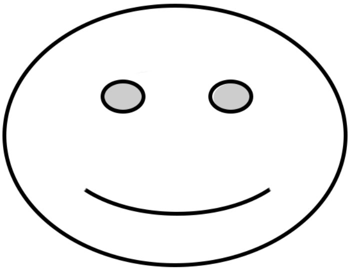
Computers are electrical devices made of wires and microchips that must communicate with electrical signals. By convention, these signals have been represented by numbers: 1 represents when a component is “on” and sending a signal, while 0 indicates when a component is “off.” Just as humans have agreed that certain letter combinations represent words, computer engineers have adopted conventions and standards that dictate how data can be represented by combinations of 1s and 0s. For instance, a standard adopted in the 1960s dictates that the letter ‘A’ is represented by the number 01000001. Computer engineers operate on the assumption that if you can express an idea or a task in these numbers, you can program a computer to perform that task.

Advancements in sensing technology allow today’s computers to scan an image, analyze it, and generate a number that represents a portion of the image. When humans see a photograph, they might see a picture of a dog. A computer, depending on how it is programmed, “sees” hundreds, thousands, or millions of pixels. Each pixel in turn, depending on the design and resolution of the imagery, is broken down into thousands of 0s and 1s. Depending on the software design, each pixel can be broken down into quadrants and numbers. These numbers are often written as binary numbers, a system of counting with those 1s and 0s. For simplicity, clarity, and brevity, the explanation that follows uses the more familiar Arabic numeric decimal system.

Let us stipulate that the number 1224 is code for a smiley face (in binary code this would be represented as 11101110111010100). In a computer, “1224” typically is not read as “one-thousand two hundred twenty-four” but as the number sequence 1-2-2-4: the numbers represent data, not a quantity. Each digit in a number represents a specific feature of the data it describes. In our case, let us further stipulate that the digits in our string of numbers describe the following features of a smiley face:

1	2	3	4
Face Shape	Left Eye Shape	Right Eye Shape	Mouth Shape

The first digit in our number, 1, fills the "Face Shape" category, indicating a circular face. If the first digit was a 0, 5, or 6 instead, it would represent a different face shape. Moving to the right, the next two digits are both 2. The computer knows that digits in these locations describe the shape of each eye: in this case, 2s mean the eyes should be circular. Finally, the last digit, 4, describes the mouth: in this case, 4 signals that the mouth should be a smile. This simple example shows how a number like 1224 can be understood by the computer to signify this:

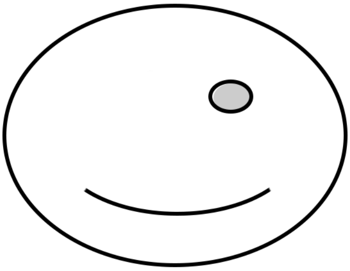


Of course, using binary numbers, 1224 would be represented in lines of 1s and 0s embedded in lines of code. Since any data can be represented in the form of a number, a computer can use math to understand and process any type of data. This concept underlies artificial intelligence and accounts for some of its strengths and weaknesses.

How AI "thinks"

At its heart, an AI algorithm is a mathematical formula that takes input (an image, video, etc.) represented as a series of numbers and analyzes those numbers to determine what they represent (as with facial recognition) or to find hidden links between the numbers (as in link analysis).

Let us say we have an AI application designed to detect simple faces. This AI is well trained and is already coded to understand that 1224 is a smiley face. However, what happens if it analyzes this slightly different face, represented by 1024?



In this case, the AI's algorithm is not robust enough to capture each variation of faces and might not know a one-eyed face is still a face. The algorithm recognizes 1224, but not 1024. To expertly identify all facial variations, the computer will need to *learn* what different varieties of faces exist, so that it can properly identify them. In theory, and sometimes in practice, computer engineers could program the algorithm to recognize every variation in facial feature: 1024, 1025, 1026, etc. In a more complex AI application, that approach might require vast amounts of data and computation. One begins to see how the volume of data—for example, the number of state driver's license pictures—can affect predictive accuracy.

This is where *deep learning* comes into play. Deep learning is a process by which a computer crafts mathematical formulas that can take in data like 1224 or 1227 and learn that each of these numbers represents a different face shape. To learn, the computer needs to be trained.

The learning process

Much AI is modeled, metaphorically, on the human brain, and like the brain, it uses "neurons" to analyze and communicate information. These interconnected neurons are referred to collectively as an *artificial neural network.* When AI receives numerical input, it passes the number through one of its artificial neurons, which analyzes the number and determines its significance. Each neuron is assigned a mathematical formula. If the result of this equation meets certain threshold qualifications, then the neuron "fires," sending its analysis to other neurons. The formulas, however, must be accurate to get accurate results. This is where training comes in.

Let us say that the neuron in our AI face identification program starts with the simple formula:

Face Shape + LeftEye + RightEye + Mouth

The neuron's coding determines that if the formula's result is greater than or equal to 9 (the sum of our face digits 1+2+2+4), then it knows to tell the computer the data is indeed a face. For 1224, this works. But if you use the number representing our one-eyed face, 1024, the coding does not work: 1+0+2+4 = 7, which is less than 9, so the neuron will not fire.

To solve this problem, the formula will need to change. That is why AI algorithms are coded to learn. To teach an algorithm, data are fed through it to teach it how to adjust its formulas to capture variations in data. The training data are often labeled, so after a neuron determines what it *thinks* the data means, it checks its answer. If the answer is wrong, it knows it must change its formula so that in the future it can get the correct answer. This"guess and check" learning process is called *supervised learning.* As explained later, it is also during this ongoing process that certain design flaws, assumptions, and biases can shape and, sometimes, undermine the accuracy of the system.

Learning from mistakes

Perhaps in its next iteration, the formula will be adjusted to this:

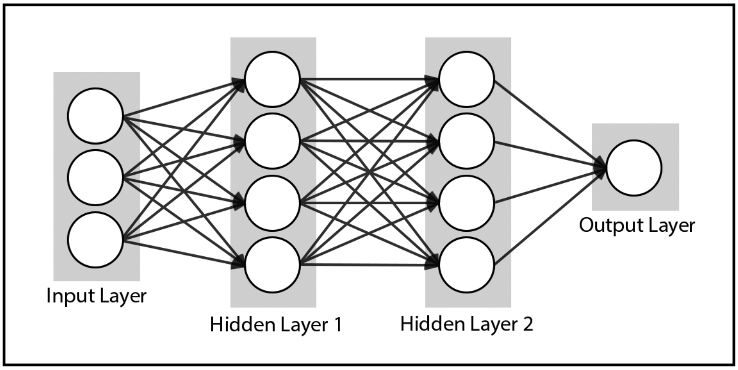
Face Shape + LeftEye + RightEye + Mouth + 2

This new formula adds a 2 to the final result, factoring in cases when the face is missing an eye. With this change, the neuron will be able to correctly identify the one-eyed face. This is a simple example of how an AI algorithm adjusts and learns.

Simplifying complexity

Of course, our smiley face example belies the immense complexity of real-world AI. We are using a simple image comprising four factors and a single neuron as an example. Most analysis is not nearly so simple. Analyzing a human face is nuanced and should require large networks of neurons to break down the face's individual features. A facial recognition algorithm might break down the eye into its individual elements, perhaps first identifying the shape of the eye socket, then the outlines of the eyes, and finally details such as the pupil and iris. The complexity of analysis commonly requires breaking data into manageable bites and organizing that data into logical categories. Breaking down data, organizing it, and recognizing it to make an identification or find a comparable match takes multiple rounds of neural analysis. Most neural networks have multiple layers, potentially hundreds and even thousands of layers, depending on their design. At a high level, these include an input layer that takes data and gives a first stab at identification, hidden layers that further analyze the data, and an output layer that spits out the final result.[25]

[25] Tim Dettmers, *Deep Learning in a Nutshell: Core Concepts*, NVIDA Developer Blog (Nov. 3, 2015), https://devblogs.nvidia.com/deep-learning-nutshell-core-concepts/#feature-learning.



[26]

Complexifying simplicity

Supervised learning is one of the easier AI methodologies to understand. As the next section explains, though deep learning and neural networks are the most common AI methods, they are not the only ones. AI is a dynamic field, and new developments could change processes at any time, accelerating how algorithms process and find meaning in data.

[26] Luke Dormehl, *What Is an Artificial Neural Network? Here's Everything You Need to Know*, Digital Trends (Jan. 6, 2019), https://www.digitaltrends.com/cool-tech/what-is-an-artificial-neural-network/.

3. Judicial Roles and Nine AI Takeaways for Judges

Judges will play at least four roles when it comes to AI in the courtroom. First, they will serve as evidentiary gatekeepers, applying the Federal Rules of Evidence (or state equivalents) to proffers of testimonial and documentary evidence, including and perhaps especially Rules 401, 402, and 403. Second, judges will serve as guardians of the law, specifically the values embedded in the Bill of Rights as well as statutes and rules of procedure and evidence. Third, judges may serve as potential AI consumers who need to decide whether to receive or rely on AI-generated outputs to inform bail, probation, and sentencing decisions. Fourth, judges will serve as communicators, translating the sometimes complex inputs behind AI into plain-language instructions for jurors and case law precedent for lawyers. The previous sections introduced the technology behind AI. This section identifies nine features of AI about which judges should be aware in their roles as gatekeepers, guardians, potential consumers, and communicators.

1. There are many different methodologies.

Because there are different AI methodologies, each application should require authentication and validation not just in concept but as applied in each context. To illustrate, so far, we have described an approach to ML called *deep learning* using the *supervised learning* method of teaching machines, so called because our algorithm is fed labeled data weighted to identify the correct answer. This method allows the machine to adjust its algorithmic equation to better predict outcomes, i.e., identify an image correctly as a face and, more specifically, a smiley face. In our scenario, that means passing along images or components of images that meet a certain threshold of accuracy or confidence, while discarding components that fall below the threshold. This process happens at each stage of the neural network inside the "black box," that part of a neural network between input and output layers. It is called the black box because engineers cannot always be sure what parameters an algorithm has relied upon within the neural network and with what weight. However, increasingly, there are methodologies engineers can incorporate to make such internal calculations more transparent, or fully transparent, to users, and judges should inquire whether the methodology employed is the most accurate, reliable, and trasparent available.

There are other theories and methods for teaching computing machines to learn, each built into the operative algorithm. These alternatives include evolutionary or genetic algorithms, inductive reasoning, computational game theory, Bayesian statistics, fuzzy logic, hand-coded expert knowledge, and analogical reasoning.[27]

As noted in the introduction, within the category of machine learning, there are multiple ways to teach a machine to learn using data.[28] The three most common are supervised learning (illustrated in the last section), unsupervised learning, and reinforcement learning. Unsupervised learning is a technique for teaching a computer to find links and patterns in large volumes of data without a determined outcome in mind. In contrast, supervised learning matches a data point, such as an image, to a known database of labeled data. The government might use an unsupervised learning methodology to search for meaningful patterns and hidden links in phone call records, travel patterns, or trade and commerce records indicating sanctions violations. Here, the algorithm is not searching for a particular number or face but for meaning in otherwise unstructured data. Importantly, it might also find connections without meaning, for example, by "matching" faces in a facial recognition application with similar backdrops or lighting. When this occurs within a neural network, it may be difficult, or impossible, to discern that the "match" is based on a factor irrelevant to the output objective.

Reinforcement learning introduces a "change agent," either an incentive or a desired goal, into the algorithmic code that might cause the machine to weight or improve its outcome on its own, as in the case of AlphaGo. A shopping algorithm, for example, might do this by automatically adjusting its code based on whether a recommendation is accepted, rejected, or ignored. In addition to deciding on the learning methodology, computer engineers must also decide how much depth and breadth to apply to a deep learning neural network—in other words, how widely the algorithm will search (breadth, also referred to as width) and how many layers of internal inputs and outputs it will employ before providing an output (depth). With facial recognition, for example, breadth might represent the number of data sets an algorithm searches. Depth might be illustrated by the number points on a face the algorithm is programmed to analyze before providing an output. Increases in network size tend to be required to capture the complexity of modern AI algorithms. Such increases create a challenge, however: the greater the depth—the number of layers in the neural network—the harder it will likely become to

[27] United Nations Institute for Disarmament Research (UNIDIR), The Weaponization of Increasingly Autonomous Technologies: Artificial Intelligence 5 (2018), http://www.unidir.org/files/publications/pdfs/the-weaponization-of-increasingly-autonomous-technologies-artificial-intelligence-en-700.pdf.

[28] *Id.* at 3.

determine which factors were determinative in the output prediction. This could become important to the extent there is risk or concern that bias or some other factor might undermine outcome accuracy. It is also why many algorithms are designed to provide output*s*, plural—for example, a range of match faces with a facial recognition algorithm, or a range of products with a shopping recommendation algorithm.

A court will need to satisfy itself that the specific AI application (as opposed to AI generally), its design, and its specific use meet the foundational requirements for the purpose for which it is being offered into evidence or used by a court. Verification will entail considering the theory and method behind the AI, including the nature of the data sets used to train, test, and validate the AI, as well as, in the case of deep machine learning, inquiring into the design of the neural network.

2. Most AI is iterative and should be tested and validated continuously.

AI/ML learns as it proceeds. That means AI systems need to be tested and validated on an ongoing basis. In other words, if a machine is learning, its variance rates and accuracy should change as well—for better or worse, depending in part on the real-world data it encounters and how well that data corresponds to training data. Experts stress the importance of three types of AI data to ongoing testing: training data, validating data, and testing data.

Training data are used to train an algorithm and thus might be curated in a particular manner to highlight features or be labeled so that the algorithm knows when it has identified the correct result and can adjust its internal weights accordingly. Validating data sets, as the name implies, are used to validate that the AI will work as intended but also to determine strengths and weaknesses in the AI. A validating data set for facial recognition may, for example, use gender- or age-based images to determine if an algorithm performs equally well across demographic parameters. Testing data are used to assess the accuracy of the AI on an ongoing basis, as well as the ability of the AI to react to the unknown (untrained data and circumstances).

Underlying data that are biased or poorly selected for these functions may undermine the accuracy of the AI or embed bias in the AI's application. In context, judges, experts, and litigators will have ample opportunities to test the reliability of any AI evidence offered in court, and judges will need to determine in context just how wide to open the door to expert testimony and discovery about matters like data sets, algorithmic design, search parameters, bias, and neural network architecture.

3. Humans are always involved.

Machines do what they are programmed to do, not because they choose to do so, but because they are programmed to do so, including learning on their own. Software drives machines. And humans, in the first instance, write software and design programs. Behind each AI application there are human choices, human values, and human bias that may impact the operation of the algorithm and the accuracy of its results. Humans select not only the data but also the metrics the algorithm uses to frame and analyze that data.[29] Algorithmic bias is such an important issue that this guide devotes a separate section to it, in chapter 4.

In the operation of AI, too, humans are involved. Under current vernacular in the AI field, humans are said to be "in-the-loop," "on-the-loop," or "out-of-the-loop." As implied, in-the-loop describes humans in functional control of an application, deciding when and how it is used. On-the-loop describes humans observing AI but not controlling it, but with the option to do so. Out-of-the-loop describes an autonomous or semiautonomous system operating automatically. These terms are imprecise in at least two regards. First, they describe a wide variance of conduct within each category and thus may convey a sense of control and oversight that is, in operation, absent. More to the point, they are insufficiently descriptive to apportion accountability and responsibility for the purpose of legal judgments. Take the example of a "driverless car." Some "driverless cars" are configured to employ a safety driver as an observer or, in the case of a semi-driverless car, a driver with shared responsibility for the operation of the vehicle. Other "driverless cars," without a human in the car, may operate under remote human control. Thus, in each of these three scenarios, at any moment in time the vehicle may be driving autonomously without human control, it may be following the explicit direction of the remote or present driver, or the human driver may be keenly observing the operation of the vehicle without overriding the car's computers. In other words, in each case, humans were out of, in, and on the loop.

However described, a human is always involved with an AI application. For courts, the factual questions will be: Who designed the seed algorithm? Using what metrics or weights? Who trained the algorithm? Using what data? Who collected the data? Who validated the data? Who used the algorithm or monitored its use? These factual questions will lead to legal questions. For example, where *Crawford*[30] applies, multiple persons might be called as witnesses regarding the design and operation of an AI algorithm. Because humans are always involved with AI, there will be persons who can, if relevant and

[29] Remarks of Nisheeth Vishnoi, *supra* note 11.

[30] Crawford v. Washington, 541 U.S. 36 (2004).

material, provide answers to the sorts of questions essential to authenticating and validating the use of AI:

- What is the AI trained to identify, how has it been weighted, and how is it currently weighted?
- Does the system have a method to transparently identify these answers? If not, why not?
- Are the false positive and false negative rates known? If so, how do those rates relate to the case at hand?
- How has AI accuracy been validated, and is the accuracy of the AI updated on a constant basis?
- What are the AI's biases?
- Is authenticity an issue?
- How do each of these questions and answers align with how the AI application is being used by the court or proffered as evidence?

Judges might also consider that a qualified AI expert or witness ought to be able to credibly answer these questions, or perhaps the expert or witness may not be qualified to address the application at issue.

4. AI predicts; it does not conclude.

As the previous example of the smiley face illustrates, AI is generally a predictive tool based on statistics. Through weighted calculation an algorithm predicts an outcome—in our case, that the image presents a smiley face. What the algorithm does not do is confirm that the image presented is a smiley face in the same way that a chemical test confirms the presence of a compound. This is why engineers use the term "confidence threshold" in describing the accuracy of an application.

In the case of a Google search algorithm, for example, the algorithm is predicting that one of the provided links will respond to the query. This is self-evident if one asks a question like, "Who was George Washington?" The algorithm is likely to provide a Wikipedia link to a webpage about the first U.S. president. It is also likely that many readers will conclude that the algorithm has *answered* the question: "George Washington was the first President of the United States." It has not. The algorithm has predicted that one or more of the links provided will answer the question, and likely in descending order of probability as the links are listed. Modify the question a bit, and the predictive aspect becomes more evident. If you ask, "Who is my friend George Washington?"—a quite different person than the first president—Google responds with sites listing the first president's friends. That is the algorithm's best prediction as to which links will answer the question based on code matching, likely use of the word "friend" and the way prior readers have responded to similar word searches. In other words, like a shopping algorithm, the search algorithm is tracking whether the searcher "bought" the response by clicking on it and measuring how long the searcher stayed. Of course, it has not answered the question at all and is nowhere near to providing a link that will answer the question about the user's friend George Washington—not without more details that can help shape the predictive outcome.

Now let us consider what this predictive quality means in a more realistic legal context. Instead of asking if a picture depicts a smiley face, an input might query, "Is this a picture of Al Capone?" Or one might input the picture of a person robbing a bank to see if there is a picture in a state driver's license database that matches the picture in some or all characteristics. The FBI facial recognition algorithm, for example, is designed not to conclusively find a match but to find pictures that might match, like Google links. As the GAO report on the subject stated, the algorithm is most accurate when offering a range of potential matches.

In a medical context (perhaps coming before a court in a malpractice case) an input might query, "Is this a picture of a benign or malignant tumor?" To respond to that question, an algorithm trained on prior pictures of tumors might break the picture into quadrants and subcomponents, as a facial algorithm might do, and then compare the picture submitted to database images of tumors. Based on all accessed images of benign and malignant tumors, the algorithm will predict whether the picture is a better match for one or the other. What the algorithm offers, which a human does not, is the capacity to search multiple databases rapidly for comparative patterns, as well as the ability to break the image into subordinate components in a way humans cannot, and thus to see connections and patterns the human eye cannot. Moreover, the algorithm is neither affected by fatigue nor subject to ordinary human distractions, pressures, and emotions.

If the algorithm has not been trained properly, or trained to identify new patterns, it is less likely than a human to identify a rare disease or new manifestation of an existing disease, raising the prospect of a false negative. One can imagine in a malpractice case how the parties might litigate the manner in which any human-machine teaming occurred. Where a tumor was not diagnosed, a plaintiff might argue that doctors placed unreasonable reliance on a "negative" AI output. Alternatively, a plaintiff might argue an unreasonable lack of reliance, if a broader use of AI databases was not employed.

5. Accuracy depends on the quality and volume of data.

If an algorithm has only been trained on one picture of a cancerous tumor or has never seen a cancerous tumor, then it will be less likely to correctly identify a tumor in response to a query. In our smiley face example, the algorithm is not capable of discerning a one-eyed face, absent the necessary training to identify a one-eyed face. This is an important limitation on the capacity and accuracy of current AI. Moreover, volume here is not measured in hundreds, but in hundreds of thousands of pictures. A human performing the same task with only one picture will more likely identify a tumor using intuition, judgment, and experience, as well as external factors the algorithm cannot assess, like the patient's unique pain threshold or situational responses to touch and feel.

