ARTIFICIAL INTELLIGENCE

Legal Issues, Policy, and Practical Strategies

Cynthia H. Cwik, Christopher A. Suarez,
and Lucy L. Thomson
EDITORS

ABA
AMERICANBARASSOCIATION
Science & Technology
Law Section

Cover design by Sara Wadford/ABA Design

The materials contained herein represent the opinions of the authors and/or the editors and should not be construed to be the views or opinions of the law firms or companies with whom such persons are in partnership with, associated with, or employed by, nor of the American Bar Association or the Science & Technology Law Section, unless adopted pursuant to the bylaws of the Association.

Nothing contained in this book is to be considered as the rendering of legal advice for specific cases, and readers are responsible for obtaining such advice from their own legal counsel. This book is intended for educational and informational purposes only.

ISBN 978-1-63905-494-7 (epub)

Discounts are available for books ordered in bulk. Special consideration is given to state bars, CLE programs, and other bar-related organizations. Inquire at Book Publishing, ABA Publishing, American Bar Association, 321 N. Clark Street, Chicago, Illinois 60654-7598.

www.ShopABA.org

Chapter 1

Artificial Intelligence: A Primer for Legal Practitioners

Hany Farid

I. Introduction

The technology revolution—spanning from (approximately) 1950 to today—has seen several distinct stages. The initial personal computer revolution spanned some five decades, from the development of large, centralized computers in 1950 to 2000, when more than half of U.S. homes had a personal computer. The next wave of the technology revolution spanned 25 years, from 1989 with Tim Berners-Lee's conception of the underlying protocols powering today's World Wide Web to 2015, when approximately half of the world's population was online. The mobile revolution took a mere five years, from the introduction of the first Apple iPhone in 2007 to 2012, when the number of mobile phone users exceeded half of the world's population. In this most recent AI-powered wave of the technology revolution, OpenAI's ChatGPT went from zero to one billion users in only one year.

Fifty, twenty-five, five, one year(s): the technology revolution is accelerating with little sign of slowing.

The term "artificial intelligence" (AI) is not new. It dates to the 1950s, when computers weighed thousands of pounds and cost today's equivalent of hundreds of thousands of dollars. Through a series of boom-bust cycles, AI has recently emerged as a significant force in the technology revolution and

has become increasingly relevant to the legal profession. This chapter will briefly review AI's seven-decade history, describe two main branches of AI (predictive and generative), and examine the potential implications of AI to our society and justice system.

A. History of AI

In his seminal 1950 article, the great Alan Turing boldly asked if a machine could think.[1] Acknowledging that "think" and "machine" are difficult to precisely define, Turing proposed an alternative way to ask this question. What eventually became known as the Turing test has a human evaluator interacting through a text-based conversation with a human and a computer, the identities of which are concealed from the evaluator. If the evaluator cannot reliably distinguish the machine from the human, the machine passes the test, a proxy for achieving humanlike thinking. Turing boldly argued that machines will be able to—by this definition—think.

Building on Turing's seminal article, John McCarthy, then an assistant professor of mathematics at Dartmouth College, coined the term "artificial intelligence" and organized a small summer workshop in 1956 with the goal of considering "the conjecture that every aspect of learning or any other feature of intelligence can in principle be so precisely described that a machine can be made to simulate it."[2]

What would follow would be two decades of boom and bust for the field of AI with ambitious efforts falling spectacularly short of bold claims of what machines would soon be capable of. The 1980s saw a resurgence in AI thanks to advances in artificial neural networks (see Section II.B), but by the 1990s AI again fell out of favor, failing to deliver on its promises. To this point in its history, the field of AI was largely focused on knowledge-driven solutions in which, as McCarthy described it, machines would be directly imbued with rules and knowledge to simulate human intelligence. By the 1990s, however, the field of machine learning (ML) emerged from AI, in which a more data-driven approach was taken where a machine analyzes large amounts of relevant data from which it learns the necessary patterns and knowledge. That is, the knowledge is extracted from data as opposed to programmed directly.

Following another decade-long AI winter, two seminal events provided a glimmer of hope for AI and ML. IBM's Deep Blue in 1997 defeated world

champion Garry Kasparov in a six-game chess match, winning three games to Kasparov's two, with one draw (by comparison, just a year earlier, Kasparov defeated Deep Blue, 4–2). Because chess has clearly defined rules and well-defined objectives, it is perhaps not surprising that a machine prevailed given its massive amounts of computing power allowing it to explore and consider many more moves than humanly possible. However, when IBM's Watson in 2011 soundly defeated two of the all-time biggest champions on the TV quiz show *Jeopardy!*, the machine exhibited a stunning ability to reason in a less structured setting about human language and knowledge.

By 2015, the current AI revolution was well under way. This revolution is being powered by access to massive and powerful computing infrastructure in the form of cloud computing, significant mathematical and algorithmic breakthroughs in machine learning, and access to more than a decade's worth of data (text, image, video) digitized and uploaded for anyone—including machines—to consume and learn from.