The quality of data is also important. Dated data, known as stale data, is more likely to generate inaccurate results. A facial recognition algorithm trained on driver's license pictures or parole pictures is more likely to identify pictures reflecting the demographics represented in the databases. This has the potential to increase the false negative rate for underrepresented groups and to increase the false positive rate for overrepresented groups.

Likewise, data may possess flaws that impact algorithms but not humans. Algorithms may discern links in data or perceive patterns in data creating matches, based on elements or numeric formulas that are unintended or that humans would not discern. In our bank robber scenario, the algorithm may match numbers and pixels based on irrelevant factors, such as a common backdrop in a photo or pattern on the robber's face mask. In either instance, there may be a match but not a meaningful match. If this occurs within the neural network, it may skew a result in a manner unseen and unknown to the user. The output is a face, but the user may not know this face has been passed through to the output stage because of similarities in the picture backdrops, not the face itself.

As will be seen, this limitation makes certain predictive algorithms particularly susceptible to error. It is essential that judges and fact finders understand the ways data can embed witting and unwitting bias, as discussed at length in the bias section, into algorithmic design, impacting the predictive accuracy of AI.

6. The heart of AI is the algorithm.

If the accuracy of an AI application often depends on the amount of data on which it is trained, it depends even more on the algorithm applied to that data. As noted previously, an algorithm is a mathematical formula that guides the software determining which data are selected and how they are weighted. The *choice* of algorithm is a choice of decisional metrics or framework—an analytical lens or value-laden perspective.

Think of an algorithm as the recipe a chef uses in a kitchen. The chef chooses not only the end-dish but also dietary restrictions (vegetarian, low sodium), flavor profile (sweet, acidic, spicy), cultural heritage, ingredient measurement system (metric, English), etc. As this is AI, not a simple algorithm, the chef supplements her recipe every time she cooks in a way that only she knows. What is more, the sous chefs supplement the recipe when no one is looking. Thus, in some cases no one can be quite sure what gives the recipe its distinctive taste; and, if the chef knew, she would not tell because she wants customers to continue to come to her restaurant. Restated, if one knew the Google search algorithm, every search platform could, in theory, be as good, provided of course that the algorithm could access and apply the same level of data (Google's data) with which to train the algorithm.

The heart of many disputes about the use of AI-generated evidence in court or the use of AI tools to inform judicial decision making will revolve around access to and disputes over the accuracy of algorithms. This is the proprietary secret most AI companies want most to protect, because it is the recipe to their market success and because too much inquiry may undermine confidence in the AI's capacity.

Here are several questions judges should contemplate before using an AI application or admitting one into evidence:

- To what extent will the court allow parties to discover the content of an algorithm? The data on which the algorithm was trained, tested, and validated?
- If discovery is permitted, what safeguards, if any, will the court use to protect the proprietary value of the discovered information?
- In the context presented, does due process require access to an underlying algorithm or its supporting training, validation, and use data?
- To what extent is such discovery necessary to apply *Daubert*? (See chapter 7.)
- In the context presented, can ex parte and in camera judicial review adequately and legally substitute for public adjudication? Or should the parties or the public have access to the algorithms and data?

These questions might lead to the further questions: Will the court or a jury be able to understand the underlying technology, and is such understanding necessary for a fair adjudication of the facts? If so, what is the appropriate mechanism to provide that understanding?

7. Narrow AI is brittle.

As noted earlier, narrow AI is not particularly good yet at situational awareness. The driverless car may not timely identify novel objects on the road. Judges will therefore have to consider whether the scenario or fact for which an AI application is offered presents questions involving situational awareness. If so, they should then ask in what manner the algorithm, the data, and the training are keyed for such circumstances and whether the accuracy rate varies in such contexts.

8. AI is also nimble.

Proponents of AI tend to emphasize its strengths, opponents its weaknesses. Of course, the strength and weakness of any AI application must be assessed on a case- and application-specific basis. One strength of AI, however, is its general capacity to identify, aggregate, and derive meaning from data in ways that humans cannot. With driverless vehicles, for example, this capacity is simply illustrated by the fact that, with the right sensors, driverless vehicles can "see" in all directions at once and calculate, with mathematical precision at speed, the distance needed to brake. AI can also see patterns, anomalies, and links in data that humans cannot. In many cases, AI is better than humans at tasks like comparing pictures of tumors to database images of benign and malignant tumors. And AI can do all this at machine speed. Of course, depending on the context, humans will need to determine whether the patterns and links that are made are relevant and reliable for the purpose presented.

9. AI is biased.

As with humans, AI has biases. Judges and litigators need to be attuned to the different ways that bias can influence AI accuracy and transparency, which is why we devote the next chapter to the topic.

4. Bias

Bias is often associated with the human application of stereotypes or prejudices to an ethnic, gender, racial, or other identity group. In U.S. law, such categories are generally recognized as “suspect classes” in equal protection law under the Fifth Amendment, as applied to the federal government, and the Fourteenth Amendment, as applied to individual states.

As judges well know, any application of law that treats classes of persons differently from the populace as a whole, if challenged in court, must pass either strict scrutiny, intermediate, or rational basis review, depending on the class. Racial classifications, for example, receive strict scrutiny requiring the government to show (1) a compelling government interest for the disparate treatment and (2) that the means used are narrowly tailored to accomplishing the compelling interest. Gender, in comparison, is subject to intermediate scrutiny, in which case the disparate treatment must further an important government interest and do so by means substantially related to the interest. An application of law that is facially neutral but adversely affects one protected group more than another might also be subject to a disparate impact claim. For example, a hiring algorithm that disproportionately favored one group over another might be subject to a disparate impact lawsuit.

With AI, bias is usually defined more broadly as a witting or unwitting (conscious or unconscious) predisposition that can undermine the accuracy of an AI application or output. Bias thus addresses a range of cognitive tendencies that can adversely affect objective analysis and technical accuracy. Significantly, AI “bias” also incorporates and describes unintentional design and data flaws that can impair the accuracy of AI outputs. Because humans design AI algorithms and choose the data that “trains” the software, developers’ biases can be baked into the algorithm’s design. Unintentional bias is often difficult to discern because it is embedded in the design of an AI system or in the data used to train an algorithm. Decision makers may subsequently place undue reliance on AI outputs predicated on biased input.

When it comes to potential algorithmic bias, there are four immediate takeaways for judges:

1. Judges (and the law) use the term *bias* in a different and more specific way than computer engineers. For AI specialists, *algorithmic bias* refers broadly to the difference between an algorithm’s output and the desired outcome, not necessarily to bias of the sort addressed by the equal protection clause.

2. Algorithmic bias can be caused by human prejudice of the sort courts typically address, cognitive bias of the sort behavioral scientists typically address, design and data flaws of the sort computer engineers address, or all of the above.
3. As bias is defined above, there is *no such thing as a bias-free algorithm*. There is a tendency to believe that "numbers are neutral" and present objective truths, but numbers may produce erroneous results.[31]
4. Through careful engineering, thoughtful use of data, and adjusted algorithmic weights, it is possible to create AI systems with lower margins of error.[32] It is also possible that reducing one form of bias by adjusting, for example, the underlying analytical framework or data sets can allow other forms of bias to creep in.

Judges, as evidentiary gatekeepers, can mitigate or bar the use of weak or biased AI by asking the right foundational questions. Knowing what to ask starts with an understanding of the forms that algorithmic bias might take.

Forms of algorithmic bias

The United Nations Institute for Disarmament Research suggests several categories and sources of algorithmic bias.[33] Starting with the Institute's findings, this guide highlights eight forms of potential AI bias: statistical bias, moral bias, training data bias, inappropriate focus, inappropriate deployment, interpretation bias, unwitting human bias, and intentional bias. This section also discusses the issue of overfitting as a potential source of bias. What is the judicial takeaway? Judges do not need to be experts on every type of bias, or for that matter AI. They do need to know that there are many different ways that bias can skew the accuracy of AI outputs. Armed with this knowledge, judges will need to ask the right questions to determine how much leeway to allow litigators to probe the accuracy of AI outputs and the algorithms, data, and training that have produced the AI outputs.

Statistical bias might occur when an algorithm's predicted outcomes devi-

[31] Joni R. Jackson, *Algorithmic Bias*, 15 J. of Leadership, Accountability & Ethics 55–65 (2018), https://search.proquest.com/docview/2170233068?accountid=14214.

[32] Jake Silberg & James Manyika, *Notes from the AI Frontier: Tackling bias in Artificial Intelligence (and in Humans)*, McKinsey Glob. Inst. (June 6, 2019), https://www.mckinsey.com/featured-insights/artificial-intelligence/tackling-bias-in-artificial-intelligence-and-in-humans.

[33] United Nations Institute for Disarmament Research (UNIDIR), *Algorithmic Bias and the Weaponization of Increasingly Autonomous Technologies: A Primer,* 9 UNIDIR Resources (2018), http://www.unidir.ch/files/publications/pdfs/algorithmic-bias-and-the-weaponization-of-increasingly-autonomous-technologies-en-720.pdf.

ate from a statistical standard, such as the actual frequency of real-world outcomes.[34] This deviation can be caused by bad statistical modeling or incorrect or insufficient data. The difficulty of calculating the infection or mortality rate of a disease such as COVID-19 illustrates the problems that can result. At the outset of the pandemic, AI-driven modeling of infection rates differed widely, in part because the models could not account for who had the disease without showing symptoms. Thus, it was only within closed data samples, such as the passengers aboard a cruise ship, that the models could account for an asymptomatic pool (because all the passengers were tested before they were allowed to leave the vessels). The risk with cruise ship findings was having too small a sample pool, and one not necessarily random or representative of a cross-section of the population.

Moral bias occurs when an algorithm's output differs from accepted norms (regulatory, legal, ethical, social, etc.).[35] For example, an algorithm may weigh factors that the law or society deems inappropriate or do so with a weight that is inappropriate in the context presented. A predictive crime algorithm might use data derived from the current prison population to "predict" rates of recidivism. Given the disproportionate number of people of color imprisoned, the algorithm likely would produce biased results by explicitly or incidentally weighing "race" or its proxies, and perhaps within the black box. AI neither thinks nor understands the world like humans, and unless instructed otherwise, its results can reflect an ignorance of norms found in the equal protection and due process clauses.

Training data bias. Like humans, AI learns from experience; however, its experience is based exclusively on data, often hand-selected by a human developer. Inaccuracies or misrepresentations in this data can perpetuate biases by embedding them in algorithmic code.[36] In other words, the results are skewed; the algorithm produces wrong answers. For example, an algorithm intended to identify potentially successful job applicants might rely on past successful job performance as an indicator of future successful job performance and derive from that data certain preferred hiring characteristics like age, school, and experience. But if the data are from a period when women or other marginalized groups were not well represented, in numbers, in the relevant employment market or educational pool, at least 50% of the potential workforce might be excluded from results. Such data might likewise incorporate human bias in the form of a past company policy to only hire persons from certain schools. The criterion might have seemed objective when the company adopted the policy,

[34] *Id.* at 2.

[35] *Id.*

[36] *Id.* at 2–3.

but it necessarily incorporates the socio-economic and other biases of the college admissions processes of the time. Thus, the algorithm might exclude candidates as good or better than those from whom the dated data set was sourced.

As discussed later, similar concerns have been raised about algorithms designed to inform parole decisions by predicting recidivism risk. These algorithms, critics argue, rely too heavily on socioeconomic status, neighborhood location, past crime statistics, policing practices, and prosecutorial decisions as predictive criteria of future criminal conduct, potentially resulting in a self-fulfilling prediction with racial and socio-economic effect, as described in the sections "Predictive Algorithms" and "Fifth and Fourteenth Amendment" of this publication.

Inappropriate focus occurs when an algorithm's training data are ill-suited to the algorithm's task.[37] This might lead the algorithm to identify factors within a neural network that, though objectively reasonable, are logically irrelevant to the desired outcome. An algorithm that matches faces based on colors, backdrops, or lighting demonstrates inappropriate focus bias.

One can readily imagine how similar bias might migrate into data sets designed to train machine-learning AI to predict terrorism recruitment or threats. To begin with, the data may rely too heavily on international versus domestic actors due to the designer's perceptions or the selection of training data. And because the amount of data may be limited (in contrast to, say, an Amazon or YouTube algorithm), human actors may put too much credence in the reliability of the predictive output. In general, the more data used to train a predictive algorithm, the more accurate the result. An algorithm trained to predict terrorism risk based on a stereotyped "terrorist profile" will, unsurprisingly, be best at locating persons who meet that profile. More persons with the profile will be identified as potential terrorists, and inevitably, more will be found to be engaged in suspicious activities because of increased scrutiny, thus appearing to validate the algorithm and the choice of criteria. Potential terrorists who don't fit the profile may be omitted.

As the example demonstrates, the risk is not just in false positives, the focus of much bias analysis to date, but in the potential failure to identify credible risks: false negatives. Disparities in facial recognition data between males and females could lead to greater inaccuracies in identifying female subjects, increasing the number of false positives—for example, the number of innocent female travelers selected for extra screening or questioning at airports. In contrast, an inability to identify known subjects or threats—for example, a missing or wanted person or Amber Alert kidnap victim on CTV camera feeds—has serious security implications.

[37] *Id.* at 4.

Inappropriate deployment happens when a system is used in a context for which it was not designed, tested, and validated.[38] For instance, a driverless car trained for driving in the United States might not be able to handle left-hand driving in the United Kingdom. A human would adapt to such a change; a driverless car algorithm would need more training. Judges and litigators will want to verify that an AI application is designed for the specific use for which it is relied on in court, one of the lingering issues presented in the *Loomis*[39] case, discussed later.)

Interpretation bias occurs where an algorithm's output is confusing or subject to incorrect interpretation by those working with the technology.[40] Users of facial recognition technology might expect singular, or perfect, matches, in contrast to what most facial recognition algorithms—including the FBI's—actually do, which is to present an array of potential matches, leaving the interpretation and conclusions to human users.

Interpretation bias can also occur because of ambiguity embedded in the algorithmic design, for instance by software designers who, unaware of cultural or linguistic cues, overlook or misuse phrases and concepts, skewing results.

Sometimes the reasoning behind a match is necessary to understand its value or import. To give a real-world example, engineers might design algorithms to search for particular words or phrases, with the goal, for example, of identifying persons engaged in radicalizing internet users. Insufficient knowledge of culture and language, however, could have unintended consequences. Phrases like "the bomb," "knock 'em dead," and "kill it," all mean something to typical American teenagers quite different from what might be intended in a terrorist cell. By the same token, an algorithm designed by an engineer who does not know the import of "the fourteen words" (which form two slogans of white supremacists) may inadvertently enable a potential data threat stream to escape detection.

Unwitting human bias refers to the unintentional infusion into an application of human preferences, stereotypes, values, fears, or knowledge. Consider an algorithm intended to predict risk. An engineer might apply engineering principles to a risk equation. But what is risk? An algorithm will almost certainly incorporate the particular fears, risk tolerances, and perceptions of its designers. (The problem may be compounded when the algorithm is both human and machine generated—a "centaur"—clouding where and how bias might have entered the system.) But the algorithmic equation does not account for human behavior, which is informed not only by the calculation of

[38] *Id.*

[39] *State vs. Loomis*, 881 N.W.2d 749, 759 (Wis. 2016), *cert. denied*, 137 S. Ct. 2290 (2017).

[40] *UNIDIR*, *supra* note 33, at 5.

objective zero-sum costs but also by the emotional impact of fear.

Use of racial, gender, and other social descriptors in algorithms is inherently risky and potentially fraught with ethical and legal issues. One can imagine how both intentional and unintentional human bias might enter the equation as a computer scientist embeds what he or she believes are traits associated with a "race" or ethnicity into facial recognition software. Racial and ethnic categories are inherently ambiguous social constructions covering wide continuums of individuals. Similarly, one can see how nuance might "fool" an algorithm intended to identify age based on the subtle distinctions of faces alone without allowing for the possibility of make-up or efforts at disguise. Bias may also occur unwittingly in machine-learning applications that may not be designed to depend on social identity descriptors but nonetheless rely on such characteristics within the neural network black box. Bias can lead to both the under- and over-inclusion of the targeted group, as "race," ethnicity, gender and other social categories are malleable concepts.

Intentional bias. Scientists, operators, and decision makers may use AI facial recognition tools or predictive algorithms to target disfavored or vulnerable groups. Algorithms can be designed to identify and select certain real and perceived social descriptors associated with "race," gender, sexuality, national origin, religion, disability, and more. Facial recognition technology can identify and track certain ethnic groups as is the case in China with "Uighur characteristics." Clearly pernicious in the profiling of Uighurs (or more accurately, a band of physical characteristics Chinese state security services associate with Uighurs), one question is whether the purposeful use of social identity descriptors is ever an appropriate search parameter. The answer may depend, in part, on purpose, how one defines "search parameter," and the level of human supervision.

On the one hand, there are qualitative differences between the reactive versus predictive uses of social identity descriptors. For example, using an individual suspect or victim description, including descriptors like "race," gender, and age, in response to a credible predicate, in certain contexts, might make sense. Of course, one needs to consider that the initial social indicator that might trigger the use of an AI application may itself be affected by cognitive, societal, or the personal bias of witnesses. Using individually based suspect descriptions is common practice. In the context of using facial recognition to search for a known suspect or a person identified in an Amber Alert, one would not necessarily expect law enforcement to employ race-gender-or-age-neutral input or an algorithm incapable of searching for the specific or reported characteristics of the suspect or victim. However, as already noted, one potential challenge to individually based suspect descriptions is that to the extent social identities are fluid rather than fixed, they may be difficult to code and for officials to implement fairly and accurately.

On the other hand, using a suspect category to identify persons who might engage in an unlawful act on a predictive basis using social identifiers—rather than relying on individualized, behaviorally based reasonable suspicion or probable cause—is an exercise in bias. Law enforcement specialists should eschew such an approach on law enforcement and not just legal and ethical grounds. Among other practical or "policy" reasons, resources are finite. Resources used this way are resources not used to address credible leads elsewhere. Further, individuals who belong to targeted groups may be less likely to share information that leads to credible threats, including threats that originate outside the group. Any terrorist or criminal group seeking to evade detection might use people who do not fit the profiled stereotype, much as, to nobler ends, the Allies used female spies in World War II where the Germans did not expect it.

Courts may need to consider whether an AI application might blur the lines between individual suspect identification and group profiling. An AI application—for example a facial recognition database—might cast a wider net than a traditional human law enforcement investigation, to the point where what started as an individual suspect identification begins to look more like group profiling, opening more people (all of whom or all but one of whom might be innocent) to suspicion and investigation.

Overfitting and outliers

One risk with machine learning is "overfitting," which occurs when the ML model is too tailored to the data it has been trained on and does not account for ambiguities or variations.[41] Generally, this problem is solved by ensuring that the data the ML algorithm is trained on are separate from the data it will encounter in use. A model thus "generalized" should be flexible enough to correctly interpret data it has not encountered.[42] If this data separation is not made, biased results may occur. Consider the following example.

If an ML sentencing algorithm is built on a training set of past offenders, the AI could design its neural network with results custom fit for those specific offenders. If a person reoffends, perhaps with a lesser crime, and his data are used to train the algorithm, there is risk that in calculating a sentence the algorithm might find and match his prior personal data and reproduce the prior sentence; statistically, the sentence will be the best match for his case. In es-

[41] IBM Cloud Education, *Overfitting* (Mar. 3, 2021), https://www.ibm.com/cloud/learn/overfitting.

[42] *See id.*

sence, the algorithm might conclude, within its black box, “For someone with this background, we give a sentence of X.” In effect, the sentencing algorithm has shown a focused bias targeting a specific person. If the individual had not been included in the training set, the algorithm would be forced to seek a more generalized result based on the cases of others and would potentially recommend a different sentence. The point here is that the two sentences are *different,* and the judge relying on the algorithm to inform a decision is unaware that the AI input (and output) is not completely reliable. This problem could apply to any pool of persons who, like reoffenders, might be in training data as well as use data.

To mitigate the risk of overfitting, it is a good idea to ask if the subject of the output prediction was in the algorithm’s training set and, if so, what steps were taken to eliminate bias toward the subject. There are, in fact, algorithms that allow engineers to scrub an individual’s data from an ML algorithm, essentially making it forget the person.[43] Courts or legislatures might also require that an algorithm for assessing risk not include in its training data any individuals to whom the application might be applied, or they might decline to use an algorithmic tool in a sentencing or similar context.

Risk in the other direction might produce an “outlier,” which occurs where the input is sufficiently distinct from the scenarios built into training sets to confuse the algorithm. The situation is analogous to sentencing for a crime that is not included in the Sentencing Guidelines and is not readily analogous to an existing offense. An algorithm designed to produce a result regardless of accuracy might attempt to force the case into an incorrect box. Outlier inputs could lead to unpredictable, biased, and incorrect results. New crimes that do not fit the algorithmic model for assessing bail or recidivism risk, or for which there is an exceedingly small data set, could produce a similar result. Consider an algorithm factoring a foreign agent registration violation (22 U.S.C. §§ 611–621), as opposed to, say, a crime like rape for which there are thousands of data points. The algorithm might match and weight the term “foreign agent” and equate the offense with espionage or even treason, not registration or ministerial failure. Depending on which terms (or factors) were weighted and how, wildly varying recommendations for significant or minimal confinement could result. Were the algorithm acting within the black box of input-output, the judge would not know why it predicted that the defendant was or was not a bail or recidivism risk or should be sentenced to so many years of confinement.

[43] Mathew Hutson, *Researchers Can Make AI Forget You*, IEEE Spectrum (Jan. 15, 2020), https://spectrum.ieee.org/tech-talk/computing/software/researchers-can-make-ai-forget-you.

Judges might decline to use an algorithmic tool, or they might attempt to mitigate the risk of outliers, by asking: Was the algorithm specifically trained for the offense or case in question, and if so, with what volume of data? Is the defendant in question an outlier, and if so, in what ways? Has the algorithm been designed to account for those possibilities? If the answer to these questions is no, then there is heightened risk the algorithm will not predict with the accuracy intended or advertised. For sentencing, the threshold for throwing out algorithmic results could reasonably be low, as the alternative—human decision making—is already the standard. In any event, it would seem incumbent on the proponents of using such an algorithm to demonstrate its validity and explain its functioning, just as a judge should (and in some jurisdictions is required to) put reasoning for a sentence on the record, allowing appellate courts and the parties to understand what occurred and why.

Mitigating bias

With AI as with people, some bias is always present. But steps can be taken to minimize the risk. One mitigator is sound process—timely, contextual, and meaningful. For policymakers and engineers, "timely" means at points where input can directly influence outcomes, i.e., at the conception, design, testing, deployment, and maintenance phases of AI development and use. "Contextual" means specific to the tool and use in question and with actual knowledge of its purposes, capabilities, and weaknesses. "Meaningful" means independent, impartial, and accountable. Specifically, the person using or designing an application should validate its ethical design and use. If a particular community or group of people is likely to be affected by the use of the tool, designers and policymakers should consult with that community or group in deciding whether and how to develop, design, or use it.[44] In addition, to the extent feasible, the system's parameters should be known, or retrievable. The system should be subject to a process of ongoing review and adjustment. The rules regarding the permissible use, if any, of social identifying descriptors or proxies should also be enunciated, clear, transparent, and subject to constitutional and ethical review. For judges and litigators, sound process also means the careful application of the Rules of Evidence to AI-generated evidence and tools on the record.

[44] Jamie Baker et al., *National Security Law and the Coming AI Revolution, Observations from a Symposium Hosted by Syracuse University Institute for Security Policy and Law Georgetown Center for Security and Emerging Technology, Oct. 29, 2020* (2021), https://cset.georgetown.edu/wp-content/uploads/Symposium-Report-National-Security-Law-and-the-Coming-AI-Revolution.pdf.

Probing for bias

Asking the right questions is crucial, not just for legal reasons but because the questions invariably underpin judgments about the reliability of the AI at issue. Here are some questions to ask:

- Who designed the algorithm at issue and subject to what process of review?
- Were stakeholders—groups likely to be affected by the AI application—consulted in its conception, design, development, operation, and maintenance?
- What is in the underlying training, validation, and testing data? How has the data chosen been cleaned, altered, or assessed for bias? How have the data points been evaluated for relevancy to the task at hand? Is the data temporally relevant or stale? Are certain groups improperly over- or under-represented? How might definitions of the data points used impact the algorithm analysis?
- Is the model the state of the art? How does it compare against any industry standard evaluation metrics or application specific benchmarks?
- How might the terms or phrasings in the user-generated prompts bias the systems' outputs? Can these prompts be phrased in a more neutral way? Do any of the terms used have alternative meanings?
- Are the algorithm's selection criteria known? Iterative? Retrievable in a transparent form? If not, why not?
- Does the application rely on a neural network? If so, are the parameters and weights utilized within the neural network known or retrievable? Does the design allow for emerging methodologies that provide for such transparency? If so, why haven't they been used? If not, what is the risk that the system will rely on parameters that are unintended or unknown to the designers or operators? How high is the risk? Is the risk demonstrated? How is the risk mitigated?
- Is the input query or prompt asking for a judgment, a fact, or a prediction? Is the judgment, fact, or prediction subject to ambiguity in response?
- Do the criteria include real or perceived racial, ethnic, gender, or other sensitive categories of social identity descriptors, or any proxies for those categories? If so, why, and do they pass ethical and constitutional review? Have engineers and lawyers reviewed the way these criteria are weighted in and by the algorithm as part of the design and on an ongoing basis? In accord with what process of validation and review?

- Is there a disparate or adverse impact on the confidence threshold based on racial classifications, ethnicity, gender, sexuality, ability/disability, nationality, and so on? If so, are there logical and objective reasons for such disparity that survive constitutional and ethical review?
- Are there situational factors or facts in play that could, or should, alter the algorithm's predictive accuracy?
- Is the application one in which nuance and cultural knowledge are essential in order to determine its accuracy or to properly query it?
- Are the search terms and equations objective or ambiguous? Can they be more precise and more objective? If not, why?
- What is the application's false positive rate? What is the false negative rate?
- What information corroborates or disputes the determination reached by the AI application? Is the application designed to allow for real-time assessment? If not, is operational necessity the reason, or is it simply a matter of design? Is there a process for such assessment that occurs after the fact?
- Is the AI being used for the purpose for which it was designed and trained?
- Is the AI being used to inform or corroborate a human decision? Are humans relying on the AI to decide or to inform and augment human decision?

5. Predictive Algorithms

Most algorithms are based on statistical prediction. In this sense, all algorithms are *predictive*. There exists a class of algorithms, however, that also seek to *make predictions about future behavior based on past data*. This happens all the time. Shopping algorithms seek to predict through data about prior purchases (past behavior) the predisposition of individuals to make additional purchases (future behavior). YouTube, which uploads 500 million hours of video a day, uses algorithms that seek to predict additional videos a viewer might watch to generate additional views and increased ratings—and thus revenue. They are called "recommendation algorithms," but what they do is push products to viewers based on predictions about their future viewing behavior. And in the case of YouTube, the algorithms are widely understood to be designed to increase viewer addiction by increasing the depth of what it is the algorithm is predicting the viewer wants, such as violence or comedy.

For judges, the question is not only whether predicting behavior is inherently good or bad but whether algorithms that seek to do so are accurate, and whether there are uses of predictive algorithms that may present issues for courts, as when, for example, a predictive algorithm embeds certain types of bias. Likewise, issues might occur because litigants are unwilling or unable to determine the parameters or data sets that informed an algorithm's prediction and thus cannot reliably evaluate its accuracy as applied to the circumstances at hand. As a result, judges need to identify the benefits and risks of relying on such algorithms.

Predictive algorithms are used, or might be used, in a variety of judicial and collateral settings. The most frequently cited applications, and potential applications, are algorithms predicting pretrial flight risk to help determine whether and at what amount to set bail, as well as those purporting to calculate the risk of recidivism to inform decisions about parole. Risk assessment tools are also used in sentencing. Other ways in which algorithms may impact judicial decisions include predictive policing and identifying "at-risk" youth. Predictive policing algorithms, for example, look at past data about the time, location, and nature of arrests to predict when and where future crimes may occur, so that patrol presence in those areas can be increased to deter or address crime. Policing algorithms are often equated with the "broken window" theory of policing. Policing algorithms are not intended to predict individual conduct, though they might include the characteristics of individual actors in an area, like registered sex offenders. Proponents of such algorithms argue that algorithmic tools better focus finite police resources on areas where crime is most likely to occur, based on “neutral” data rather than the hunches, perceptions, or potential biases of police officers. The argument against them is at least twofold. First, such algorithms can generate their own reinforcing and circular logic. The algorithm predicts criminal conduct, police patrols are increased, and additional arrests occur, validating the accuracy of the algorithm. Second, the underlying data may not, in fact, be neutral. Such algorithms may have a disproportionate racial and socioeconomic impact where they generate increased patrols in poorer neighborhoods with historically higher recorded crime rates and larger concentrations of minorities. In this way, they may also reflect past police practices and prosecutorial decisions focusing on communities and people of color. They may also have intentional racial impact to the extent they use “race” or socioeconomic status as predictive factors.

“At-risk” youth. In the United Kingdom, some local governments use algorithms to identify at-risk youth for the purpose of social intervention before more drastic law-enforcement remedies are triggered. These algorithms identify and weigh risk based on data from police reports, such as parental involvement in domestic incidents, social benefits and other government records, and school attendance records, among other sources.[45] Proponents of their use argue that they pick up data-based cues faster and more comprehensively than human actors—social workers and school counselors—who would not discern the same data patterns and do not have the time or data access, in any event, to find the same connections. Advocates claim that identifying vulnerable youth early allows for timelier and more beneficial social, rather than criminal, intervention.

Opponents argue that aggregate data collection amounts to an invasion of the privacy of a fragile class of citizens: children from lower economic strata. Further, opponents believe the algorithms and corresponding interventions unfairly stigmatize youth by speculating about what *might* happen based on statistics rather than individual characteristics. There is also concern that socioeconomic parameters can embed latent and unintended racial and economic bias. While all sides would acknowledge that early intervention with at-risk youth is beneficial, they would not necessarily agree that the factors relied on by algorithms are the most relevant to identifying the young people who would benefit from earlier intervention.

As these examples indicate, several generalized arguments for and against the use of predictive algorithms emerge. Proponents might argue:

[45] Cade Metz and Adam Satariano, *An Algorithm That Grants Freedom, or Takes It Away: Across the United States and Europe, Software Is Making Probation Decisions and Predicting When Teens Will Commit Crime. Opponents want more human oversight.* N.Y. Times (Feb. 7, 2020), https://www.nytimes.com/2020/02/06/technology/predictive-algorithms-crime.html.

- Predictive algorithms can identify patterns and trends humans cannot see, thus curtailing additional risk or harm.
- Predictive AI rests on the premise that neither judges nor law enforcement personnel can reasonably predict conduct based on judgment and intuition alone. AI simply has more data and excels at statistics.
- In the courtroom, predictive AI could add data to human judgments about risk assessment, informing decisions on bail, parole, and sentencing. Moreover, because AI is data driven, some argue that a well-designed algorithm could in theory be more neutral or objective than a human. While AI invariably contains bias, a particular application might, in theory, be less biased than a human subject to implicit or express bias.

All of these assumptions can be contested, in the abstract as well as with reference to specific AI applications, which is why courts should hear arguments from both sides where predictive algorithms, especially those driven by AI, are concerned.

Opponents of predictive algorithm use meanwhile make the following arguments:

- Western law and criminal procedure are premised on individualized suspicion. This means an individual should be investigated or prosecuted based on articulable facts about them, not patterns found in data about the past conduct of other persons who may simply share one or more social descriptors, or data about past police practices and prosecutorial decisions.
- All algorithms are biased in some way by the choices their human designers make: what metrics are used to evaluate data to make predictions, what data the algorithm is trained on, and what data it is tested on. Further, algorithms reflect human bias and can multiply and magnify bias by repeating it at scale. One of the most common criticisms of criminal risk assessment tools, for example, is that they rely on historical records of arrests, charges, convictions, and sentences, though "[d]ecades of research have shown that, for the same conduct, African-American and Latinx people are more likely to be arrested, prosecuted, convicted, and sentenced to harsher punishments than their white counterparts."[46]
- Predictive algorithms focus on characteristics that are, at least purportedly, readily discerned and susceptible to data adaptation and

[46] *Technical Flaws of Pretrial Risk Assessments Raise Grave Concerns*, Berkman Klein Cter for Internet & Soc'y at Harv. Univ. (July 17, 2019), https://cyber.harvard.edu/story/2019-07/technical-flaws-pretrial-risk-assessments-raise-grave-concerns.

recording. Classifications such as "race," gender, marital status, family status, address, and education likely play a disproportionate role in algorithm design and operation. Conversely, in operation or design, algorithms are less likely to include subjective weights like role models and community connections and participation that might also predict behavior and perhaps do so more accurately.

- Classifications used as factors in algorithmic predictions are subject to all the risk of bias, intended and unintended. Even when unintentional, this bias may infiltrate an application through training data, how computer engineers assign weights to factors, or "learning" the AI does on the job.
- Some factors do not account for variation or nuance. Factors that appear to be subject to yes/no answers, and thus data scoring, may in reality be more complex and fall along a continuum. "Race" and ethnicity are good examples, even if not used intentionally in predictive tools, other than to counter historical or algorithmic bias. Marital status, for example, a seemingly objective data point, may fall on a contextual continuum ranging from stable to unstable, happy to unhappy. Depending on what an algorithm is intended to predict, nuance can make all the difference in outcomes.

All these factors are compounded where there is an inability to understand or challenge the underlying algorithm. Judges will have to determine when algorithm transparency is required as a matter of law, including due process. Lack of transparency undermines the ability of judges and litigators to assess the accuracy and meaning of an algorithmic output by asking questions like: What factors did the algorithm rely on? How were they weighted? Do those factors in fact reflect the case and parties in question?

When considering whether AI outputs should be admitted as evidence or used to inform judicial decisions, judges should do the following:

- Require corroboration before relying on an algorithm to inform a decision. Judges might consider whether the algorithm's statistical prediction aligns with their own understanding of the facts. If so, how so? And if not, why not?
- Give more (if any) deference to algorithms that are transparent in their (1) function; (2) underlying training, validation, and testing data; (3) weighting factors; and (4) methodology of weighting. Where such factors are not discernible or understandable, ask why and if better technology is available; and if a determination is made to use the algorithm anyway, state why on the record.

- Consciously and purposefully distinguish between data that are generated based on group characteristics and data that are specific to the individual in question.
- Insist that any AI utilized by the court include a mechanism to evaluate its accuracy on an ongoing basis, specifically one to identify false positive and false negative rates, along with the trends associated with each.
- Determine whether the AI application incorporates biased data inputs or design, or creates biased outputs, as discussed in chapter 4.
- Know when "race," gender, or other suspect class factors—or inputs that may function as proxies for those factors, such as housing and employment status[47]—are incorporated into algorithmic designs, and determine on the record why those factors are relevant to the purpose and function of the AI use in question. (Judges of course must also evaluate whether the use of such factors passes constitutional and ethical review.)
- Where AI is used to make judicial decisions, or not used but available, consciously determine whether that choice should be determined by legislative direction or judicial discretion.
- Clearly state on the record when, how, and to what extent an algorithm informed a decision. Appellate courts give trial judges greater deference when evidentiary rulings are made on the record and explained. One question appellate judges will need to address is when and whether to give such deference where AI is concerned. To what extent, for example, should or must "on the record" include exploration of the underlying AI elements—design, data, algorithm, bias—in addition to a clear statement as to why and with what legal analysis AI evidence has been admitted into evidence or used to inform a judicial decision?

[47] Chelsea Barabas, et al. *An Open Letter to the Members of the Massachusetts Legislature Regarding the Adoption of Actuarial Risk Assessment Tools in the Criminal Justice System*, Berkman Klein Ctr. for Internet & Soc'y 3 (Nov. 9, 2017), http://nrs.harvard.edu/urn-3:HUL.InstRepos:34372582.

6. Deepfakes

AI's capacity to convert symbolic language (coded numbers) into natural language and to discern, recognize, and formulate patterns at the pixel level makes it a tool of choice not only for identifying voices and pictures but also for mimicking voices and altering images. Moreover, AI can do so with real-life precision, creating images or recordings known as "deepfakes." Hollywood has, of course, known about deepfakes for years, though in movies they're called "special effects," as in *Star Wars* or *Forrest Gump*. What makes deepfakes noteworthy for courts is not only the lifelike quality attainable but the accessibility of this capability to the general population. Tools readily available on the internet allow nonspecialists to alter photographs and mimic speech with startling realism, capable of fooling practically everyone—including triers of fact. Luckily, there are also tools and methods to authenticate images like digital IDs and cryptographic hashes. The question is when courts should require such authentication before admitting images or voices into evidence.

As is often the case with image technology, the deepfake found one of its first manifestations in pornography and pornographic revenge, with digital editors grafting one person's face onto another's body. In contrast to some areas of AI, some state legislatures were relatively quick to consider regulating certain deepfake pornography through criminal sanction.[48] Thus, state courts, but also federal courts in the context of the Consumer Privacy Protection Act, will likely confront increasing use of AI to generate fantasy porn, revenge porn, and child porn. The questions for courts will include: Is the particular deepfake porn criminal? Or does it fall under some rubric of First Amendment protection?[49]

The same capabilities enabling creation of lifelike pornography can already be used to convincingly generate or alter evidence. Judges, in their capacity as evidentiary gatekeepers, can expect to engage in new areas of inquiry and debate involving authentication.

[48] Matthew F. Ferraro, *Deepfake Legislation: A Nationwide Survey*, WilmerHale (2019), https://www.wilmerhale.com/en/insights/client-alerts/20190925-deepfake-legislation-a-nationwide-survey.

[49] For a helpful overview of deep fake issues, see Danielle K. Citron & Robert Chesney, *Deep Fakes: A Looming Challenge for Privacy, Democracy, and National Security*, 107 Cal. L. Rev. 1753 (2019), https://scholarship.law.bu.edu/faculty_scholarship/640.

7. Judges as AI Gatekeepers

As evidentiary gatekeepers, judges will need to determine whether and when AI evidence will assist the fact finder and is admissible in court. The Federal Rules of Evidence and their state equivalents will help guide this determination. The Supreme Court's *Daubert,*[50] *Crawford,*[51] and *Carpenter*[52] cases may also inform the evidentiary questions presented by AI. Neither these cases nor the Rules, however, were written with AI in mind. And currently few federal or state cases or jury instructions address AI. The following discussion is intended to help judges spot AI-specific issues beyond the ordinary evidentiary questions that judges address. Judges will, of course, interpret and apply these cases and rules to AI in the specific contexts presented and do so consistent with the law of the jurisdiction in which they practice.

Federal Rules of Evidence 401–403, 702, 902(13) and (14)

As judges well know, under Federal Rule of Evidence 401, evidence is relevant if "(a) it has any tendency to make a fact more or less probable than it would be without the evidence; and (b) the fact is of consequence in determining the action."[53] Rule 402 states that relevant evidence is admissible unless the Constitution, a federal statute, the other Federal Rules of Evidence, or other rules prescribed by the Supreme Court apply and would exclude the evidence.[54] Due process or confrontation clause concerns, for example, might bar or limit certain AI evidence from admission. Statutes addressing data privacy and use may do so as well. Rule 403 allows a court to exclude relevant evidence if its probative value is substantially outweighed by a danger of creating unfair prejudice, confusing the issues, misleading the jury, causing undue delay, wasting time, or needlessly presenting cumulative evidence.[55]

Many of the threshold evidentiary issues associated with AI will be litigated under Rules 402 and 403 or their state equivalents. Relevancy in most or all jurisdictions is broadly defined, and most AI applications are essentially tools for assessing probability, in theory, making them inherently relevant in assessing whether something is "more or less probable." The primary issues, then, are (1) the reliability of AI generally and (2) the appropriateness of use in the context presented. Rules 402 and 403 are pertinent because the evi-

[50] Daubert v. Merrell Dow Pharms., Inc., 509 U.S. 579 (1993).

[51] Crawford v. Washington, 541 U.S. 36 (2004).

[52] Carpenter v. United States, 138 S. Ct. 2206 (2018).

[53] Fed. R. Evid. 401.

[54] Fed. R. Evid. 402.