B. Overview

Arguably today's use of the term "AI" is not faithful to Turing's and McCarthy's conception of machine intelligence. Most of what today is termed AI is more akin to data-driven ML. For expedience, however, I will continue to use the term "AI" to describe this broad field, which encompasses a range of different computational techniques.

I will describe two distinct branches within AI. The first, predictive AI, embodies a class of techniques for extracting patterns from data for the purpose of categorizing or characterizing data. This can range from diagnosing cancer from a CT scan to recognizing people in images and predicting who may default on a loan or recidivate if released on bail. Predictive AI is probably closest to what Turing and McCarthy envisioned—machines making humanlike decisions or analyses. The second, generative AI, embodies a class of techniques for creating content (text, audio, image, or video) that mimics the human content-creation process. Generative AI is particularly intriguing because of its ability to mimic not just human decision-making but human creativity, which was thought—until recently—to be well out of reach of machines.

II. Predictive AI

A predictive AI model takes as input diagnostic data and outputs a prediction. This prediction can be characterized as continuous or categorical. A continuous model may, for example, take as input the current unemployment rate, average temperature, and current number of citywide felonies and predict how many felonies will be committed in the next month. A categorical model may take as input an individual's employment status, marital status, and previous number of convictions and predict if they are at high risk of committing a felony in the next month. The former is continuous because the model's output is any numeric value, whereas the latter is categorical because the model's output is one of only two values, yes/no.

Building a continuous or categorical predictive-AI model typically follows three distinct steps involving inputs and outputs: (1) collect training data consisting of previous observations of the desired paired input/output relationship—this data is split into a training set and an evaluation set, each used in the next two steps; (2) train the computational model to learn the desired input/output relationship specified in the training set; and (3) evaluate the model to determine how well the model generalizes to previously unseen data in the evaluation set.

Evaluation of a continuous model is typically reported as mean absolute error (MAE). The MAE does not distinguish between over- and underpredicting. The mean signed difference (MSD) can be used alongside the MAE to reveal a potential bias in the model.

If, for example, a trained model predicts the number of monthly felonies for the first quarter of a year to be 150, 200, and 80 and the actual counts are 180, 170, and 120, then the MAE is $1/3 \times (|180 - 150| + |170 - 200| + |120 - 80|) = 33.3$, where $|\cdot|$ corresponds to absolute value. In other words, the average monthly prediction of the model is expected to be accurate to within plus or minus 33.3 felonies (or a 21 percent error relative to the average monthly felony rate).

The MSD for the previous example is $1/3 \times ((180 - 150) + (170 - 200) + (120 - 80)) = 13.3$, revealing a tendency of the model to underpredict the number of felonies (if this MSD was less than zero, then the model would be biased to overpredict).

Evaluation of a categorical model is typically reported in terms of overall accuracy and in terms of false positive and false negative rates. For predicting if an individual is likely to commit a felony in the next month, the overall accuracy specifies how many individuals in the dataset are correctly

classified. If we refer to the classification of high risk as a "positive," then the false positive rate specifies how many individuals were incorrectly classified as being high risk (i.e., they did not commit a felony), and the false negative rate specifies how many individuals were incorrectly classified as low risk (i.e., they did commit a felony). While overall accuracy is an important evaluation metric, these false positive and negative rates are critical, as they characterize the nature and consequence of the model's mistakes.

For both continuous and categorical models, it is important to report both training and evaluation accuracy as a measure of how well a trained model will generalize to new data. Importantly, however, because the training data and evaluation data are typically pulled from the same overall dataset, generalization on the evaluation data does not necessarily imply that the model will generalize once deployed in the wild. For example, if the model used to predict if a person will commit a felony in the next month was trained on only one narrow demographic, then it is likely to fail to generalize to different demographics. It is critical, therefore, that the assessment of a model's accuracy also be determined by the size and diversity of the training and evaluation datasets.

Predictive models range in computational complexity from techniques dating back to the early 1800s to more modern techniques. The next two sections describe the computational methodology underlying representative examples of these predictive models, along with general guidelines for assessing their reliability. The first class of regression techniques described next are not generally categorized as AI but form the basis for the description of artificial neural networks that form the basis of many AI-based predictive models.

A. Regression

1. Continuous

Consider a continuous model to predict the number of daily citywide shooting victims from the maximum daily temperature. Shown in Figure 1.1(a) is the relationship between these two variables collected over a random sampling of 50 days (i.e., each data point corresponds to a single day). We seek a model that captures the relationship in this data.

The simplest model is a linear model that relates the temperature (t) to the number of shootings (s) through a two-parameter equation of a line: $s = at + b$, where the model parameters a and b correspond to the slope and intercept of a line. The model parameters are estimated by finding the values of a and b that minimize the mean distance between the line and the data. With an appropriate distance metric, this estimation—termed linear regression—can be done using standard software and with no more computing power than found on any laptop.

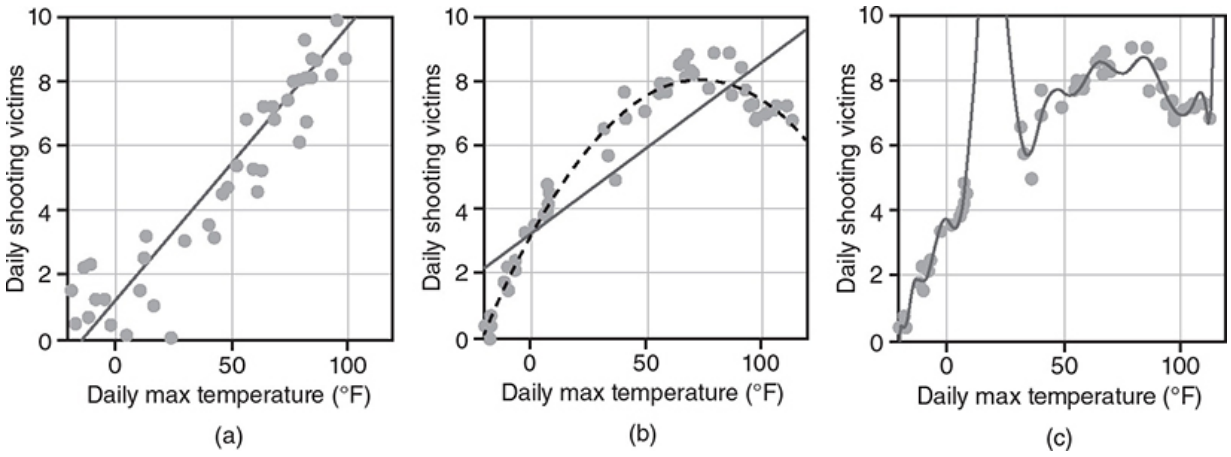

Figure 1.1: Continuous-valued linear regression in which the daily number of shooting victims is predicted from the daily maximum temperature using (a) a first-order linear model (gray line), (b) a first-order linear model (gray line) and second-order parabolic model (dashed curve), and (c) an overfitted 20th-order polynomial model (gray curve). These data are synthetically generated for the purpose of illustration and do not correspond to real statistics.

The gray line shown in Figure 1.1(a) is the resulting fitted model. With the model parameters (a, b) estimated, the number of shootings s on a given day can be predicted from the expected maximum temperature t as $s = at + b$.

Consider now the slightly more complex relationship between temperature and shootings shown in Figure 1.1(b), in which the number of shootings begins to decrease at around 75°F. A linear model (solid gray line) does a poor job of capturing the relationship, underestimating the number of shootings at around 50°F and overestimating around 100°F and beyond. This data can be better modeled with a higher-order model with three parameters (a parabolic model) of the form $s = at^2 + bt + c$. The same regression techniques used to estimate the linear model are used to estimate the parabolic model. Shown as a gray dashed curve is the estimated model that more accurately captures the underlying data.

Generally speaking, the addition of more model parameters improves the ability of the model to capture more complex patterns in the data. Care, however, must be taken to ensure that increasingly larger models do not overfit to the data. The data shown in Figure 1.1(c), for example, is fitted with a 20th-order polynomial (a line is a first-order polynomial, and a parabola is a second-order polynomial). Because of a dearth of data around 25°F, the model is highly unreliable in this region, as it has been overfitted to the remaining data (note that the simpler model in panel (b) does not overfit to the data in the same way). Notice also that the model struggles to extrapolate beyond the provided temperature range, with erratic predictions beyond 110°F.

In the previous examples, only one input (max daily temperature) was correlated against the desired predicted variable (daily shooting victims). Linear regression can be used to model the relationship between any number of inputs and the desired predicted variable.

Linear regression, dating back to the early 1800s, is extremely well understood both mathematically and computationally and has been widely deployed across the sciences and social sciences. It is, therefore, highly desirable in terms of its simplicity and explainability. The drawback of this method is that unlike more modern techniques (see Section II.B), linear regression may not be able to extract highly complex patterns in data.

2. Categorical

Consider now a categorical model to predict the likelihood that an individual, if released on bail, will commit a crime in the next six months based on their total number of prior convictions. Unlike the continuous model in the previous section that predicted a numeric value, this model's output is only one of two values: yes/no.