[55] Fed. R. Evid. 403.

dentiary use of AI will invariably present questions about discovery and due process, such as whether there is a right to access an underlying algorithm or data used to generate evidence or inform judicial decisions. Another issue is the risk that litigation over AI will present the figurative "trial within a trial" and potentially confuse the jury under Rule 403. Also, courts might apply Rule 403 to exclude AI evidence that is biased or otherwise unreliable. Inquiry is prudent; otherwise, juries may assume AI evidence has the imprimatur of "science" or "technology" in the context presented, potentially lending it false authority or undue weight, or permitting its use in a manner for which it was not intended.[56]

Judges will need to decide in what manner and to what extent to require authentication of the AI evidence offered and how, if at all, to validate its reliability. These criteria will bring Federal Rules of Evidence 702 and 902 into play, as well as *Daubert* and *Crawford*.

Rule 702 governs the admissibility of expert witness testimony. It provides:

> A witness who is qualified as an expert by knowledge, skill, experience, training, or education may testify in the form of an opinion or otherwise if:
>
> (a) the expert's scientific, technical, or other specialized knowledge will help the trier of fact to understand the evidence or to determine a fact in issue;
>
> (b) the testimony is based on sufficient facts or data;
>
> (c) the testimony is the product of reliable principles and methods; and
>
> (d) the expert has reliably applied the principles and methods to the facts of the case.

Rule 902 covers self-authenticating evidence, such as official records and newspapers. In 2017, subparagraphs (13) and (14) were added to Rule 902 to address, among other things, the admission of digital evidence and machine-generated records, which in theory now are self-authenticating:

> (13) *Certified Records Generated by an Electronic Process or System.* A record generated by an electronic process or system that produces

[56] Andrea Roth, *Machine Testimony*, 126 Yale L.J. 1972, 2043 (2017) ("Moreover, just as the Framers were concerned that factfinders would be unduly impressed by affidavits' trappings of formality, 'computer[s] can package data in a very enticing manner.' The socially constructed authority of instruments, bordering on fetishism at various points in history, should raise the same concerns raised about affidavits.")(internal citations omitted).

> an accurate result, as shown by a certification of a qualified person that complies with the certification requirements of Rule 902(11) or (12). The proponent must also meet the notice requirements of Rule 902(11).
>
> (14) *Certified Data Copied from an Electronic Device, Storage Medium, or File.* Data copied from an electronic device, storage medium, or file, if authenticated by a process of digital identification, as shown by a certification of a qualified person that complies with the certification requirements of Rule 902(11) or (12). The proponent also must meet the notice requirements of Rule 902(11).

These rules cover digital photographs and other digital documents as well as data "generated by an electronic process or system." In other words, the language of these paragraphs might cover AI-generated outputs and data, potentially applying to evidence as diverse as the output from an AI-driven radiological machine or the results of a hiring algorithm for sorting job applicants. Judges will need to decide whether certain outputs, such as an imagery assessment of a medical picture or the program history of a driverless vehicle, qualify for admission under FRE 902(13). If they do qualify, the next issues are whether the AI application produces accurate results and who the "qualified person" is to make that certification. On this latter question alone, there are many options including: the software engineer, design engineer, data engineer, or company CEO. (Courts will need to determine whether a "custodian of the records," generally, or in each instance, is in fact competent to authenticate evidence derived from AI.)

Artificial Intelligence and the interpretation of AI outputs is complex. Courts will have to determine the appropriate means to verify AI outputs. This might involve expert testimony, or it might be done through technical means, such as cryptographic hashes embedded in an image at the time it is created. Courts will need to determine who is qualified to testify about the accuracy and fairness of an AI application. Of course, steady, purposeful, and consistent application of the Federal Rules of Evidence or their state equivalents on the record is a good place to start.

Crawford, *Daubert*, and *Frye*

In 2004, the Supreme Court held in *Crawford v. Washington* that in certain contexts documentary evidence should no longer be considered a business record when used as criminal evidence at trial but rather as testimony for the purpose of triggering the Sixth Amendment right of cross-examination.[57] The

[57] Crawford v. Washington, 541 U.S. 36 (2004).

Court left it to lower courts to determine when machine-generated evidence should be treated as "testimonial" rather than as a business or other written record. As we will see, algorithm-generated outputs used in court offer ample ground to continue this debate, not only regarding whether the output is testimonial but, if so, who should be cross-examined.

Daubert v. Merrell Dow Pharmaceuticals, Inc. (1993) and, in certain states, its predecessor, *Frye v. United States* (1923), govern the admission of expert testimony based on scientific methodology. *Daubert* uses a "factors" approach, while *Frye* uses a "general acceptance" standard. Each of the *Daubert* factors opens wide the door to debate over many AI attributes. *Frye* is likely more complicated, asking judges to determine when a scientific method is "sufficiently established to have gained general acceptance in the particular field to which it belongs."[58] In theory, making such a determination will entail examining not only the specific algorithm and use in question but also identifying the relevant field of acceptance and what acceptance means for something like facial recognition or behavior prediction—all in a context where algorithms are iterative and changing.

It is intuitive, but worth remembering that proponents of AI-generated evidence will seek to simplify its admission by limiting or eliminating as many threshold foundational requirements as possible. Opponents of admission will seek to undermine its relevance and reliability in general or for the purpose for which it is offered. To challenge relevance and accuracy opponents will seek access to the underlying algorithm, the data on which it was trained, and knowledge of what occurs and what is weighted inside any machine-learning black box. Thus, courts will face a layered adjudicative challenge each time AI-generated evidence is offered.

Where AI outputs are admitted, opponents will seek to cross-examine the software engineers responsible for its design. Each AI application is different. It will have a different purpose, rely on a different algorithm, use a different machine learning methodology or methodologies, and will train, test, and validate using different data. Consequently, AI issues are generally not subject to resolution through the application of case law precedent in the same way that, for example, DNA analysis is now widely accepted in court. Adjudication is to be expected for each application and in each context for which the application is offered as evidence. As noted at the outset, AI is a constellation of technologies and applications, not a single process or technology that can be validated once and generally adopted. In each instance where AI evidence is offered, there may be a legitimate need to explore the underlying technology, and different components of that technology, for use in that instance or for the proffered purpose.

[58] Frye v. United States, 293 F. 1013, 1014 (D.C. Cir. 1923).

We review a few of the potential adjudicative issues below to help judges and litigators develop context-relevant questions as well as realize the importance of probing beyond confidence thresholds and false positive rates before using AI applications or admitting AI evidence to inform legal judgments.

Salient issues

Context. Courts should pay attention to whether a particular AI application is a good "fit"[59] for the purpose for which it is proffered. Some criminal risk assessments, for example, are designed for the purpose of determining which individuals might benefit from alternatives to incarceration, such as parole or counseling. These algorithms might have less relevance and reliability when used to determine sentencing.[60] That will depend on all the factors noted above, including the input factors, the weight assigned to those factors, the data on which the algorithm was trained, and the nature of the confidence thresholds applied to the output. Courts should pause and ask not only whether the AI at issue is relevant and material to the matter before the court but for what purpose the AI was specifically designed and whether the outputs will materially and fairly inform the fact finder.

Case-specific reliability. Even when an algorithm is being used for the purpose for which it was designed, there may be data or design reasons why output reliability will decrease in a specific context. An AI algorithm may have been designed for and tested on a population substantially different from the population for which the output is offered, with less accurate results than the lab-tested confidence threshold[61] ("inappropriate deployment" as discussed in the bias section). The Government Accountability Office (GAO) stated that the FBI's facial recognition application has an 86% percent match rate (confidence threshold) when an input image is compared to at least fifty potential output matches drawn from state license data bases. However, the same algorithm would not have the same match accuracy if run against a different input demographic—say, the population of another country—not because the algorithm is necessarily intentionally biased but because it has not been trained against a comparative population pool. In fact, output disparity across gender

[59] *See* Daubert v. Merrell Dow Parms., Inc., 509 U.S. 579, 591–92 (1993).

[60] *See* Christopher Bavitz et al., *Assessing the Assessments: Lessons from Early State Experiences in the Procurement and Implementation of Risk Assessment Tools.* Berkman Klein Center for Internet & Soc. research publication, 6–7 (Nov. 2018), http://nrs.harvard.edu/urn-3:HUL.InstRepos:37883502 (discussing the Wisconsin Supreme Court's warning in *State v. Loomis*, 881 N.W.2d 729 (Wis. 2016) that the risk assessment tool COMPAS was not developed for use at sentencing).

[61] *See* Barabas et al., *supra* note 47, at 3, and Bavitz et al., *supra* note 60, at 7.

and ethnicity has been an issue with some facial recognition algorithms.[62] As a result, facial recognition accuracy has been a focal point of AI design initiatives, and we anticipate future American facial recognition applications will largely address this issue.

Inapt factors. There is a risk with ML that a neural network will rely on inapt factors in making its output predictions. Judges will want to know whether this is possible and, if so, regarding which factors, before allowing a jury to assess the weight of AI evidence or before using an algorithm themselves to assess bail or recidivism risk. For example, a judge would want to determine, consistent with case law and the Constitution, which factors were included and weighted within any AI-driven bail, parole, confinement, or sentencing tool, to ensure that inappropriate, inapt, or unconstitutional factors were not included and, if factors were appropriate, not given undue weight by the neural network. A judge would also want to know if any factors might be working as proxies[63] for suspect categories.

Bias. Courts will want to investigate the ways in which a given AI application is biased before admitting its outputs into evidence or relying on it to inform a judicial decision. (Refer to the discussion of "Bias" earlier.)

Crawford. The Sixth Amendment provides that "[i]n all criminal prosecutions, the accused shall enjoy the right … to be confronted with the witnesses against him … ." This right is understood to encompass the right to cross-examine witnesses at trial. An algorithm is not "a witness," but in *Crawford*, the Supreme Court held that the right to cross-examine witnesses extends, in some cases, to certain out-of-court "statements" introduced at trial, including statements to the police (as was the case in *Crawford*) as well as "statements that were made under circumstances which would lead an objective witness reasonably to believe that the statement would be available for use at a later trial."[64] Significantly, the Court subsequently held that certain lab reports were testimonial and thus the technician or scientist who compiled the report was subject to examination. Before *Crawford*, many of these statements were admitted into evidence as business records or under generally recognized exceptions to the hearsay rules. In the absence of clarifying guidance from the Supreme Court, lower courts have struggled to apply *Crawford* to documentary data and other information later introduced as criminal evidence, like lab reports and photographs. In short, *Crawford* is applied inconsistently and on a case-by-case basis.

[62] *NIST Study Evaluates Effects of Race, Age, Sex on Face Recognition Software: Demographics Study on Face Recognition Algorithms Could Help Improve Future Tools* (Dec. 19, 2019), https://www.nist.gov/news-events/news/2019/12/nist-study-evaluates-effects-race-age-sex-face-recognition-software.

[63] Barabas et al., *supra* note 47.

[64] Crawford v. Washington, 541 U.S. 36, 52 (2004).

AI-generated information later used as evidence is fertile ground for a *Crawford* challenge, including litigation over just who or what is "bearing witness." Where AI data are used as evidence in a criminal trial against an accused, the defendant may seek to assert a Sixth Amendment right to question the author of the algorithm. The designer of the software, the data selector, and the author of the learning algorithm are all candidates for cross-examination. Whether *Crawford* is applicable or not, some scholars and practitioners argue that litigants should be able to impeach machines at trial, just as they would human witnesses.[65] The argument is rooted in the Sixth Amendment to be sure, but more generally it arises from uncertainty about the accuracy of AI-driven machines. One scholar argues that judges—and if not judges, legislators—"should allow the impeachment of machines by inconsistency and incapacity, as well as by evidence of bias or bad character in human progenitors."[66] Whether required by *Crawford* or not, legislators and judicial rule-making bodies might require live testimony "for human designers, inputters, or operators in certain cases where testimony is necessary to scrutinize the accuracy of inputs."[67] Of course, judges might already allow such a process through the application of the existing Rules of Evidence, as well as due process. The public policy question is whether the law or the Rules should require such inquiry, or whether inquiry should be left to the discretion of individual judges to determine.[68]

Daubert. One way to conceptualize AI evidence is to apply the (nonexhaustive) list of factors the Supreme Court developed in *Daubert*[69] to determine whether expert testimony based on a specific, scientific methodology should be admitted. These include[70]

- whether the theory or technique in question can be and has been tested
- whether it has been subjected to peer review and publication, its known or potential error rate
- the existence and maintenance of standards controlling its operation
- whether it has attracted widespread acceptance within a relevant scientific community

With AI, these factors would need to be applied to individual algorithms

[65] Roth, *supra* note 56.

[66] *Id.*

[67] *Id.*

[68] *See* id., at 2031 (discussing other nations' choices).

[69] Daubert v. Merrell Dow Pharms., Inc., 509 U.S. 579 (1993).

[70] *Id.* at 593–95.

and applications rather than "AI" generally, which term generically describes a constellation of technologies and methodologies.

Testing. The first step suggested by *Daubert* is to identify the theory, technique, or component that is subject to evaluation. There are many options with AI. Is it: The sensor(s) that fed data to the AI system? The algorithm? The math behind the algorithm? The data set used to train the algorithm? The training methodology? Or is it the system as an integrated whole that is subject to review?

The second step is to decide what test is appropriate and what baseline to use to establish accuracy. Medical diagnostic AI, for example, might be compared to physician-diagnosed outcomes. It is true that medical diagnostics are subject to social influence and human and machine bias. But in medicine there is often a fixed data point, an established fact or yes-no answer to whether a disease or tumor is present, against which testers can measure the algorithm's accuracy.

In contrast, an algorithm intended to predict future behavior, such as a criminal assessment tool, cannot be tested with the same degree of scientific or evidence-based meaning, given the weight placed on social factors. Recidivism algorithms attempt to predict future human behavior, using circumstantial factors drawn from a base population. In such contexts, there is no certain result and no control group, and confirming predictions is difficult. Human circumstances are endlessly complex, creating multiple influences on behavior—without necessarily *determining* behavior. Nor is there a way to verify, after an individual has been jailed or sentenced, how an individual's future behavior is affected by imprisonment. The experience of imprisonment itself might turn a person toward or away from future crime, making it difficult or impossible to verify the machine's prediction. In short, predictive algorithms in the criminal context are especially difficult to test, to peer review, and to assess for accuracy and error rates.

Peer review. A recent innovation in AI-enabled medicine highlights the question of machine reliability and illustrates the importance of peer review. In April 2019, NPR reported that Stanford computer scientists had created an algorithm for reading chest X-rays to diagnose tuberculosis.[71] They hoped to use it to diagnose the disease in HIV patients in South Africa, and the machine's results were already better than doctors'.[72] To corroborate their success,

[71] Richard Harris, *How Can Doctors Be Sure a Self-Taught Computer Is Making the Right Diagnosis?,* NPR (Apr. 1, 2019), https://www.npr.org/sections/health-shots/2019/04/01/708085617/how-can-doctors-be-sure-a-self-taught-computer-is-making-the-right-diagnosis.

[72] *Id.*

the Stanford scientists submitted their results to other scientists for review.[73] One noticed a peculiarity in the AI's decision making.

> [The peer reviewers] Zech and his medical school colleagues discovered that the Stanford algorithm to diagnose disease from X-rays sometimes "cheated." Instead of just scoring the image for medically important details, it considered other elements of the scan, including information from around the edge of the image that showed the type of machine that took the X-ray. When the algorithm noticed that a portable X-ray machine had been used, it boosted its score toward a finding of TB.
>
> Zech realized that portable X-ray machines used in hospital rooms were much more likely to find pneumonia compared with those used in doctors' offices. That's hardly surprising, considering that pneumonia is more common among hospitalized people than among people who are able to visit their doctor's office.
>
> "It was being a good machine-learning model and it was aggressively using all available information baked into the image to make its recommendations," Zech says. But that shortcut wasn't actually identifying signs of lung disease, as its inventors intended.[74]

The machine was making a correlational, rather than causal, connection between the use of a portable machine and TB. Without informal peer review, humans might not have discovered that aspect of how the AI algorithm was making decisions, a clear example of both how AI adapts and the fact that it often does so in the black box. The TB-scan example also demonstrates the disruptive role of unknowns, here an unwitting, algorithmic bias ("inappropriate focus"). The original programmers evidently did not anticipate that the machine would teach itself to evaluate information beyond the scan itself. It is impossible for a programmer to anticipate every real-world factor a machine will encounter and attempt to interpret.

Error rates. Judges will also need to ask the right questions to determine whether error rates are accurate and meaningful. For example, will, or might, error rates vary depending on whether the AI application is tested and reviewed using the relevant local population (database) to which it will be applied, as opposed to a national population, or perhaps a more idealized lab database?[75] What types of bias might be affecting the accuracy of any reported error rates? (See suggested questions under "Probing for Bias.")

Standards controlling an AI application's operation and maintenance. AI im-

[73] *Id.*

[74] *Id.*

[75] *See* Barabas et al., *supra* note 47, at 3, and Bavitz et al, *supra* note 60, at 7.

poses operational and maintenance obligations. At this point in time, however, operational standards, if any, are set voluntarily. The intelligence community and the Department of Defense have each published principles for the ethical use of AI, while many companies have their own internal standards. In the absence of uniform statutory standards, courts might begin by asking: What dataset is used? Is that dataset updated appropriately? Is the machine learning monitored by continued testing against known results to ensure the machine is not learning bad habits? Courts might also ask all the questions about bias suggested in "Probing for Bias," page 40.

Acceptance. Courts will also need to determine what widespread acceptance within the relevant scientific community means in the context of AI. There is a big difference between general acceptance of the field and acceptance of a specific application. Many computer engineers and government actors accept the premise and use of facial recognition, but privacy advocates do not. Skepticism will remain with any specific application. The point is also illustrated by driverless cars. General acceptance of the concept has not at present translated to acceptance of a model of autonomous driverless car that is ready for commercial sale and public use. What then would constitute appropriate general acceptance?

Proprietary algorithms. How does one test the accuracy or conduct a peer review of a proprietary algorithm or an iterative or evolving ML algorithm? Google is not likely to disclose its search algorithm for public or peer inspection and risk its market dominance in the search engine arena. Unless courts can demonstrably protect such trade secrets while also testing their validity, applying the *Daubert* factors to many or most AI applications in open court may be difficult. (As discussed in chapter 8, jurists or lawmakers[76] may determine that defendants or the public should have access to certain underlying algorithms and data, such as in instances where liberty interests are at stake. Courts will need to determine whether the Fifth, Sixth, and Fourteenth Amendments require it.)

In other contexts, where courts seek to allow litigants to test the validity of AI applications while still protecting proprietary information, they might exercise their general power to oversee how evidence is entered, to enforce rulings, and to seal records. A parallel can be found in the way classified information is protected while still allowing certain litigation to proceed, with

[76] For example, an Idaho law, Section 19-1910 of the Idaho Code, states, "All pretrial risk assessment algorithms shall be transparent, and all documents, records, and information used to build or validate the risk assessment shall be open to public inspection, auditing, and test. No builder or user of a pretrial risk assessment algorithm may assert trade secret or other protections in order to quash discovery in a criminal matter by a party to a criminal case." https://legislature.idaho.gov/wp-content/uploads/sessioninfo/2019/legislation/H0118.pdf.

records reviewed by judges and sometimes cleared counsel. Also relevant is the 1996 Defend Trade Secrets Act (18 U.S.C. § 1835), which specifically directed federal courts to protect trade secrets in proceedings arising under Title 18 of the U.S. Code. Specifically, section (a) states,

> In any prosecution or other proceeding under this chapter, the court shall enter such orders and take such other action as may be necessary and appropriate to preserve the confidentiality of trade secrets, consistent with the requirements of the Federal Rules of Criminal and Civil Procedure, the Federal Rules of Evidence, and all other applicable laws.

In context, specific statutes also provide intellectual property protections for AI, such as those protections found in § 705 of the Defense Production Act, which allow the president in the first instance and courts in the second instance, through the power of contempt and jurisdiction found in § 706, to protect intellectual property relevant to DPA enforcement or defend against DPA actions.

8. AI in the Courtroom

AI issues will arise in virtually all areas of law. This section illustrates potential litigation scenarios and some of the complexities that AI will create.

Tort

Litigants might seek to introduce AI-derived evidence in a variety of tort law contexts, both as proof of probability and causation and as the underlying source of tort, as in cases alleging medical malpractice and involving AI-operated vehicle accidents. Consider a hypothetical malpractice example involving a hospital that uses AI to diagnose a particular disease. The AI diagnoses more accurately than most doctors. It is self-taught, and even its developer does not know exactly how it has learned to make decisions within its black box.

If the AI, or the healthcare provider interpreting its output, fails to catch a fatal case, who, if anyone, is liable? Plaintiff and defendant(s) might be interested in submitting evidence about the reliability of the AI. They might contest whether the AI developers or the hospital adequately protected against its potential for error and against any social and algorithmic bias. Was the AI properly vetted? Was it submitted to peer review?

The opposite scenario might arise too. An AI diagnostic might over-read a mammogram and provide a false positive result. If a subsequent surgery suggests there was never any cancer, should the medical provider be liable for using and then relying on the AI to inform the patient's decision?

Does informed consent address liability in either scenario? Where AI is used to make decisions, what does "informed consent" consist of? Does it require educating the patient about the error rate of the algorithm? What level of detail is adequate to inform a patient about the AI application used, or not used, in medical care?

Current law suggests that ultimately the healthcare provider as defendant would need to show that the use of the AI application in a particular circumstance was deemed acceptable by the medical community as a standard of care, that is, that it had come into generally accepted use for making predictions equal or superior to human diagnoses.

AI complicates the legal standard while also posing new issues for informed consent. Negligence in medical malpractice often comes down to whether the practitioner provided reasonable care under the circumstances. A practicing physician cannot always research issues to the *nth* degree and will lean on the practices and knowledge of other, equally trained physicians in the medical community. Even if the physician could research each AI application before relying on it, whatever the machine learns or simply processes inside its black box creates a moving target for the physician to understand, much less communicate to patients. It is therefore up to the medical community to establish best practices for testing and relying on individual AI applications. Courts in turn must determine whether and when AI applications might be appropriate to predict or assess contributing percentages of negligence. The law may not change, but it will need to keep pace.

The autonomous or semiautonomous car case might seem straightforward by comparison. Does the driver who does not brake fast enough pass the liability-buck to the car maker that promised automatic breaking but failed to deliver? If a car typically beeps to warn of another car in the driver's blind spot but fails to do so at the critical moment, can the driver successfully sue the car company for damages? The 2018 case of an Uber test vehicle that killed a pedestrian suggests that these issues may initially be settled out of court. (Prosecutors also did not find evidence to charge Uber with a crime.)[77] Assuming semi-autonomous and fully autonomous vehicles are here to stay, we can expect to see more accidents and lawsuits.[78] The potential defendants in autonomous vehicle accident cases are myriad. Who owned the vehicle? Who, if anyone, was driving or riding in it? Who manufactured it? Who subcontracted to manufacture its parts? To develop the software? To install the software? Traditional liability schemes—contributory and comparative negligence, strict liability and the "assumption of risk" defense, and vicarious liability—will be tested. A complicating question will be: What, if anything, went wrong inside the machine's black box; what blame, if any, lies there?

First Amendment

Every time the government, in law or practice, takes an action that can be construed as impeding, restricting, chilling, or favoring one voice or view over another, there is space for a First Amendment challenge. Inventors seeking patents, for example, might assert that the government is chilling free speech by preventing them from talking about their inventions under the Invention Secrecy Act. Consider the issues that might arise if the government sought to review and regulate Facebook postings for foreign interference or undertook

[77] Mihir Zaveri, *Prosecutors Don't Plan to Charge Uber in Self-Driving Car's Fatal Accident*, N.Y. Times (Mar. 5, 2019), https://www.nytimes.com/2019/03/05/technology/uber-self-driving-car-arizona.html.

[78] *See* Fredrick Kunkle, *Fatal Crash with Self-Driving Car Was a First—Like Bridget Driscoll's Was 121 Years Ago with One of the First Cars*, Wash. Post (Mar. 22, 2018), https://www.washingtonpost.com/news/tripping/wp/2018/03/22/fatal-crash-with-self-driving-car-was-a-first-like-bridget-driscolls-was-121-years-ago-with-one-of-the-first-cars/.

to validate the authenticity of political ads. Imagine the potential disputes arising over government funding for AI development, which, depending on how it was allocated or withheld, could create First Amendment issues. Some scholars think that the threat of government regulation, followed by a social media response to head off regulation, is sufficient governmental conduct to implicate the First Amendment.[79] Courts will likely have the opportunity to address each of these questions.

Think, too, of the effect of constant or perfect surveillance on First Amendment freedoms. Facial recognition applications are already in use, and many cities use security cameras extensively. Some, like London, are experimenting with allowing police to use facial recognition technology, while others, like San Francisco, have banned its use by government and law enforcement.[80] Real-time video surveillance devices are able to make predictive identity matches based on photo-memories no human mind could ever catalogue.[81]

It is easy to imagine the chilling effect AI surveillance may have on an individual's willingness to speak freely in public, to assemble with political or religious groups, or to worship as they wish.[82] One need only google China's use of AI for surveillance and, in some contexts, "social credit scores" to start worrying about First Amendment implications. One scholar, Margot Kaminski, has argued that the government has an interest in preventing the chilling effect of surveillance to foster a culture of free discourse and truth telling.[83] At the same time, she and others recognize the potential First Amendment interests of private actors, such as journalists or real estate professionals, in the developing right to record,[84] possibly with AI-enhanced technologies, such as

[79] Jed Rubenfeld, *Are Facebook and Google State Actors?* LAWFARE (blog), Nov. 4 2019.

[80] Evan Selinger & Woodrow Hartzog, Opinion, *What Happens When Employers Can Read Your Facial Expressions?*, N.Y. TIMES (Oct. 17, 2019), https://www.nytimes.com/2019/10/17/opinion/facial-recognition-ban.html.

[81] Gregory Barber & Tom Simonite, *Some US Cities Are Moving Into Real-Time Facial Surveillance*, WIRED (May 17, 2019), https://www.wired.com/story/some-us-cities-moving-real-time-facial-surveillance/.

[82] *See* Jennifer Lynch, *Face Off: Law Enforcement Use of Face Recognition Technology*, ELECTRONIC FRONTIER FOUNDATION 1, 8–10 (Feb. 12, 2018), https://www.eff.org/wp/law-enforcement-use-face-recognition#_idTextAnchor004.

[83] *See Margot E. Kaminski*, *Regulating Real-World Surveillance*, 90 WASH. L. REV. 1113, 1136–37, 1155–58 (2015) (Kaminski "conceptualize[s] privacy harm as interference in an individual's ability to dynamically manage disclosure and social boundaries. Stemming from this understanding of privacy, the government has two related interests in enacting laws prohibiting surveillance: an interest in providing notice so that an individual can adjust her behavior; and an interest in prohibiting surveillance to prevent undesirable behavioral shifts.")

[84] *See id.* at 1117. Some privacy torts and criminal prohibitions, such as the eavesdropping nuisance, Peeping Tom laws, and the tort of intrusion upon seclusion, might in turn butt up against any "right to record," but they tend to be applied to private rather than public settings.

facial-recognition enabled cameras and drones.[85]

The European Commission's proposed regulation of AI would prohibit "the placing on the market, putting into service or use of AI systems by public authorities or on their behalf" intended to evaluate the trustworthiness of individuals "based on their social behaviour or known or predicted personal or personality characteristics," where that "social score" might lead to certain, delineated unfavourable treatment.[86] The regulation would also prohibit the use of real-time biometric surveillance in public places "unless and in as far as such use is strictly necessary for one of the following objectives:

> (i) the targeted search for specific potential victims of crime, including missing children;
>
> (ii) the prevention of a specific substantial and imminent threat to the life or physical safety of natural persons or of a terrorist attack;
>
> (iii)the detection, localisation, identification or prosecution of a perpetrator or suspect of a criminal offence referred to in Article 2(2) of Council Framework Decision 2002/584/JHA 62 and punishable in the Member State concerned by a custodial sentence or a detention order for a maximum period of at least three years, as determined by the law of that Member State."[87]

Europe, it would seem, is concerned about AI surveillance.

Some of the most contentious First Amendment debates about AI may concern the threshold for initiating investigation of criminal conduct involving domestic extremism. Because AI-driven search engines and tools likely will be a key measure in identifying potential threats, Executive Branch lawyers and, subsequently, the courts will need to address the way First Amendment principles are embedded in code and whether the First Amendment "constraint" occurs when the algorithm identifies a posting of interest or when a law enforcement officer first looks at the posting and determines whether it meets the threshold for investigation. The FBI *Domestic Investigations and Operations Guide* states that "… investigative activity may not be based solely on the exercise of rights guaranteed by the First Amendment…."One pending question is: What constitutes a sufficient predicate beyond "solely First Amendment activities" to initiate investigation? *Brandenburg v. Ohio* (1969)

[85] *Id.* at 1122.

[86] European Commission, Proposal for a Regulation of the European Parliament and of the Council Laying Down Harmonised Rules on Artificial Intelligence (Artificial Intelligence Act) and Amending Certain Union Legislative Acts, title II, art. 5(1)(c) (Apr. 21, 2021), https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A52021PC0206.

[87] *Id.* at title II, art. 5(1)(d).

provides a partial answer.[88] First Amendment principles, the Court concluded, "do not permit a State to forbid or proscribe advocacy of the use of force or of law violation except where such advocacy is directed to inciting or producing imminent lawless action *and* is likely to incite or produce such action."[89] But *Brandenburg* predates social media and machine-speed search algorithms. No doubt, courts will be asked to address these AI predicates in the context of criminal prosecutions and First and Fourth Amendment motions to suppress evidence.

Deepfakes present another area that might implicate First Amendment freedom of speech. First Amendment speech protections would preclude an outright ban on deepfakes, protecting Hollywood productions, artistic performances, and comedy routines, but legislatures might take a more "surgical approach" to imposing criminal and civil liability for harmful deepfakes, such as those intended to incite violence, defame private persons, or sexualize children.[90]

Fourth Amendment

The Fourth Amendment bolsters First Amendment rights, and vice versa. The founders drafted the Fourth Amendment in response to the British use of the "general warrant" (and its equivalent "writs of assistance" in America) to search private premises at will, including the homes and shops of "dissidents, authors, and printers of seditious material."[91] The Fourth Amendment introduced reasonableness, probable cause, a neutral magistrate, and particularity, providing:

> The right of the people to be secure in their persons, houses, papers, and effects, against unreasonable searches and seizures, shall not be violated, and no Warrants shall issue, but upon probable cause, supported by Oath or affirmation, and particularly describing the place to be searched, and the persons or things to be seized.

AI-enabled data aggregation and data mining, link analysis, cameras, drones, facial recognition, et cetera have the potential to create a system of what Chief Justice Roberts might recognize as "near perfect surveillance," as he described cell phone location tracking in *Carpenter v. United States*.[92] Such

[88] Brandenburg v. Ohio, 395 U.S. 444 (1969).

[89] *Id.* at 447 (emphasis added).

[90] Citron, *supra* note 49, at 1790–91.

[91] William C. Banks & M.E. Bowman, *Executive Authority for National Security Surveillance*, 50 Am. U. L. Rev. 1, 2–3 (2000).

[92] 1138 S. Ct. 2206, 2210 (2018). *See also* Ryan Calo, *Artificial Intelligence Policy: A Primer and Roadmap*, 51 U.C. Davis L. Rev. 399, 423 (2017) ("Even assuming away the likely false positives, a reasonable question for law and policy is whether we want to live in a society with

surveillance is possible not only in public spaces but also in our homes and offices,[93] via the Internet of Things, including connected cars and appliances, wearable heath monitors, home security systems, and much more. AI is not the first technology to pose Fourth Amendment questions and challenges related to invasiveness and government incursions on privacy; however, excepting perhaps the telephone, AI's potential scope and impact seems unrivaled.

Fourth Amendment analysis about modern technologies has turned largely on whether use constitutes a search, a question courts generally have addressed by applying the reasonable expectation of privacy test and the third-party doctrine.

The reasonable expectation of privacy test emerged in 1967 with *Katz v. United States*. In determining that police needed a warrant to tap a public phone booth, the Supreme Court extrapolated from the framers' "persons, houses, papers, and effects" points of reference. The Court held that warrantless wiretaps were unreasonable, reasoning that the Fourth Amendment protects people, not places.[94] In a concurring opinion, Justice Harlan authored the reasonable expectation of privacy test still in use today.[95] That test considers whether an individual has a subjective expectation of privacy that society also recognizes to be reasonable; if so, that interest is constitutionally protected, and any government intrusion on it is presumptively unreasonable in the absence of a warrant.[96]

The reasonable expectation test has the advantage of being capacious and dynamic as technology improves.[97] On the other hand, it arguably is not very protective as society's expectations of privacy dwindle in the age of AI and the Internet of Things. That potential shortcoming is traceable in part to the third-party doctrine arising in the late 1970s in *Smith v. Maryland* and *United States v. Miller.*

The third-party doctrine posits that someone who voluntarily shares information with a third party loses any objectively reasonable expectation in its privacy and "assumes the risk" the third party may share that information with the government.[98] *Miller* held that law enforcement's acquisition of financial

perfect enforcement.").

[93] There is a growing body of literature on AI in the workplace. *See, e.g.*, Karen E. C. Levy, *The Contexts of Control: Information, Power, and Truck-Driving Work,* 31 Info. Soc'y 160–74 (2015), https://www.tandfonline.com/doi/full/10.1080/01972243.2015.998105 (last visited Oct. 27, 2019).

[94] Katz v. United States, 389 U.S. 347, 351 (1967).

[95] *Id.* at 360 (Harlan, J., concurring).

[96] *Id.* at 360–61.

[97] Stephen Dycus et al., National Sec. L., Sixth Ed., Wolters Kluwer (2016) Teachers' Manual, 24-3.

[98] *Smith v. Maryland*, 442 U.S. 735, 744 (1979) (citing United States v. Miller, 425 U.S. 435, 442–44 (1976)).

information conveyed by a bank depositor to his bank was not a search within the meaning of the Fourth Amendment.[99] *Smith* held that the police's request that a phone company install a pen register at its central office to record the numbers a suspect dialed was likewise not a search for Fourth Amendment purposes.[100] Third-party doctrine draws a distinction between content information, in which one has a reasonable expectation of privacy, and business records. The *Smith* Court argued that "a pen register differs significantly from the listening device employed in *Katz*, for pen registers do not acquire the *contents* of communications."[101]

With today's technologies, the relevant questions are what information we "voluntarily" convey to service providers and whether the Court will treat that information as content requiring a warrant or as business records exempt under the third-party doctrine. AI raises the stakes by potentially allowing private actors or the government to compile and analyze data at tremendous speed and scale, deriving content-like meaning from what was heretofore treated as telephonic or location metadata. Another question is the extent, if any, to which courts will allow "retroactive warrants," i.e., the search of stored data potentially going back years, if not decades. With the advent of cloud computing, that data can now be stored indefinitely, and it can be aggregated and searched with AI tools.

The Supreme Court has considered the Fourth Amendment implications of modern technologies in two broad categories: (1) where the government uses technology to surveil people directly and (2) where the government obtains data via the third-party doctrine from private actors who have collected it.

Direct surveillance. A series of 1980s aerial surveillance cases may be of interest to courts facing questions about AI-enabled drones. Three cases held that certain aerial surveillance by law enforcement from publicly navigable airspace did not constitute a search within the meaning of the Fourth Amendment. In 1986, the Court decided that criminal defendants did not have a reasonable expectation of privacy that would preclude surveillance of the curtilage of the home by plane at 1,000 feet altitude[102] or open areas of an industrial complex by plane at 1,200 feet.[103] In 1989, the Court concluded in *Florida v. Riley*[104] that surveillance of a backyard by helicopter at 400 feet was not a search within the meaning of the Fourth Amendment.

[99] *Smith*, 442 U.S. at 744.

[100] *Id.* at 737.

[101] *Id.* at 741 (emphasis in original).

[102] California v. Ciraolo, 476 U.S. 207, 215 (1986).

[103] Dow Chem. Co. v. United States, 476 U.S. 227, 239 (1986).

[104] 488 U.S. 445, 455 (1989).

These cases may become specifically relevant in the context of domestic drones used by law enforcement (or by private actors whose records law enforcement subpoenas), or generally relevant as courts consider evolving concepts of privacy.[105] Drones may be equipped with AI-enabled operating systems, allowing them to fly autonomously or semi-autonomously to gather evidence, or with AI-enabled sensors such as facial recognition. Will police need a warrant to use those drones in the publicly navigable airspace above or near a home or business? [106]

Concurring in the judgment in *Florida v. Riley*, Justice O'Connor observed that "public use of altitudes lower than [400 feet]—particularly public observations from helicopters circling over the curtilage of a home—may be sufficiently rare that police surveillance from such altitudes would violate reasonable expectations of privacy . . ."[107] In a dissenting opinion, Justice Brennan wrote,

> Imagine a helicopter capable of hovering just above an enclosed courtyard or patio without generating any noise, wind, or dust at all—and, for good measure, without posing any threat of injury. Suppose the police employed this miraculous tool to discover not only what crops people were growing in their greenhouses, but also what books they were reading and who their dinner guests were. Suppose, finally, that the FAA regulations remained unchanged, so that the police were undeniably "where they had a right to be."[108]

We need no longer imagine such "miraculous tools."[109] They are here, and they are called drones. FAA regulations currently allow for commercial small drone flight below 400 feet, with certain conditions such as the operator keeping the drone in line of sight.[110] (Operators must apply for a waiver for flights over 400 feet.) Law enforcement may fly drones under those same

[105] *See* Troy A. Rule, *Airspace In An Age Of Drones*, 95 B.U. L. Rev. 155, 172–74 (2015); Gregory S. McNeal, *Drones and the Future of Aerial Surveillance*, 84 Geo. Wash. L. Rev. 354, 373–83 (2016).

[106] *See* Kyllo v. United States, 533 U.S. 27, 33 (2001) (quoting *Dow Chemical*, 476 U.S. at 237 n. 4) ("We have previously reserved judgment as to how much technological enhancement of ordinary perception from such a vantage point, if any, is too much. While we upheld enhanced aerial photography of an industrial complex in *Dow Chemical*, we noted that we found 'it important that this is *not* an area immediately adjacent to a private home, where privacy expectations are most heightened[…]'").

[107] *Florida v. Riley*, 488 U.S. at 455 (O'Connor, J., concurring); *see* McNeal, *supra* note 104, at 377.

[108] *Riley*, 488 U.S. at 462 (Brennan, J., dissenting).

[109] *See* McNeal, *supra* note 105, at 383; Rule; *supra* note 105, at 174.

[110] Small Unmanned Aircraft Systems, 14 C.F.R. Part 107, https://www.ecfr.gov/cgi-bin/text-idx?node=pt14.2.107&rgn=div5.

conditions or apply for a waiver for public drone use.[111] Drones are potentially more discreet than manned airplanes and helicopters, able to approach a residence more closely and quietly, and unlike street cameras, they are mobile. If using AI facial recognition or making their own operating decisions, such as how to tail a suspect,[112] drones may be even more invasive. Some, but not all, states are moving toward warrant requirements for drones. Again, AI magnifies and complicates the privacy implications of technology with its capacity to aggregate and search data permanently stored in the cloud for meaning that previously would have been retrievable only through warrant-authorized content searches.

In a more recent line of cases, the Supreme Court has tended toward requiring a warrant to use modern technology in criminal searches or to search the technology itself. In 2001, in *Kyllo v. United States*, the Court held that law enforcement needed a warrant before using a thermal-imaging device to detect heat prints emanating from a private home, where, the Court observed, the technology was not yet in "general public use."[113] The case turned on whether the thermal search did or did not penetrate into the home, a question on which the Court divided 5-4. *Kyllo* might limit police use of AI-enabled technology, at least so long as the relevant AI application is not in general public use. But given the iterative nature of most AI applications and thus the difficulty of pinpointing if an application has become commonplace, the "general public use" dictum would seem to offer only modest guidance. In *United States v. Jones* (2012), the Court applied a trespass theory of the Fourth Amendment (concurring opinions applied a reasonable expectation of privacy theory) in deciding that law enforcement needed a warrant to attach a GPS tracker to a suspect's vehicle and track its movements for over four weeks.[114] In *Riley v. California* (2014), the Court held that police could not search a person's cellphone pursuant to the "search incident to arrest" exception to the warrant clause, concluding that a digital search of a cell phone was much more invasive than a physical search of the materials on a person's body.[115] Presumably, enhancing technologies and searches with AI will only increase the individual privacy interests at stake. But courts will still need to address competing governmental interests potentially achieved by AI on a case-by-case, or AI

[111] *Drones in Public Safety: A Guide to Starting Operations*, Federal Aviation Administration (Feb. 2019), https://www.faa.gov/sites/faa.gov/files/uas/public_safety_gov/public_safety_toolkit/Law_Enforcement_Drone_Programs_Brochure.pdf.

[112] Cade Metz, *Police Drones Are Starting to Think for Themselves*, N.Y. Times, Dec. 5, 2020, https://www.nytimes.com/2020/12/05/technology/police-drones.html.