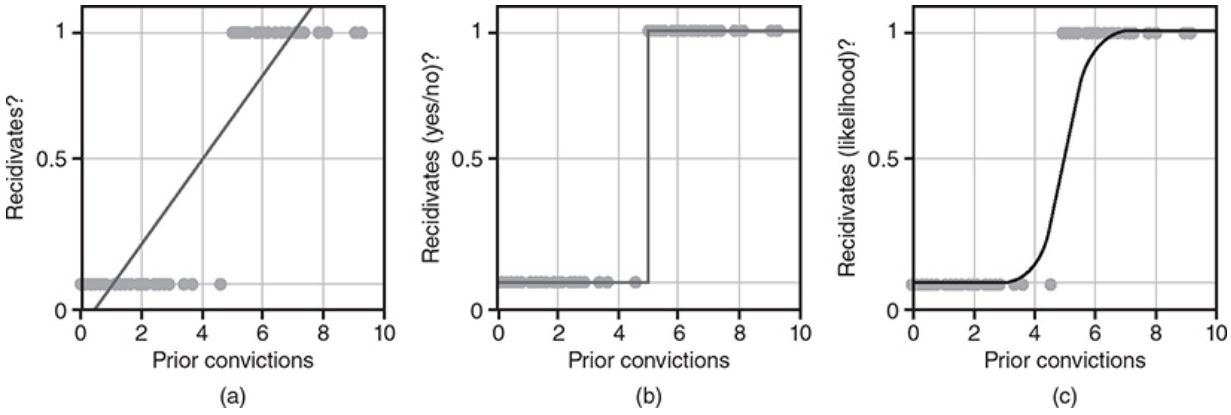

Figure 1.2: Categorical logistic regression in which the chance of recidivating after arrest is predicted from the number of prior convictions using (a) a first-order linear model, (b) a step function that forces the output of the linear model to be 0 or 1, and (c) a sigmoidal function that forces the output of the linear model to be bounded between 0 and 1, interpretable as a probability of recidivating. These data are synthetically generated for the purpose of illustration.

As before, we begin with a historical dataset in which defendants released on bail are tracked for six months to determine who does and does not recidivate and how this data correlates to the number of prior convictions.

Shown in Figure 1.2(a) is the relationship between these two variables collected for 50 individuals in which recidivism is encoded with a value of 1 and a lack of recidivism is encoded with a value of 0. We can, as described in the previous section, employ a linear model that relates number of prior convictions (p) to recidivism (r) through a two-parameter equation of a line: $r = ap + b$, where the model parameters a and b correspond to the slope and intercept of a line. This continuous model (gray line), however, generates arbitrary numeric values, whereas we seek a yes/no (1/0) from the categorical model.

By passing the output of this continuous model through a step function, Figure 1.2(b), we can force all values less than a threshold to be 0 and all values greater than this threshold to be 1. The model that combines a continuous linear model with this step function can be directly trained using well-understood techniques—termed logistic regression. The model is trained by finding the parameters that maximize the classification accuracy for individuals who do and do not recidivate. As with the previous linear regression, this estimation can be performed using standard software and with no more computing power than found on any laptop.

An alternative formulation passes the output of a continuous model, Figure 1.2(a), through a function with a gentler transition (termed a sigmoidal function), Figure 1.2(c). This approach has the benefit of forcing the model to generate—to the question of recidivism—a value bounded by 0 (no) and 1 (yes) with intermediate values between 0 and 1 corresponding to a probability of recidivating. With respect to the data shown in Figure 1.2(c), for example, a defendant with between zero and three prior convictions will confidently be predicted as unlikely to recidivate (the model's output is 0), a defendant with greater than seven prior convictions will confidently be predicted as likely to recidivate (the model's output is 1), and in between at five priors, the likelihood of recidivating is 50 percent (the model's output is 0.5). Although this model generates a range of numeric values, unlike the continuous model in Figure 1.2(a), these values are bounded between 0 and 1, and the numeric values can therefore be interpreted as a likelihood of the predicted variable.

In the previous example, only one input variable was correlated against the desired predicted variable. Logistic regression can be used to model the relationship between any number of inputs and the desired predicted variable.

This approach is categorized as supervised learning because it is explicitly trained on known input (e.g., convictions) and output (e.g., recidivism) data. By comparison, a class of techniques termed unsupervised learning begins with unlabeled data where the output variable is unknown and then automatically clusters the data into two (or more) categories to learn the relationship between the input variable and this learned categorization.

Like linear regression, logistic regression is extremely well understood both mathematically and computationally and has been widely deployed across the sciences and social sciences. It is, therefore, highly desirable in terms of its simplicity and explainability. The drawback of this method is that unlike more modern techniques (described next), logistic regression may not be able to extract highly complex patterns in data.

B. Artificial Neural Networks

The grayed portion of Figure 1.3 is a graphical depiction of a categorical model consisting of one output (y) and two inputs (x_1, x_2) of the form $y = w_1x_1 + w_2x_2 + w_3$ followed by a sigmoidal function that forces the output y to

be bounded between 0 and 1 (see also Figure 1.2). As described in Section II.A.2, the model parameters (w_1, w_2, and w_3) can be estimated using logistic regression.