[113] Kyllo v. United States, 533 U.S. 27, 31–41 (2001).

[114] 565 U.S. 400, 404–05 (2012). For a discussion of *Jones*, see Baker, *supra* note 14, at 113–14.

[115] 573 U.S. 373, 385–98 (2014).

application-by-application, basis. As the Court caveated in the 2018 *Carpenter v. United States*[116] decision, context matters. That context might be the type of information searched, or the government's purpose in searching, such as for criminal law enforcement or national security ends.

The third-party doctrine. The 2018 *Carpenter* decision did not involve AI but appears most apt for AI. With that decision, the Supreme Court continued its trend of requiring a warrant to use or search with a modern technology. *Carpenter* "declin[ed] to extend" the third-party doctrine to "a new phenomenon: the ability to chronicle a person's past movements through the record of his cell phone signals," specifically, 127 days' worth of cell-site-location-information (CSLI) that the government had subpoenaed from Carpenter's service provider.[117] It was not enough for law enforcement to obtain a court-ordered subpoena, based on the reasonable suspicion and relevancy standard[118] in the Stored Communications Act; rather, law enforcement use of historical CSLI required a warrant based on probable cause.

The Court described CSLI information as being like the GPS vehicle tracking in *Jones*: "detailed, encyclopedic, and effortlessly compiled."[119] Chief Justice Roberts, quoting Justice Sotomayor in *Jones*, wrote, "As with GPS information, the time-stamped data provides an intimate window into a person's life, revealing not only his particular movements, but through them his 'familial, political, professional, religious, and sexual associations.'"[120] The Court distinguished the "exhaustive chronicle" and "revealing nature" of information provided by CSLI records from "the limited types of personal information" collected by pen register and in bank records in *Smith* and *Miller*.[121]

The Court noted, too, that most people carry cell phones everywhere, and that CSLI records are typically held by wireless carriers for up to five years, suggesting that law enforcement could look back retrospectively.[122] "Given the unique nature of cell phone location information," the Court concluded, "the fact that the Government obtained the information from a third party does not overcome Carpenter's claim to Fourth Amendment protection."[123] The Court limited its holding, however, to the facts before it:

[116] 138 S. Ct. 2206 (2018).

[117] *Id.* at 2206.

[118] The Stored Communications Act, as amended in 1994, "permits the Government to compel the disclosure of certain telecommunications records when it 'offers specific and articulable facts showing that there are reasonable grounds to believe' that the records sought 'are relevant to an ongoing criminal investigation.'" *Id.* at 2212 (citing 18 U.S.C. § 2703(d)).

[119] *Carpenter*, 138 S. Ct. at 2216.

[120] *Id.* at 2217.

[121] *Id.* at 2219.

[122] *Id.* at 2218.

[123] *Id.* at 2220.

> Our decision today is a narrow one. We do not express a view on matters not before us: real-time CSLI or "tower dumps" (a download of information on all the devices that connected to a particular cell site during a particular interval). We do not disturb the application of *Smith* and *Miller* or call into question conventional surveillance techniques and tools, such as security cameras. Nor do we address other business records that might incidentally reveal location information. Further, our opinion does not consider other collection techniques involving foreign affairs or national security.[124]

Given this narrowing language, it will be up to lower courts to determine how *Carpenter* applies to new, AI-enabled technologies or applications. The trend in the last two decades points to the Court favoring a warrant requirement for invasive emerging technologies or technologies capable of collecting aggregate data over time. Security cameras, it appears, are still covered by the plain-sight doctrine, but what of security cameras (or drones) with AI-enabled facial recognition? What if those cameras can instantly search their archives for all pictures of a person, creating a historical record across a web of cameras of comings and goings, perhaps for the past five years? At least with respect to CSLI, the Court required a warrant for a retrospective search. The Court made clear it was not opining on real-time CSLI. Can law enforcement subpoena security cameras in real time and connect them to other AI-enabled databases that combine facial recognition with instant feedback on a person's criminal and financial records? In either instance, retrospective or real time, the plain view captured on camera is no longer so plain.

Carpenter may well signal the beginning of the end of the third-party doctrine. Even outside the criminal context, it may suggest implications for the data used in ML. If the Supreme Court was nervous about the aggregation of cell tower data in *Carpenter*—data collected pursuant to legislative authorization—imagine the Court's concern when it looks at data collection and use for ML. United States constitutional and statutory law traditionally address limiting the role of government; they do not address privacy. That may have to change in the context of AI.

It is all but certain that no single case or principle will, or ever can, "address AI." Consider alone the myriad caveats in *Carpenter* limiting its reach, each caveat now requiring its own resolution. Rather, courts will need to address AI in its separate parts across a constellation of technologies. Absent overriding statutory guidance, the technology's complexity will make it harder to discern and apply black-letter rules to AI's development and use. In the box on page 70 we suggest legal policy questions that may help legislators and

[124] *Id.*

The Fourth Amendment and AI: Some General Questions

- Should courts rely on AI outputs as predicates for Fourth Amendment search warrants? If so, with what underlying inquiry into the algorithm's design; training and other data; and output accuracy?
- Should the federal and state governments be able to use AI-enabled technologies to surveil Americans? Should private persons and companies be permitted to do so? If so, under what predicate conditions? With what limitations on data collection, use, retention, and dissemination?
- Under what conditions should the government have access, via the third-party doctrine and its attendant subpoena statutes, to the business records and metadata of private parties using AI-enabled technologies that collect consumer or employee data? Should the government be able to purchase third-party data for AI development purposes as private actors do? Should these conditions be determined by legislative enactment or judge-made doctrines and rules?
- To what extent should the government be permitted, via AI, to combine and then use different types of information? Should the law permit AI sorting through databases that include not only faces, but also financial, tax, travel, internet search, DNA, and driving records, and any other data citizens might have "voluntarily" disclosed to third parties or the government?
- Is a particular AI reliable, and does its use change accepted norms or principles of our justice system? What safeguards, if any, are required or warranted? Many facial-recognition systems, including the FBI's, do not make exact matches; rather, given a fixed data set, they determine and rank which photos within that set are most likely to match.[125] Facial recognition might flag someone five states away from the crime scene, but make them subject to a probable-cause search warrant, shifting the burden from the state proving guilt to the suspect proving innocence.[126]
- Is there any social or other algorithmic bias in the AI used or accessed by law enforcement potentially leading to inaccurate predictions that undermine probable cause and other Fourth Amendment predicates?[127]

[125] Lynch, *supra* note 82.

[126] *Id.* "False positives can alter the traditional presumption of innocence in criminal cases by placing more of a burden on suspects and defendants to show they are not who the system identifies them to be. This is true even if a face recognition system offers several results for a search instead of one; each of the people identified could be brought in for questioning, even if there is nothing else linking them to the crime. Former German Federal Data Protection Commissioner Peter Schaar has noted that false positives in face recognition systems pose a large problem for democratic societies: '[I]n the event of a genuine hunt, [they] render innocent people suspects for a time, create a need for justification on their part and make further checks by the authorities unavoidable.'" *Id.*

[127] For two excellent studies on the use of facial recognition in policing, see Lynch, *supra* note 80, and Claire Garvie, Alvaro M. Bedoya & Jonathan Frankle's *The Perpetual Lineup: Unregulated Police Face Recognition in America*, Georgetown Law Center on Privacy & Technology (Oct. 16, 2016), https://www.perpetuallineup.org/.

judges assess the Fourth Amendment implications of AI tools, as well as when to statutorily permit or prohibit such use.

Fifth and Fourteenth Amendments

Machine learning black boxes, or "legal" black boxes, where parties are not permitted to inquire into an algorithm's parameters and weights, raise specialized due process and equal protection concerns. Two categories of cases where an individual's life or liberty may be at stake illustrate how an AI application might determine or effect due process: (1) criminal justice risk assessments and (2) government watch lists. These cases may also present equal protection and First Amendment issues.

Criminal justice risk assessments

Many police departments and courts across the country use algorithmic risk assessments.[128] Police use such tools to predict where crime might occur and by whom.[129] Some courts use them in pretrial release, probation, and sentencing decisions; parole boards also make use of them.[130] Some of these algorithmic risk assessments are capable of machine learning,[131] a capacity that will increase with time. Private companies may develop the risk assessment algorithms; Northpointe developed the COMPAS system used by several states.[132] These companies may not release the underlying code for the algorithms for defendants to test and challenge.[133]

Due Process. Using risk assessment algorithms to make or inform liberty decisions creates potential Fifth and Fourteenth Amendment due process issues. (It also creates Sixth Amendment Confrontation Clause questions, as discussed earlier.) To identify a few:

- May a defendant meaningfully challenge the logic of an algorithm if the source code is kept from them? Is it enough for them to have access only to the inputs and outputs the algorithm processes and generates but not the decisional framework it uses?

[128] Randy Rieland, *Artificial Intelligence is Now Used to Predict Crime. But Is It Biased?* Smithsonian Mag. (Mar. 5, 2018), https://www.smithsonianmag.com/innovation/artificial-intelligence-is-now-used-predict-crime-is-it-biased-180968337/; *AI in the Criminal Justice System*, Elec. Priv. Info. Ctr., https://epic.org/algorithmic-transparency/crim-justice/ (last visited June 4, 2021).

[129] Rieland, *supra* note 128.

[130] *AI in the Criminal Justice System*, *supra* note 128.

[131] Danielle Kehl, Priscilla Guo, & Samuel Kessler, *Algorithms in the Criminal Justice System: Assessing the Use of Risk Assessments in Sentencing*, Responsive Cmtys. Initiative (July 2017), https://dash.harvard.edu/bitstream/handle/1/33746041/2017-07_responsivecommunities_2.pdf.

[132] *AI in the Criminal Justice System*, *supra* note 128.

[133] *See id.*

- May a defendant meaningfully challenge an algorithmic risk assessment without access to its training, testing, and real-world-use data?
- If the algorithm uses machine-learning, and no one, not even the developer, understands its "analysis," can courts ensure due process of law?
- How many courts test algorithms for accuracy, especially when they predict future (i.e., unrealized) human behavior?

Equal Protection. Racial and other biases contained in or produced by algorithms present equal protection issues. The adoption of risk assessment tools has caused much controversy in this context,[134] and there is a rich academic literature on the efficacy and fairness of these tools.[135] Lawmakers or police departments using these tools might seek to replace, improve, or inform judicial decisions with "evidence-based"[136] algorithmic recommendations, or to decrease the incarceration rate by releasing more people before trial and during probation.[137] Critics argue that risk assessment tools not only have racially biased results but, through the ML process, exacerbate racial inequalities in the criminal justice system. To quote from *MIT Technology Review*, "Using historical data to train risk assessment tools could mean that machines are copying the mistakes of the past."[138] Over one hundred civil rights groups issued a joint statement detailing their concerns with pretrial risk assessments.[139]

[134] For example, a 2016 ProPublica study determined that COMPAS was almost twice as likely to falsely identify a black person as a repeat violent offender as it was to falsely identify a white person as a repeat offender. The company contested this finding. Julia Angwin et al., *Machine Bias: There's Software Used Across the Country to Predict Future Criminals. And It's Biased Against Blacks*, ProPublica (May 23, 2016), https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing. *See also* Sam Davies-Corbett et al., *A Computer Program Used for Bail and Sentencing Decisions Was Labeled Biased Against Blacks. It's Actually Not That Clear.*, Wash. Post (Oct. 17, 2016), https://www.washingtonpost.com/news/monkey-cage/wp/2016/10/17/can-an-algorithm-be-racist-our-analysis-is-more-cautious-than-propublicas/.

[135] For an introductory overview of that literature, see *A Letter to the Members of the Criminal Justice Reform Committee of Conference of the Massachusetts Legislature Regarding the Adoption of Actuarial Risk Assessment Tools in the Criminal Justice System* (Feb. 9, 2018), https://medium.com/berkman-klein-center/a-letter-to-the-members-of-the-criminal-justice-reform-committee-of-conference-of-the-massachusetts-2911d65969df.

[136] State v. Loomis, 881 N.W.2d 749, 759 (Wis. 2016), *cert. denied*, 137 S. Ct. 2290 (2017).

[137] Derek Thompson, *Should We Be Afraid of AI in the Criminal-Justice System?* Atlantic (June 20, 2019), https://www.theatlantic.com/ideas/archive/2019/06/should-we-be-afraid-of-ai-in-the-criminal-justice-system/592084/.

[138] Karen Hao, *AI Is Sending People to Jail—and Getting It Wrong*, MIT Tech. Rev. (Jan. 21, 2019), https://www.technologyreview.com/s/612775/algorithms-criminal-justice-ai/.

[139] The Use of Pretrial "Risk Assessment" Tools: A Shared Statement of Civil Rights Concerns, http://civilrightsdocs.info/pdf/criminal-justice/Pretrial-Risk-Assessment-

While the ideal of "evidence-based" practice may be appealing, the risk that an assessment tool may cause disparate treatment under a mantel of "data-driven" legitimacy warrants careful consideration. Whether any type of unbiased machine neutrality or fairness is possible is a matter of debate.[140]

State v. Loomis. Appendix B lists federal and state cases addressing AI. However, one Wisconsin Supreme Court case, *State v. Loomis,*[141] is worth highlighting because of the attention it has received in the AI literature from legal practitioners and scholars.[142]

The case addressed the use of the Correctional Offender Management Profiling for Alternative Sanctions (COMPAS) risk-need assessment tool produced by Northpointe, a Michigan-based computer hardware and software company. (Northpointe later merged with two other companies to become Equivant.) Known as a case management suite, COMPAS became controversial as a predictive algorithm for at least two reasons. First, the trial court used the algorithm during sentencing rather than to assess the parole risk or treatment needs for which the tool was designed. Second, as one of the first opinions in the United States addressing AI, the Wisconsin Supreme Court decision became a focal point of academic attention. (The U.S. Supreme Court declined to hear the case.) The State of Wisconsin charged the defendant, Eric Loomis, with five criminal offenses deriving from a drive-by shooting in 2013 involving a stolen car. As stated in his Supreme Court cert petition, "Mr. Loomis denied that he had any involvement in the drive-by shooting and maintained only that he later drove the car after the shooting." Loomis pleaded guilty to two lesser included offenses: attempting to flee or elude a traffic officer, as a repeater, and operating a motor vehicle without the owner's consent, as a party to a crime and as a repeater. He also agreed that the state could read-in the dismissed charges, a procedure where the defendant does not admit guilt but the trial judge may consider the charges, in effect, as aggravating evidence for sentencing.

During sentencing the state argued that the circuit court (trial court) should use the COMPAS report when determining an appropriate sentence and that the report indicated the "high risk and high needs of the defendant."[143] The trial judge stated, "In terms of weighing the various factors, I'm

Full.pdf (last visited Oct. 27, 2019).

[140] *See, e.g.*, Bavitz *et al.*, *supra* note 60, at 20–21, and Craig Smith, *Dealing with Bias in Artificial Intelligence: Three Women with Extensive Experience in A.I. Spoke on the Topic and How to Confront It,* N.Y. Times (Nov. 19, 2019, updated Jan. 2, 2020), https://www.nytimes.com/2019/11/19/technology/artificial-intelligence-bias.html.

[141] 881 N.W. 2d 749 (Wis. 2016); *cert. denied*, 137 S. Ct. 2290 (2017).

[142] *Loomis* was decided on due process rather than equal protection grounds.

143 All quotes are from the decision of the Wisconsin Supreme Court unless otherwise indicated.

ruling out probation because of the seriousness of the crime and because your history, your history on supervision, and the risk assessment tools that have been utilized, suggest that you're extremely high risk to re-offend." The court sentenced Loomis to six years' confinement and five years of supervised release.

The defendant filed two motions requesting a sentence rehearing. One challenged the trial judge's use of the "read-in" charges as part of sentencing. The other challenged the use of COMPAS in sentencing on Fourteenth Amendment due-process grounds. In support of the COMPAS motion, the defendant called an expert witness who testified, "The Court does not know how the COMPAS compares that individual's history with the population that it's comparing them with. The Court doesn't even know whether that population is a Wisconsin population, a New York population, a California population.... There's all kinds of information that the court doesn't have...." The Wisconsin Supreme Court stated, "In denying the post-conviction motion, the circuit court explained that it used the COMPAS risk assessment to corroborate its findings and that it would have imposed the same sentence regardless of whether it considered the COMPAS risk scores." Loomis appealed to the Court of Appeals, which certified the case to the Wisconsin Supreme Court without an opinion.

Loomis made three arguments before the Wisconsin Supreme Court: First, the use of the COMPAS tool for sentencing purposes violated his due process rights to be sentenced based on accurate information, "in part because the proprietary nature of COMPAS prevents him from assessing its accuracy." Second, use of an algorithm based on group statistics violated his right to individualized sentencing. And third, the algorithm "improperly used gender assessments in sentencing"—as the Wisconsin Supreme Court noted, the risk assessment tool "compares each offender to a 'norming' group of his or her own gender."

In its decision, the Wisconsin Supreme Court described COMPAS as

> a risk-need assessment tool ... to provide decisional support for the Department of Corrections when making placement decisions, managing offenders, and planning treatment. The COMPAS risk assessment is based upon information gathered from the defendant's criminal file and an interview with the defendant. A COMPAS report consists of a risk assessment designed to predict recidivism and a separate needs assessment for identifying program needs in areas such as employment, housing and substance abuse. The risk assessment portion of COMPAS generates risk scores displayed in the form of a bar chart, with three bars that represent pretrial recidivism risk, general recidivism risk, and violent recidivism risk. Each bar indicates a

> defendant's level of risk on a scale of one to ten. (Paras. 13–16)

The court's opinion also stated,

> COMPAS provides a prediction based on a comparison of information about the individual to a similar data group.
>
> The PSI [Presentence Investigation Report] also cautions that a COMPAS risk assessment should not be used to determine the severity of a sentence or whether an offender is incarcerated. (Paras. 51, 54, 58)
>
> A recent analysis of COMPAS's recidivism scores based upon data from 10,000 criminal defendants from Broward County, Florida, concluded that Black defendants 'were far more likely than white defendants to be incorrectly judged to be at a higher risk of recidivism.' Likewise, white defendants were more likely than Black defendants to be incorrectly flagged as low risk. Although Northpointe disputes this analysis, this study and others raise concerns regarding how a COMPAS assessment's risk factors correlate with race. [Citations omitted.] (Para. 63)

The court concluded with two justices concurring that

> using a risk assessment tool to determine the length of a sentence is a poor fit. As scholars have observed, "[a]ssessing the risk of future crime plays no role in sentencing decisions based solely on backward-looking perceptions of blameworthiness, … is not relevant to deterrence, … and should not be used to sentence offenders to more time than they morally deserve." (Para. 97)
>
> Thus, a sentencing court may consider a COMPAS risk assessment at sentencing subject to the following limitations. As recognized by the Department of Corrections, the PSI instructs that risk scores may not be used: (1) to determine whether an offender is incarcerated; or (2) to determine the severity of sentence. Additionally, risk scores may not be used as the determinative factor in deciding whether an offender can be supervised safely and effectively in the community. Importantly, a circuit court must explain the factors in addition to a COMPAS risk assessment that independently support the sentence imposed. A COMPAS risk assessment is only one of many factors that may be considered and weighed at sentencing." (Paras. 98–99)

The court then directed that any PSI filed with a circuit court "must contain a written advisement listing the limitations" along with five advisements stating among other things that "the proprietary nature of COMPAS has been invoked to prevent disclosure of information relating to how factors are weighted or how risk scores are determined," and that "risk assessment scores are based on group data … not a particular high risk individual." (Paras. 99–100)

Applying its analysis to the facts of the case, the court concluded:

> The circuit court here was aware of the limitations.
>
> [T]he court essentially gave it little or no weight.
>
> The circuit court … used the COMPAS risk assessment to corroborate its findings and that it would have imposed the same sentence regardless of whether it considered the COMPAS risk scores.
>
> Ultimately, although the circuit court explained at the post-conviction hearing, it would have imposed the exact same sentence without it. Accordingly, we determine that the circuit court's consideration of COMPAS in this case did not violate Loomis's due process rights. (Para. 110)

The court analyzed COMPAS and its use in sentencing but did not affirm its use in sentencing. Rather, the court appears to have engaged in harmless-error analysis (concluding that if there was an error, it was harmless). The court included so many caveats and cautions, including a statement that "COMPAS should not be used for sentencing," it is hard to see how a trial court might successfully use COMPAS for sentencing without substantial risk of error. These caveats also limit *Loomis*'s value as precedent for the use of AI tools in sentencing. Although the court caveated that COMPAS might be used for sentencing as "one of many factors," its statements to the contrary (may not be used "to determine the severity of a sentence") suggest caution.[144] Restated, *Loomis* is a better vehicle to evaluate legal and policy considerations about the use of AI assessment tools than it is a precedent on which to build an AI case law foundation.

Academic commentary about *Loomis* focuses on at least four lines of inquiry: (1) the risk that the existence of risk-assessment tools will place pressure on courts to use the tools, whether they are accurate or not; (2) the risk of embedded racism within algorithms based on demographic, location, and socioeconomic factors, which can serve as proxies for race; (3) the psychological bias toward relying on empirical evidence more heavily than nonempirical evidence ("anchoring bias"); and (4) the risk that "most judges are unlikely to understand algorithmic risk assessments," and therefore may misuse them or

[144] In *State v. Jones*, No. 2015AP2211-CRNM, 2016 WL 8650489 (Wis. Ct. App. Nov. 29, 2016), *State v. Spivery*, No. 2015AP2565-CRNM, 2016 WL 8650373 (Wis. Ct. App. Nov. 18, 2016), and *State v. Booker*, No. 2015AP1253-CRNM, 2016 WL 8614037 (Wis. Ct. App. Sep. 14, 2016), all cases tried before *Loomis* was decided, the Wisconsin Court of Appeals subsequently upheld the use of COMPAS during sentencing citing *Loomis* and noting that the "trial court commented on the [COMPAS] report only briefly" and the COMPAS report "was one of many factors" the trial court considered.

give them inappropriate weight.[145]

In addition to highlighting some of the arguments advanced by litigants and commentators about risk assessments, the *Loomis* case prompts several questions about whether to use AI-enabled tools for judicial decision making or to admit AI-generated outputs into evidence.

We would encourage judges to look under the hood. While we believe judges can understand AI driven tools and evidence, we are skeptical they can reasonably understand an AI tool or admit AI outputs into evidence without knowing not only the AI inputs—in the *Loomis* case the interview questions posed to the defendant, of which the defendant and the trial court were aware, and also information gleaned from his criminal file, the specifics of which it is unclear whether the defendant and trial court were aware (para. 54)—but also the weights that were attached and allocated to each input, what data COMPAS was trained on, what data was in the "similar" data group, how the corresponding outputs were compared to the "similar" data group along, and an explanation of the methodologies used for prediction. Judges might ask the following questions:

- For what purpose was the AI designed, trained, tested, and validated? Is that the purpose for which the court is considering its use? If not, why is the court admitting or using the AI for an alternative purpose? What safeguards is the court using or imposing to ensure appropriate use in context. Will they suffice?
- On what data inputs was the AI trained, tested, and validated? Did the data inputs or labels include suspect category information or proxies for it?
- What parameters did the algorithm search and what weight was given to those parameters in the AI output? Are such parameters and weights discoverable? Do those parameters include race, gender, other suspect categories, or their proxies as factors? If so, why, and do they pass ethical and constitutional review?
- Does the AI have equal or disparate error rates across different racial, gender, or other suspect categories? (*See* "Probing for Bias").
- Was the defendant's data included in the training data for the algorithm in question? If so, is it possible to expunge that data?

[145] *State v. Loomis: Wisconsin Supreme Court Requires Warning Before Use of Algorithmic Risk Assessments in Sentencing, Recent Case,* 130 Harv. L. Rev. 1530, 1535 (Mar. 2017), https://harvardlawreview.org/wp-content/uploads/2017/03/1530-1537_online.pdf; Ellora Israni, *Algorithmic Due Process: Mistaken Accountability and Attribution in State v. Loomis*, JOLT Digest (Aug. 31, 2017), https://jolt.law.harvard.edu/digest/algorithmic-due-process-mistaken-accountability-and-attribution-in-state-v-loomis-1.

- Does the application rely on a neural network? If so, what is the risk that the system will rely on parameters that are unintended or unknown to the designers or operators? Is it possible to identify those potential parameters?
- Does the algorithm use the most accurate, reliable, and transparent methodologies available?
- If the answers to these questions implicate proprietary information, what would prevent the court from hearing answers to the questions in camera, or even ex parte, with appropriate protective orders and the power of contempt to enforce the court's orders? Does due process require that the defendant have access to all of the information at issue?

If the moving party cannot answer these questions to the satisfaction of the court, or is not prepared to answer these questions, judges might well be skeptical about the reliability of the evidence proffered.

Watch lists

With its ability to aggregate, sort, search, and analyze large quantities of data, AI has application to myriad national security contexts. The government might use AI, for example, to generate and maintain watch lists. AI watch-listing raises procedural due process issues under the Fifth Amendment. Depending on the inputs selected as well as what training and testing data are used, the application could also raise equal protection and First Amendment issues.

Even without AI as a factor, some courts have found due process violations in the nomination process for various government watch lists and in the government's redress process for individuals denied or delayed flight boarding.[146] Courts addressing watch-listing have applied the *Mathews v. Eldridge*[147] three-factor test to decide what process is constitutionally due, balancing:

1. the private interest that will be affected by the official action;
2. the risk of an erroneous deprivation of such interest through the procedures used, and the probable value, if any, of additional or substitute procedural safeguards; and

[146] *E.g.*, Ibrahim v. Dep't of Homeland Sec., 62 F. Supp. 3d 909 (N.D. Cal. 2014); *Latif v. Holder*, 28 F. Supp. 3d 1134 (D. Or. 2014); *but see* Elhady v. Kable, 993 F.3d 208 (4th Cir. 2021) (reversing district court finding of due process violation, where plaintiffs' travels were delayed but not precluded); Abdi v. Wray, 942 F.3d 1019 (10th Cir. 2019); Beydoun v. Sessions, 871 F.3d 459 (6th Cir. 2017). For a discussion of the government database at issue in *Elhady*, see Jeffrey Kahn, *Why a Judge's Terrorism Watchlist Ruling Is a Game Changer: What Happens Next*, Just Sec. (Sept. 9, 2019), https://www.justsecurity.org/66105/elhady-kable-what-happens-next-why-a-judges-terrorism-watchlist-ruling-is-a-game-changer/.

[147] 424 U.S. 319, 335 (1976).

3. the Government's interests, including the function involved and the fiscal and administrative burdens that additional or substitute procedural requirement[s] would entail.

The courts have considered, in greater nuance than presented here, the individual's right to travel and to be free from incarceration and from the stigma of being denied boarding or being watch-listed.[148] The courts have also considered the government's strong national security interests in watch-listing. Where courts have determined that an individual's liberty interest has been infringed, cases have turned on the second factor, the risk of erroneous error and the probable value of additional or substitute procedural safeguards.[149] No doubt adding AI to the equation will increase emphasis on that factor and the relative adjudicative transparency of applicable algorithms.

In *Ibrahim v. Department of Homeland Security* the government acknowledged that an FBI agent mistakenly nominated the plaintiff to the No-Fly list by marking a checklist form in exactly the opposite way it was intended.[150] The Northern District of California held that due process required a correction of "the error and all of its echoes" in all government records and "interlocking databases."[151] If an AI algorithm had nominated the plaintiff to the list, proving what factors the application had considered, erroneously or not, in the black box might be difficult or impossible.

But due process might require just that. In *Latif v. Holder*, the District of Oregon held due process required the government to provide the plaintiffs, who had been denied flight boarding, notice whether they were on the No-Fly list and the reasons for their placement on that list.[152] The notice had to be reasonably calculated to permit plaintiffs to submit evidence rebutting the government's reasons for their inclusion.[153] The executive, or a court reviewing the executive's actions, might consider whether an AI algorithm could document exactly what factors it considered in nominating a person to a watch list,

[148] *Compare Ibrahim*, 62 F. Supp. 3d at 928 *and Latif*, 28 F. Supp. 3d at 1148–51, *with Elhady*, 993 F.3d at 226–27, *Beydoun*, 871 F.3d at 469, *and Abdi*, 942 F.3d at 1033–34 (all determining plaintiffs could not establish the "plus" parts of their "stigma plus" claims because their placement on watch lists did not result in the denial or alteration of any previously held legal right).

[149] *See Latif*, 28 F. Supp. 3d at 1160–61 and *Ibrahim*, 62 F. Supp. at 929; *see also* Dycus, et al., *supra* note 96, at 26-6. *But see Elhady*, 993 F.3d at 228 (finding "the weight of the private interests at stake … comparatively weak" where plaintiffs' travels were only delayed).

[150] *Ibrahim*, 62 F. Supp. 3d at 928.

[151] *Id.* at 929.

[152] *Latif*, 28 F. Supp. 3d at 1162.

[153] *Id.* The *Latif* court left it to the government to fashion the appropriate procedures, but suggested it might provide unclassified summaries or share the classified reasons with cleared counsel.

and whether meaningful human review of the AI (or meaningful AI review of human nominations) was possible.

Latif, and more recently the district court in *Elhady v. Kable*,[154] noted the low standard—the executive's reasonable suspicion standard—for inclusion on the lists. In *Elhady*, which was later reversed, the Eastern District of Virginia determined that the central national database from which all other, shorter lists are derived, the Terrorist Screening Database (TSDB), posed due process issues. The court cited the vague and low standard for including an individual on the TSDB, noting the plaintiffs' assertions that the Terrorist Screening Center "may consider a wide range of factors in determining whether an individual belongs on the Watchlist, including an individual's 'race, ethnicity, or religious affiliation,' beliefs and activities protected by the First Amendment, travel history, personal and professional associations, and financial transactions."[155] Moreover, the court found "there is no independent review of a person's placement on the TSDB by a neutral decisionmaker," which "coupled with the limited disclosures and opportunity to respond by a person who requests that his status be reviewed," creates a substantial risk of erroneous deprivation.[156] (The Fourth Circuit, however, reversed the district court's finding of a due process violation in *Elhady*, because the plaintiffs' travels were delayed but not precluded.[157] The appellate court, citing similar holdings about additional security screenings and delays by the Tenth and Sixth Circuits, "declined the invitation to create a circuit split."[158])

Assuming a due process harm, the standard used in the watch list nomination process remains an issue. Depending on its quality, AI presents the possibility of either sharpening or blurring the standard and the factors considered in a nomination, whether by human or machine. AI might be used after the fact to check human nominations for errors or to sort through masses of data to bring certain records to human attention for human nomination. But, as in other scenarios, any AI will be subject to historical data sets (or learning sets for the AI) that may bias the AI's predictions of future behavior. Historical watch list data sets might, for example, include a disproportionate number of Muslim or Arab individuals, training the machine to factor religion or ethnicity into future nominations, even if neither is an intentional or at least explicit input. As ever, the programmers' own biases will inevitably be reflected in the

[154] *Elhady v. Kable*, 391 F. Supp. 3d 562 (E.D. Va. 2019).

[155] *Id.* at 582 (citing Pls.' Statement of Material Facts).

[156] *Id.*

[157] Elhady v. Kable, 993 F.3d 208 (4th Cir. 2021).

[158] *Id.* at 212, 222–23 (citing Abdi v. Wray, 942 F.3d 1019 (10th Cir. 2019); Beydoun v. Sessions, 871 F.3d 459 (6th Cir. 2017)).

algorithms in ways perhaps unknown to them, highlighting the need for judges to ask the right questions to test the accuracy, bias, and relevance of any AI application or output.

Other Fifth Amendment issues

The black box aspect of machine learning makes it especially susceptible to due process issues. The future may see litigation surrounding *any* government use of an AI application, for example, AI-driven DNA testing. Takings cases may present another Fifth Amendment context for AI court appearances, for example, if the federal government invokes the Invention Secrecy Act to prevent disclosure of private-sector AI inventions in the interest of national security.

This section has delved into such topics as liability for AI products and constitutional issues raised by AI-driven surveillance, criminal risk assessments, and government watch-listing. Those are a just a few areas where AI might foster litigation, however. We might also expect lawsuits about AI-empowered smart contracts, intellectual property rights to AI, AI allocated government budgets, and employment issues created by AI-tracking in the workplace, among other areas that we have not covered. We hope the illustrations we have provided will be useful in highlighting issues that will cross litigation genres.

Final Thoughts

As we hope is evident from the prior discussion, the multidisciplinary AI field presents a myriad of complex evidentiary challenges. The law rarely, if ever, keeps pace with technology. The legislative and appellate processes simply do not move at the same pace as technological change and could not if they tried. Moore's Law is faster than case law. Likewise, scholars and commentators are currently better at asking questions than they are at answering them. Artificial Intelligence itself is a fast-moving field encompassing a constellation of technologies.

Judges and lawyers do not need to be mathematicians or coders to understand AI and to wisely adjudicate the use of AI in courts or by courts. Judges need to define and understand their roles as evidentiary gatekeepers, constitutional guardians, and in some cases, potential AI consumers. Then they have to ask the right questions. That is what judges do.

As stated at the outset, we have not sought to provide legal judgments nor endorse the use of particular AI applications in context. AI, and its many applications, however, will present judges with many judicial determinations in the days and years ahead. With respect to the judicial use of AI:

- Judges might use or forgo AI algorithms when making bail, sentencing, and parole decisions and do so with or without first validating the underlying AI.
- Judges also might decide that where an AI application is used to inform or decide questions of liberty—bail, sentencing, and parole—only publicly provided and disclosed AI systems should be used, or only applications that are also transparent to the defendant should be used.
- Judges might examine whether an AI application incorporates biased data (training, validation, or testing) or design or produces biased outputs that favor or disfavor particular social groups, raising equal protection and due process concerns.

With respect to the introduction of AI-generated evidence, courts have even more choice ahead, at least until the applicable rules of evidence change, binding precedent is set, or legislative bodies define a judicial range of choice.

- A judge will also have to decide whether to accept statistical assertions alone in validating the use of an algorithm, such as false positive and false negative rates, or require in-person testimony from experts or software engineers before allowing a jury to rely on an AI output as evidence.

- Most judges, we would surmise, would want to ensure that not only the AI algorithm was apt for the purpose at hand, but also the data, factors, and weighting in the case at hand.
- Judges may also decide that the moving party behind AI evidence bears the burden for demonstrating not only its admissibility, but also its validity.

In the immediate future, perhaps the most important thing courts can do is ask careful and informed questions. Judges should also put their analysis and application of the answers on record as to whether, why, how, and subject to what evidentiary standards and determinations AI has been admitted into evidence or used by courts, allowing full and informed appellate review. We hope this guide removes some of the mystery around AI and helps judges continue to build a common law of AI.

Appendix A. Key Terms, Concepts, and Issues

Artificial intelligence (AI)

There is no agreed–upon or general definition of AI; however, one practical definition is that artificial intelligence is any machine that can "perform tasks that would otherwise require human intelligence."[159] The NSCAI expands on this idea noting, "AI is not a single piece of hardware or software, but rather, a constellation of technologies that gives a computer system the ability to solve problems and to perform tasks that would otherwise require human intelligence."[160] AI can be implemented in computers as **algorithms based on models** such as **artificial neural networks (ANN)** which are often iteratively designed through methods such as **machine learning (ML)** or **deep learning (DL)**. One statutory definition of AI is found in the National Defense Authorization Act of 2020, which states:

> The term "artificial intelligence" means a machine-based system that can, for a given set of human-defined objectives, make predictions, recommendations, or decisions influencing real or virtual environments. Artificial intelligence systems use machine and human-based inputs to
>
> - perceive real and virtual environments;
> - abstract such perceptions into models through analysis in an automated manner; and
> - use model inference to formulate options for information or action.[161]

Algorithm

"[A] step-by-step procedure for solving a problem or accomplishing some end."[162] A familiar example is a recipe, which details the steps needed to prepare a dish. In a computer, an algorithm is implemented in computer code and details the discrete steps and calculations a computer needs to implement to complete a task. An algorithm is the "engine" an AI uses to "think" and make predictions. In the field of AI, the term "algorithm" is often used synonymously with "computer program." A program, however, is a more specific

[159] Baker*, supra* note 14, at 21.

[160] NSCAI, *supra* note 1, at 8.

[161] National Artificial Intelligence Act of 2020 §5002(3), Pub. L. No. 116-283, 134 Stat. 4524, 15 U.S.C. §9401.

[162] Definition of Algorithm, https://www.merriam-webster.com/dictionary/algorithm (last visited May 20, 2021).

term, referring to an algorithm written in computer code and packaged for execution.

Black box

A term used to describe the often-mysterious nature of AI decision-making and the problem of AI explainability.[163] Most AI write their own algorithms through machine learning, that can result in complex code and decision-making processes indecipherable even to engineers. Such complexity limits human ability to understand how an AI makes decisions, and what factors, including biases, may have influenced those decisions. However, considerable research is underway by organizations such as NIST to enable more transparent neural networks, which may allow judges and lawyers to more fully understand the parameters and weights applied within.[164]

Narrow AI

"[T]he ability of computational machines to perform singular tasks at optimal levels, or near optimal levels, and usually better than, although sometimes just in different ways, than humans."[165] Under this umbrella falls many single- or limited-purpose AI technologies such as facial recognition algorithms, driverless cars, and drones, among others. These technologies are intelligent in one or a few domains, limiting their ability to handle complexity or tasks outside of their intended purpose. All AI currently in use falls in this category.

Artificial general intelligence (AGI)

In the future it is possible we will move past narrow AI and develop Artificial General Intelligence that does not have a narrow function and can serve multiple purposes. AGI can be conceived of as an AI system that equates the general purpose intelligence of the human brain.[166] AGI does not have a precise definition and the line that divides it from narrow AI is gray. Thus, the introduction of AGI will likely be a gradual process and it is unlikely there will be a precise "Sputnik moment" that introduces the age of AGI.

Superintelligence (SI)

AI philosophers also contemplate the emergence of Superintelligence, a stage of AI evolution marked as beyond human intelligence.[167] SI has sparked

[163] Ariel Bleicher, *Demystifying the Black Box That Is AI*, Sci. Am. (Aug. 9, 2017), https://www.scientificamerican.com/article/demystifying-the-black-box-that-is-ai/.

[164] NIST (June 16, 2022), https://www.nist.gov/artificial-intelligence/ai-fundamental-research-explainability.

[165] Baker, *supra* note 14, at 29.

[166] IBM Cloud Education, *Artificial Intelligence (AI),* IBM (June 3, 2020), https://www.ibm.com/cloud/learn/what-is-artificial-intelligence.

[167] *Id.*

widespread concern and its risks and benefits are unclear. Stephen Hawking highlighted this uncertainty in 2015 exclaiming that "AI may be the best thing to ever happen to humanity or the worst."[168] A malicious SI could cause incalculable harm and a beneficial SI could prove an invaluable tool. It must be noted that many engineers and government officials disdain consideration of Superintelligence as the stuff of science fiction as well as a distraction from the real and immediate challenges of narrow AI today.

Machine learning (ML)

A method of creating AI that relies on data, algorithms, and learned experience to refine algorithms and form intelligence.[169] The premise of machine learning is that "intelligence" is not innate but must be learned through experience. Machine-learning AI algorithms are "trained" by engineers who feed it mass amounts of data which it slowly learns to interpret and understand. In response to the data, the AI gradually tweaks its code to steadily improve its abilities. These tweaks add up over time, helping the AI create stronger predictions. Forms of machine learning include:

Supervised learning ("learning through instruction")

A form of machine learning where engineers specify a desired outcome and feed the AI algorithm curated and labeled data to guide AI towards that outcome.[170] For example, to teach a facial recognition AI to match names and faces, labeled facial data would be fed to its algorithm so it could learn which faces correspond to which names. This method is ideal for tasks with agreed-upon "correct" answers or decisions.

Unsupervised learning ("self-taught learning")

A form of machine learning where unstructured and uncurated data is fed to a machine-learning algorithm which finds trends, patterns, and relationships in that data.[171] This is useful for finding insights humans may have overlooked or cannot perceive. This method is ideal for applications without a firm "answer" and general data analysis.

Reinforcement learning ("learning by doing")

A form of machine learning where the algorithm learns through trial

[168] Alex Hern, *Stephen Hawking: AI Will Be "Either Best or Worst Thing" for Humanity*, Guardian (Oct. 19, 2016), http://www.theguardian.com/science/2016/oct/19/stephen-hawking-ai-best-or-worst-thing-for-humanity-cambridge.

[169] IBM Cloud Education, *Machine Learning,* IBM (July 15, 2020), https://www.ibm.com/cloud/learn/machine-learning.