The grayed portion of Figure 1.3 also represents a perceptron (the simplest form of an artificial neural network (ANN)). Conceived of in 1958, the perceptron[3] is a categorical model (see Section II.A.2) in which the relationship between the input and output remains the same, but the way in which the model parameters are estimated is different. Because the underlying model is the same as logistic regression, the perceptron can struggle to capture complex patterns in the training data.

However, two or more perceptrons combined as shown in Figure 1.3 have been found to have more inference power. In this multilayer perceptron, the inputs are first fed into two perceptrons with model parameters w_1, w_2, w_3 and w_4, w_5, w_6. The outputs of these perceptrons are then fed into a third perceptron with model parameters w_7, w_8 to yield the output z. The entire model consists of all eight of these parameters.

This combination of simpler models has proven to be surprisingly effective at learning complex patterns. Shown in Figure 1.3 is a tiny ANN by comparison to today's networks with millions to billions of parameters (the term "deep learning" refers to neural architectures with a large number of layers). As ANNs have grown in size (measured in terms of numbers of parameters) and novel architectures (how the inputs are combined and recombined), they have also grown in their ability to extract increasingly more complex patterns relating the inputs to the desired output(s). Today's ANNs, for example, are capable of seemingly human-level face recognition, interpreting diagnostic medical images, and traffic flow prediction. The downside of ANNs is that the estimation of the model parameters requires more complex algorithmic optimization, more computational power, and more training data.

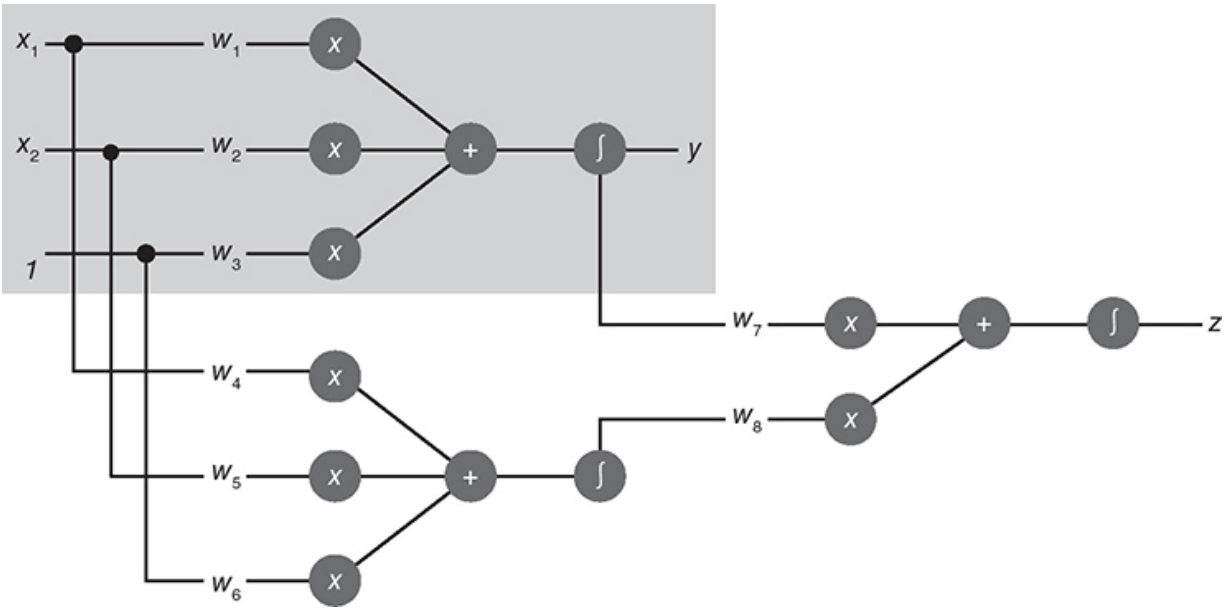

Figure 1.3: The grayed portion is a graphical depiction of a simple three-input, one-output perceptron of the form $y = w_1x_1 + w_2x_2 + w_3$. Three (or more) perceptrons can be combined to form a multilayer perceptron (also known as an artificial neural network (ANN)).

The general trend with all predictive models (from regression to ANNs) is that the larger the model, the more data is required to produce reliable models. As a very rough guideline, a typical rule of thumb calls for 10 to 100 data points for each model parameter. With billions of parameters, a typical ANN needs massive amounts of data to ensure that it does not overfit to its training data and thus will not be able to generalize when deployed (see, for example, Figure 1.1(c)). In addition, this training data needs to be representative of the data that the deployed model will eventually encounter. A common limitation of a trained ANN is that it struggles to generalize to data not represented in the training or validation datasets. For example, because white male faces have traditionally been overrepresented in large face datasets, ANN-powered face recognition has historically been less accurate when classifying women or people of color.[4]

As ANNs have grown in size, datasets have correspondingly grown. As a result, it has become increasingly more difficult to verify the integrity and appropriate representation of these massive datasets. This complexity can be partially alleviated using foundation models that are trained on large and diverse datasets so that they can be applied in a wide range of applications.[5]

Foundation models supported by large datasets have, therefore, become increasingly important in AI.