[170] *Id.*

[171] *Id.*

and error.[172] Its learning is often guided by a goal—winning a game of chess for instance—and adjusts its parameters to better reflect trials that helped it to reach or come close to that goal. This method is useful for discovering optimal solutions in rule-based systems such as chess, chemistry, physics, traffic pattern analysis, and many others.

Parameter

When a machine learning algorithm is fully trained it uses an equation it has discovered to create its predictions. The variables in this equation are called parameters and each represents some aspect of the data the model was trained on.[173] For instance, if a model was trained to predict recidivism, a potential parameter the algorithm could use to predict recidivism might be number of past offenses. Each algorithm will have many, often hundreds, of parameters. The importance of a given parameter to the algorithm is determined by a **weight**. If a parameter is highly important, it will be highly weighted in determining the result; if it is unimportant its weight will be very low or even zero.

Deep learning

A machine-learning approach characterized by the use of a multi-layered Artificial Neural Networks. This approach[174] has exploded in popularity over the last decade and is the predominant form of machine-learning AI. Common applications use deep learning including many driverless cars and voice-recognition AI.

Artificial neural network (ANN)

The model (or "tool") used in deep-learning AI best defined as a computer system that works to achieve intelligence through a network structure that works to simulate the human brain.[175] An ANN analyzes data by passing it through multiple layers of **artificial neurons** which sift through and decipher the data. This layered network structure allows the system to analyze discrete data elements, draw connections between discovered data patterns, and ultimately derive meaning and form predictions. Neural networks can be **wide**, meaning each layer has large numbers of neurons, or **deep,** meaning data must pass through many layers of neurons before a final conclusion is drawn. Engineers determine the width and depth of the network based on their interpretation of the tools and structures a specific AI application needs for success.

[172] *Id.*

[173] Jason Brownlee, *What is the Difference Between a Parameter and a Hyperparameter?*, Machine Learning Mastery (July 26, 2017), https://machinelearningmastery.com/difference-between-a-parameter-and-a-hyperparameter/.

[174] IBM, *supra* note 169.

[175] IBM Cloud Education, *Neural Networks,* IBM (Aug. 17, 2020), https://www.ibm.com/cloud/learn/neural-networks.

Confidence score

Any expression of certainty in the predictive accuracy of an AI or ML application.[176] AI applications are imperfect and offer approximate results, decisions, or predictions that can be provided with a level of confidence. Few, if any, results an AI produce should be treated as a certainty. For example, the FBI facial identification software mentioned in the introduction is not designed or intended to match a single identity with a face. Rather it offers the user a range of potential matches based on potential pattern similarities or matches. The algorithm is reported to be accurate 86% of the time when the algorithm output offers the user at least fifty potential match pictures.[177] Put another way, the AI has 86% confidence that the match will be one of the fifty given matches.

Facial recognition

A prominent class of AI applications that can detect a face and analyze its features (or "biometrics") and even predict the identity of that face. These AI applications are notable for their common use in criminal justice and national security as a means of identifying suspects or threats. Facial recognition algorithms can also be used to surveil more generally. Facial recognition may also be used as a biological "password" to authenticate an individual's identity (for example, to unlock a smartphone).

Autonomous systems

AI-controlled machines and vehicles such as driverless cars and aerial drones that can operate and make decisions with little or no human control. Such systems already exist; however, in most cases stringent safety demands have forestalled widespread use. **Lethal autonomous weapons systems** (LAWS or simply AWS), or autonomous systems that can use deadly force, have received outsize legal, ethical, and political attention given widespread concerns about giving inhuman systems the power to take a human life.

Natural language processing (NLP)

AI algorithms designed to process, analyze, and recognize written or verbal human speech at human levels.[178] NLP has a wide variety of applications. Familiar applications include virtual assistants such as Amazon's Alexa or Apple's Siri. In national security and criminal justice, NLP can be used to analyze

[176] Jason Brownlee, *Confidence Intervals for Machine Learning,* Machine Learning Mastery (Jul. 26, 2017), https://machinelearningmastery.com/confidence-intervals-for-machine-learning/.

[177] GAO, *supra* note 4, at 14.

[178] IBM Cloud Education, *Natural Language Processing (NLP*), IBM (July 02, 2020), https://www.ibm.com/cloud/learn/natural-language-processing.

and understand language recordings and written information, drawing conclusions, insights, and patterns from that data. This offers a powerful intelligence and investigative tool. It can also be used to match a voice to an identity (much like facial recognition) and language translation.

Algorithmic bias

According to McKinsey, "[w]hile 'bias' can refer to any form of preference, fair or unfair" undesirable AI bias is bias that leads to "discrimination against certain individuals or groups of individuals based on the inappropriate use of certain traits or characteristics."[179] As AI is designed by humans, AI will always be biased and will assume the biases of its engineers potentially leading to poor or discriminatory results. As noted throughout this guide, causes of bias can stem from the statistical, such as errors in model design, to the social, to the use of inapt data to the context presented. Bias can also be caused by incomplete data sets. For instance, a facial recognition AI trained only on male faces may perform poorly when analyzing female faces.

Human in-the-loop

An autonomous AI system designed to work cooperatively with a human to complete its tasks. Often these AI defer to human judgment when making certain decisions, especially those with significant consequences or moral weight. Human in-the-loop systems generally seek a "best of both worlds" approach that maximizes the benefits of both human and AI decision making.

Human on-the-loop

An autonomous AI system designed to work under human oversight, allowing the human to easily intervene if the AI's decisions are in error, pose significant danger, or are ethically compromising.

Human out-of-the-loop

An autonomous AI system designed to operate without human oversight or involvement. Such systems do not facilitate easy human intervention if unethical or dangerous decisions are made.

[179] Silberg, Jake, *Notes from the AI frontier: Tackling bias in AI (and in humans),* Mckinsey Global Institute (Jun. 2019), https://www.mckinsey.com/featured-insights/artificial-intelligence/tackling-bias-in-artificial-intelligence-and-in-humans#.

Appendix B. Illustrative AI and AI-Related Cases

Federal cases

Bertuccelli v. Universal City Studios LLC, No. 19-1304, 2020 WL 6156821, at *1 (E.D. La. Oct. 21, 2020) (applying *Daubert* to allow expert testimony by an individual who "applied mathematical analysis using artificial intelligence and target algorithms to predict human response to similarity" between two artistic facial images in a copyright infringement case).

Force v. Facebook, 934 F.3d 53 (2d Cir. 2019) (holding that Section 230 of the Communications Decency Act barred civil terrorism claims against social networking website Facebook where plaintiffs argued Facebook should be liable for hosting content posted by Hamas members that allegedly inspired attacks on plaintiffs in Israel, and for using algorithms that directed such content to the personalized news feeds of the individuals who harmed the plaintiffs).

Gonzalez v. Google LLC, Nos. 18-16700, 18-17192, 19-15043, 2021 WL 2546675 (9th Cir. June 22, 2021) (affirming that most of plaintiffs' claims that Google, Twitter, and Facebook aided and abetted in acts of international terrorism were barred by Section 230 of the Communications Decency Act, but calling upon Congress to revisit Section 230 in light of advances in machine learning and other algorithms).

Henderson v. Stensberg, No. 18-CV-555-JDP, 2020 WL 1320820 (W.D. Wis. Mar. 20, 2020) (denying motion to dismiss plaintiff's equal protection claim of alleged racial bias in COMPAS risk assessment tool used in parole decisions; distinguishing the procedural due process claim about sentencing in *State v. Loomis* (noted below) from plaintiff's equal protection claim).

Henderson v. Stensberg, No. 18-CV-555-JDP, 2021 WL 1221249, at *6 (W.D. Wis. Mar. 26, 2021) (granting summary judgment for defendants on two grounds: "First, some research suggests that COMPAS has a disparate impact on Black offenders, but it does not directly support a claim of intentional race discrimination, which is what Henderson must show here. Second, and more important to this case, Henderson fails to present evidence showing that his COMPAS assessment worked against him in the parole hearing. Henderson's COMPAS recidivism score was the lowest possible, so he cannot show that his COMPAS recidivism score was the reason he was denied parole.")

*Houston Federation of Teachers, Local 2415 v. Houston Independent School Dist*rict, 251 F. Supp. 3d 1168 (S.D. Tex. 2017) (denying defendant's motion for summary judgment because use of privately developed algorithms to terminate public school teachers for ineffective performance may violate procedural due process, where teachers were denied access to computer algorithms and data necessary to meaningfully challenge terminations).

In re Ashley Madison Customer Data Security Breach Litigation, 148 F. Supp. 3d 1378 (J.P.M.L. 2015) (centralizing various class action claims against dating website for married persons, ashleymadison.com, for data security breach and for fraud based on the use of artificial intelligence "bots" and other mechanisms to mimic fake female users to induce actual, predominantly male users to make purchases).

In re Search of a Residence in Oakland, 354 F. Supp. 3d 1010 (N.D. Cal. 2019) (denying Government's request for search warrant to compel suspects of a crime to press a finger, or utilize other biometric features, to unlock digital devices, reasoning that such a search would violate the Fourth Amendment, because the Government lacked sufficient probable cause, and the Fifth Amendment, because the proposed used of biometric features would be testimonial).

In re Search of a White Google Pixel 3 XI Cellphone in a Black Incipo Case, 398 F. Supp. 3d 785 (D. Idaho 2019) (holding, contrary to *In re Search of a Residence in Oakland*, *supra*, that a requested warrant to compel the defendant to press a finger to unlock a cell phone did not violate the Fifth Amendment because it did not require defendant to provide any testimonial evidence).

Leaders of a Beautiful Struggle v. Baltimore Police Department., 2 F.4th 330 (4th Cir. 2021) (citing *Carpenter* in requiring a search warrant for a (now discontinued) aerial surveillance program operated by the Baltimore police, where contractors used data from planes equipped with high tech cameras to "track individuals and vehicles from a crime scene and extract information to assist BDP in the investigation of Target Crimes"; reports included analysis before and after the crime occurred but no real-time analysis).

Luokung Technology Corp. v. Department of Defense, 538 F.Supp.3d 174

(D.D.C. 2021) (granting a preliminary injunction enjoining the Department of Defense from enforcing its designation of Luokung as a Communist Chinese military company (CCMC) where Luokung is a technology company that makes navigation and mapping technology that is used in autonomous vehicles).

Patel v. Facebook, Inc., 932 F.3d 1264, 1273 (9th Cir. 2019) (determining that plaintiffs, who claimed Facebook's facial recognition and scanning technology violated the Illinois Biometric Information Privacy Act, alleged a concrete injury-in-fact for purposes of Article III standing: "the development of a face template using facial-recognition technology without consent (as alleged here) invades an individual's private affairs and concrete interests.").

Ross-Hime Designs, Inc. v. United States, 151 Fed. Cl. 241 (2020) (concluding the National Aeronautics and Space Administration's use of two robotic hand-like manipulators did not infringe on plaintiff's patents).

Stein v. Clarifai, Inc., 526 F. Supp. 3d 339, 345 (N.D. Ill. 2021) (granting Clarifai's motion to dismiss for lack of personal jurisdiction where Clarifai, a technology company, used photos from OKCupid to create a face database to develop and train algorithms in its facial recognition programs, because plaintiff did not demonstrate Clarifai directed its suit-related actions at Illinois).

Thornley v. Clearview AI, Inc., 984 F.3d 1241, 1249 (7th Cir. 2021) (holding that plaintiffs did not allege a concrete and particularized harm and therefore lacked Article III standing where they asserted a violation of the Illinois Biometric Privacy Act, where Clearview scraped photos from the internet to harvest biometric facial scans and associated metadata and offered access to its database to users to find information about someone in a photograph).

Vance v. Amazon.com Inc., No. C20-1084JLR, 2021 WL 1401633 (W.D. Wash. Apr. 14, 2021) (holding that users of photo-sharing website Flickr sufficiently pleaded claims against Amazon for violating the Illinois Biometric Information Privacy Act and for unjust enrichment, when Amazon used their biometric data in its facial recognition product sold to consumers and law enforcement).

Vance v. Microsoft Corporation, No. C20-1082JLR, 2021 WL 1401634 (W.D. Wash. Apr. 14, 2021) (dismissing, with leave to amend, claim by users of photo-sharing website Flickr against Microsoft for violating the Illinois

Biometric Information Privacy Act because their allegations did not establish that Microsoft disseminated or shared access to their biometric data through its products; further determining that plaintiffs did state a claim for unjust enrichment under applicable Illinois law).

United States v. Moore-Bush, 381 F. Supp. 3d 139 (D. Mass. 2019) (holding a warrantless video log of the defendants' travels in and out of their home over the course of eight months, created by a camera affixed to a utility pole that could also read the license plates of their guests, violated the Fourth Amendment).

United States v. Wilson, No. 18-50440, 2021 WL 4270847 (9th Cir. Sept. 21, 2021)(holding that the government violated Fourth Amendment rights of defendant where defendant uploaded images of apparent child pornography to his email, Google filed an automated report with the National Center for Missing and Exploited Children, based on an automated assessment of the images, and a government officer received the report and viewed the attachments).

WeRide Corp. v. Huang, 379 F. Supp. 3d 834 (N.D. Cal. 2019) (determining that "deep learning" source code for autonomous driving is a protectable trade secret).

United States v. Wright, 431 F. Supp. 3d 1175 (D. Nev. 2020) (finding, in agreement with *In re Search of a Residence in Oakland*, *supra*, that forcing defendant to unlock his cell phone using the facial recognition feature violated his Fifth Amendment right against self-incrimination).

State cases

Malenchik v. State, 928 N.E.2d 564, 565 (Ind. 2010) (affirming sentence where trial court's consideration of the defendant's assessment model scores was only supplemental to other sentencing evidence ("Legitimate offender assessment instruments do not replace but may inform a trial court's sentencing determinations ....")).

People v. Superior Court (Dominguez), 28 Cal. App. 5th 223 (Cal. Ct. App. 2018) (holding that the government could not be compelled to produce software program and an algorithm source code related to DNA testing because the research institute that developed the materials sought was not a member of the prosecution team; among other things, defendant's contention that software program rendered "machine testimony" did not adequately account for human input by lab analyst).

People v. Wakefield, 2019 N.Y. Slip Op 06143 (N.Y. App. Div. 2019) (holding that not having access to a software program's source code, which was used to identify the defendant's DNA on the victim's body and belongings, was not a violation of the Confrontation Clause; while the report generated by the program was "testimonial," the source code, even through the medium of a computer, could not be considered a declarant, where the program relied on human input and the program's creator testified at length in court ("This is not to say that an artificial intelligence-type system could never be a declarant, nor is there little doubt that the report and likelihood ratios at issue were derived through distributed cognition between technology and humans ....")).

State v. Guise, 921 N.W.2d 26 (Iowa 2018) (vacating court of appeals decision finding there is no legislative authority supporting the use of algorithmic risk assessment tools at sentencing and affirming judgment of district court because defendant failed to preserve due process claim on direct appeal and record was insufficient to reach the due process claim on direct appeal under the rubric of ineffective assistance of counsel).

State v. Headley, 926 N.W.2d 545 (Iowa 2019) (holding district court did not abuse its discretion in considering risk assessment tools at sentencing).

State v. Loomis, 881 N.W.2d 749 (Wis. 2016); *cert. denied* 137 S. Ct. 2290 (2017) (holding, *inter alia*, that the use of an algorithmic risk assessment tool as a nondeterminative factor in sentencing does not violate a defendant's due process right to be sentenced based on "accurate information," despite the defendant's limited ability to challenge the scientific validity of the risk assessment due to its proprietary nature).

Appendix C. Resources for Tracking AI-Related Legislation

Federal

Legislation related to artificial intelligence compiled by the Center for Data Innovation: https://datainnovation.org/ai-policy-leadership/ai-legislation-tracker/

State

State Legislation Related to Artificial Intelligence compiled by the National Conference of State Legislatures: https://www.ncsl.org/research/telecommunications-and-information-technology/2020-legislation-related-to-artificial-intelligence.aspx

State and local laws on facial recognition compiled by the Electronic Privacy Information Center: https://epic.org/state-policy/facialrecognition/

European

European Commission, "Press Release: Europe fit for the Digital Age: Commission proposes new rules and actions for excellence and trust in Artificial Intelligence," April 21, 2021, https://ec.europa.eu/commission/presscorner/detail/en/ip_21_1682 (provides overview and links to proposed regulation).

About the Authors

Judge James E. Baker is director of the Syracuse University Institute for Security Policy and Law and a professor at the Syracuse College of Law and the Maxwell School of Citizenship and Public Affairs. He previously served as a judge and chief judge on the U.S. Court of Appeals for the Armed Forces. As a career civil servant, Baker served as legal adviser and deputy legal adviser to the National Security Council. He has also served as counsel to the President's Foreign Intelligence Advisory Board and Intelligence Oversight Board, an attorney in the U.S. Department of State, an aide to Sen. Daniel Patrick Moynihan, and a Marine Corps infantry officer. In 2017–18, Baker was the Robert E. Wilhelm Fellow at the Center for International Studies, MIT. In addition to teaching at Syracuse University, Baker has taught at Yale, Iowa, Pittsburgh, Washington University (St. Louis), and Georgetown University. He is the author of numerous articles and three books: *The Centaur's Dilemma: National Security Law for the Coming AI Revolution* (Brookings, 2021), *In the Common Defense: National Security Law for Perilous Times* (Cambridge, 2007), and, with Michael Reisman, *Regulating Covert Action* (Yale, 1992).

Laurie Hobart is an associate teaching professor at Syracuse University College of Law, where she teaches national security law and related subjects. She has worked as an honors attorney and assistant general counsel within the intelligence community, serving in litigation, administrative law, and contract law divisions, and receiving a Harvard Heyman Fellowship for federal service. She has also worked as an associate at King & Spalding LLP, and as a law clerk to Hon. Charles F. Lettow of the U.S. Court of Federal Claims. She received a B.A., *summa cum laude*, from Cornell University's College of Arts & Sciences, a J.D. from Harvard Law School, and an M.F.A. in fiction writing from Syracuse University.

Matthew Mittelsteadt is an AI policy fellow for the Institute for Security Policy and Law (SPL) at the Syracuse University College of Law and research fellow at the Mercatus Center at George Mason University. His research focuses on issues of AI, arms control, and decentralized finance regulation. He holds an MS in Cybersecurity from New York University, an MPA from Syracuse University, and a BA in both Economics and Russian Studies from St. Olaf College.

Acknowledgments

The views expressed in this guide are our own, as are any errors. However, we would like to thank our research assistants over three years, Thomas Clifford, Shannon Cox, Thomas Finnigan III, Hannah Gabbard, Rickson Galvez, Alyssa Kozma, Margaret Santandreu, and Michael Stoianoff, for their hard work and good humor; Kristen Duda, for all her help in managing our team; John Cooke and the FJC for the opportunity to publish this book, and all of the FJC team, including Jason Cantone, Nathan Dotson, Meghan Dunn, José Idler, and Beth Wiggins, for their work on this project; Hon. Curtis Collier, who provided helpful feedback as part of the FJC review process; and reviewers and editors of a shorter version of this guide, published by the Georgetown Center for Security and Emerging Technology, Chuck Babington, Keith Bybee, Tobias Gibson, Danny Hague, Matt Mahoney, Hon. John Sparks, and Lynne Weil. Thank you all.

The Federal Judicial Center

Board

Director

John S. Cooke

Deputy Director

Clara J. Altman

About the Federal Judicial Center

The Federal Judicial Center is the research and education agency of the federal judicial system. It was established by Congress in 1967 (28 U.S.C. §§ 620–629) on the recommendation of the Judicial Conference of the United States.

By statute, the Chief Justice of the United States chairs the Center's Board, which also includes the director of the Administrative Office of the U.S. Courts and seven judges elected by the Judicial Conference.

The organization of the Center reflects its primary statutory mandates. The Education Division plans and produces education and training for judges and court staff, including in-person and virtual programs, videos and podcasts, publications, curriculum packages for in-district training, and web-based resources. The Research Division examines and evaluates current and alternative federal court practices and policies. This research assists Judicial Conference committees, who request most Center research, in developing policy recommendations. The Center's research also contributes substantially to its educational programs. The Federal Judicial History Office helps courts and others study and preserve federal judicial history. The International Judicial Relations Office provides information to judicial and legal officials from foreign countries and informs federal judicial personnel of developments in international law and other court systems that may affect their work. Two units of the Director's Office—the Information Technology Office and the Editorial & Information Services Office—support Center missions through technology, editorial and design assistance, and organization and dissemination of Center resources.

FJC
Federal Judicial Center

One Columbus Circle NE
Washington, DC 20002-8003
fjc.dcn • fjc.gov