C. Case Study: Recidivism

Over the past two decades, automated risk assessment has become more common in the criminal justice system. A common use case asks whether someone with a criminal offense will recidivate at some point in the future. These tools rely on an individual's criminal history, personal background, and demographic information to make these risk predictions. One widely used criminal risk assessment tool, Correctional Offender Management Profiling for Alternative Sanctions (COMPAS, Northpointe), has been in use since 2000.

In May 2016, writing for ProPublica, Angwin et al.[6] analyzed the efficacy of COMPAS in the pretrial context on over 7,000 individuals arrested in Broward County, Florida, between 2013 and 2014. The analysis indicated that the predictions were unreliable and racially biased. The authors found that COMPAS's overall accuracy for white defendants is 67.0 percent, slightly higher than its accuracy of 63.8 percent for black defendants (and only somewhat better than chance performance of 50 percent on this particular dataset). The mistakes made by COMPAS, however, affected black and white defendants differently: black defendants who did not recidivate were incorrectly predicted to reoffend at a rate of 44.9 percent, nearly twice as high as their white counterparts at 23.5 percent; and white defendants who did recidivate were incorrectly predicted to not reoffend at a rate of 47.7 percent, nearly twice as high as their black counterparts at 28.0 percent. In other words, COMPAS appears to favor white defendants over black defendants by underpredicting recidivism for white defendants and overpredicting recidivism for black defendants.

What is peculiar about this racial asymmetry is that the predictive model does not consider the race of the defendant as input. Because the makers of COMPAS have not revealed the details of their predictive model, it is not immediately obvious why their model is racially biased. Indeed, a common criticism of COMPAS and many AI tools is that their algorithms are a "black box," and so there is limited ability to understand why or how they are biased.

Absent transparency, the only way to discern these biases is through reverse engineering. Effectively reverse engineering the COMPAS model has shown that this model reduces to classifying defendants at high risk of recidivism simply if they have a high rate of prior convictions (i.e., number of convictions relative to their age).[7] While conviction rate may be a reasonable indicator for risk of recidivism, we also know that asymmetries throughout the criminal justice system lead to higher policing, arrests, prosecutions, and convictions for black defendants.[8] As a result, the AI model latched onto prior convictions—a proxy for race—as an indicator for risk of recidivism probably because no other input data provided a better indicator.

Overall, these models correctly predict recidivism only 65 percent of the time and are significantly biased against black defendants. In addition, it has been shown that these model predictions are no more accurate or less biased than asking nonexperts to make the same prediction based on a limited amount of information about a defendant.[9] Here, and in general, the deployment of AI does not guarantee more objective, accurate, or fair decisions.

This case study provides several important lessons for the deployment of predictive AI in the criminal justice system and beyond:

1. Reporting only the overall accuracy of a predictive model can hide problematic bias. Bias is apparent when the false positives (incorrectly classifying defendants as high risk) and false negatives (incorrectly classifying defendants as low risk) are reported and correlated against different demographic groups.
2. Even if a predictive model is not provided with specific demographic information (age, gender, identity, race), demographics can creep into the model through proxies. It is unlikely, therefore, for AI models to be entirely blind to protected attributes such as race, gender, and disability.
3. As noted earlier, COMPAS is an example of what is termed a black-box model in which the details of how the model works are opaque and the explainability of the model is weak. The model does not, for example, explain why a particular defendant is rated as high or low risk. By contrast, the reverse engineering performed in the Dressel

and Farid analysis[10] yielded an explainable model in which the criterion for assessment was accessible. Explainability is important to understanding the fairness and accuracy of AI-powered decision-making.

4. Predictive AI will not necessarily create a utopian future. Instead, because AI models are trained on historic data, predictive AI is prone to repeat history.

III. Generative AI

Generative AI embodies a class of techniques for creating content (text, audio, image, or video) that mimics the human content-creation process.[11] This section describes how these techniques work and how (and if) AI-generated content can be distinguished from real content.

A. Language

Ask ChatGPT just about anything and you will receive a surprisingly coherent—if not always accurate—response. ChatGPT is a form of generative AI commonly referred to as a large language model (LLM).[12] The first, and perhaps most important, thing to understand about LLMs is that they build a response one word at a time. For example, when asked, "Harley-Davidson is . . . ," ChatGPT will compare this prompt to its internal representation of billions of webpages and documents scraped from the internet for a similar text fragment and add a likely next word: "a." This process is repeated adding "legendary," followed by "American," "motorcycle," and so on to produce a response like "Harley-Davidson is a legendary American motorcycle brand known for its distinctive design, deep cultural impact, and a loyal following among enthusiasts around the world."

At each step, ChatGPT computes a rank-ordered list of possible next words and selects a word that is likely, but not necessarily the most likely. It has been observed—although not fully understood—that by adding randomness to which top-ranked word is selected, ChatGPT produces more interesting results (this is why ChatGPT will produce different responses when repeatedly prompted with the same question).

What is particularly surprising about ChatGPT is that it produces coherent responses even though it is generating a response one word at a time

with no explicit knowledge of what is being asked: ChatGPT is, in a sense, a sophisticated autocomplete similar to your email or text messaging app autocompleting a word for you as you are typing a message.

A critical difference, however, between a simple autocomplete and an LLM is that the LLM represents text differently, allowing it to match text fragments based not on identical words but on a transformed representation of the words that appears to capture something akin to meaning. For example, the text fragments "Harley-Davidson is a legendary American motorcycle brand . . . ," "The legendary American motorcycle brand Harley-Davidson . . . ," and "Famed U.S.-based Harley motorcycles . . ." share a common representation in the LLM-transformed representation of text fragments.

This word-by-word autocomplete also explains why today's LLMs are prone to fabricating facts—they simply do not (yet) have a mechanism to fact-check their responses. If, for example, ChatGPT scraped a large number of documents that claim that Harley-Davidson is a Canadian company, then ChatGPT would happily regurgitate this falsehood.

ChatGPT, and LLMs in general, is powered by a large neural network (Section II.B) that is designed to handle language as input and output—a so-called transformer. The model is trained on a large corpus of text, mostly scraped from a variety of websites (in many cases in violation of a site's terms of service and/or copyright law—a conversation, no doubt, for another time). The model begins by transforming a text fragment into a numeric representation—a so-called embedding. The model output consists of approximately 50,000 numeric values, each corresponding to a word, and with each numeric output value corresponding to the likelihood that the corresponding word follows the input text fragment. The large corpus of training text is used to train the model weights. Once the model is trained, a text fragment (e.g., "Harley-Davidson is a legendary . . .") is provided to it as input, and the model then outputs a rank ordering of likely next words, settling on "American" as the next word in the sequence.

As previously discussed in Section II.B, the general trend has been that larger networks lead to more powerful AI models. Over four iterations, ChatGPT has grown in size from 1.5 billion model parameters in GPT-2 to 175 billion (GPT-3), 350 billion (GPT-3.5), and just over 1 trillion parameters (GPT-4). There is good reason to believe that future versions with even more parameters will exhibit increasingly more sophisticated behaviors.

Artificial general intelligence (AGI) corresponds to a machine being capable of performing any intellectual task performed by humans. It is unclear if the path to AGI (if it exists) passes through LLMs or if fundamentally different approaches to AI will be required to achieve true AGI. Regardless, today's LLMs exhibit hints of human-level thinking as conceived of by Turing and McCarthy.

B. Visual

Before the more respectable term "generative AI" took root, AI-generated audio/visual content was referred to as deepfakes, a term derived from the moniker of a Reddit user who in 2017 used the then-nascent AI-powered technology to create nonconsensual sexual imagery. This section will describe how deepfake images and videos are constructed and different interventions for detecting them.

1. Image

A generative adversarial network (GAN)[13] is a common computational technique for synthesizing images of people, cats, planes, or any other category: *generative* because these systems are tasked with generating an image; *adversarial* because these systems pit two separate components (the generator and the discriminator) against each other; and *network* because the computational machinery underlying the generator and discriminator is a neural network (see Section II.B).

StyleGAN[14] is one of the most successful systems for generating realistic human faces. When tasked with generating a face, the generator starts by laying down a random array of pixels and feeding this first guess to the discriminator. If the discriminator, equipped with a large database of real faces, can distinguish the generated image from the real faces, the discriminator provides this feedback to the generator. The generator then updates its initial guess and feeds this update to the discriminator in a second round. This process continues with the generator and discriminator competing in an adversarial game until an equilibrium is reached when the generator produces an image that the discriminator cannot distinguish from real faces.

The various versions of StyleGAN are open source, with a fully functioning version of StyleGAN2 available at https://thispersondoesnotexist.com, where each page reload yields a new synthesized face. Shown in the top row of Figure 1.4 are representative examples of StyleGAN2-generated faces.

A recent perceptual study[15] found that when asked to distinguish between a real and AI-generated face, participants performed no better than chance. In a second study in which participants were provided with training prior to completing the task, their performance improved only slightly. AI-generated faces are highly realistic and extremely difficult to perceptually distinguish from reality.

Although highly realistic, GANs do not afford much control over the appearance or surroundings of the synthesized face. By comparison, more recent text-to-image (or diffusion-based) synthesis affords more rendering control. Trained on billions of images with an accompanying descriptive caption, the model progressively corrupts each training image until only visual noise remains. The model then learns to denoise each image by reversing this corruption. This model can then be conditioned to generate an image that is semantically consistent with a text prompt like "a person on a university campus," as shown in the bottom row of Figure 1.4.

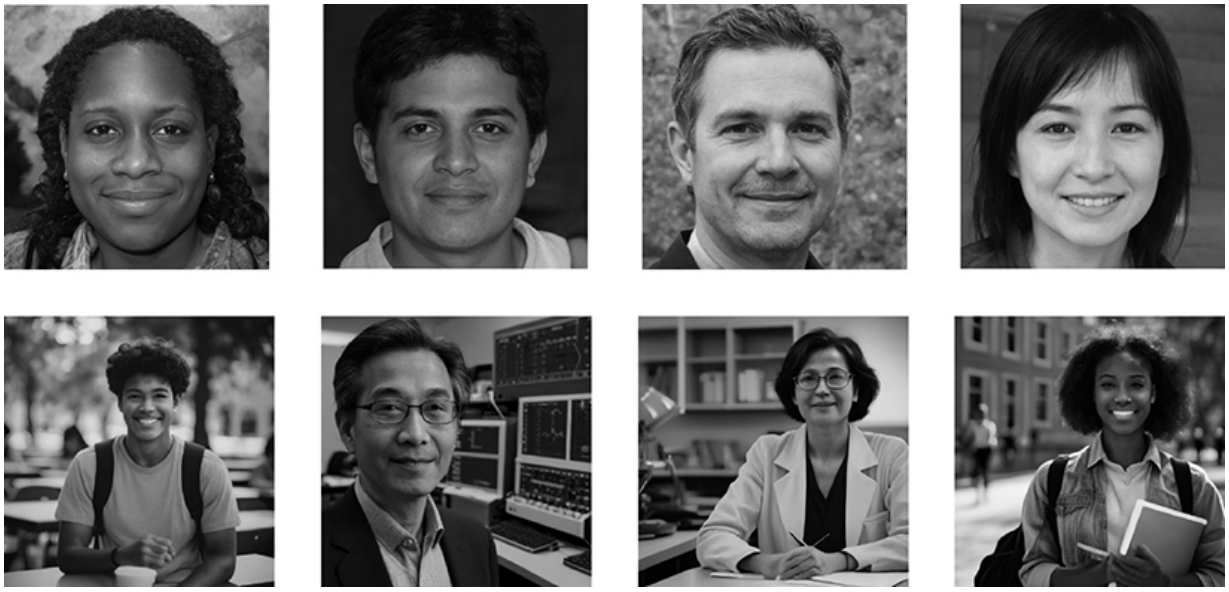

Figure 1.4: AI-generated images created by a generative adversarial network (top) and text-to-image synthesis (bottom) with the prompt "a person on a university campus."

2. Video

Although there are several different incarnations of video deepfakes, two of the most popular are so-called lip sync and face swap.

Given a reference video of a person talking and a new audio track (either AI-generated or impersonated), a lip-sync deepfake generates a new video track in four basic steps: (1) a neural network is trained to learn a mapping between an audio track and an outline of the mouth shape consistent with the audio; (2) a detailed image of the mouth region including nose, cheeks, mouth, and chin is synthesized; (3) the synthesized mouth region is blended onto a retimed reference video modified so that the head motion is consistent with the audio (e.g., the head is typically still when there is a pause in the speech); and (4) the jaw line is warped to match the shape and position of the chin. Using this technology, a video can be altered to make someone say things they never did.

A face-swap deepfake is a modified video in which one person's identity, from eyebrows to chin and cheek to cheek, is replaced with another identity. For each video frame of the original identity *A*, a video frame is synthesized where *A*'s face is swapped with a new identity *B*. The creation of this deepfake consists of three basic steps: (1) an image of the identity *B* is synthesized in the same head pose and expression as *A*; (2) any missing facial or hair pixels that arise from the synthesis step are filled in; and (3) the synthesized face *B* is blended into the original frame, replacing the identity of *A*. Repeating this process frame after frame yields a video in which one person's identity is swapped with another. This technique is at its best with access to many images of the co-opted identity *B* with different facial expressions and head poses but can also be performed from only a single image of *B*.

Several open-source implementations for both lip-sync and face-swap deepfakes are freely available online, in addition to numerous commercial implementations.

Today's deepfake videos animate a person from the neck up. We are, however, already seeing early examples of full-body deepfakes in which a person's entire body and movement can be manipulated. Although these videos currently have fairly obvious visual artifacts, this full-body puppeteering will soon be mastered, leading to even more realistic deepfakes. Adding to the realism, a person's voice can be cloned from only a

minute or two of prerecorded audio. This means that a person's face and voice can be fully co-opted.

Although computationally more demanding, face-swap deepfakes can also be created in real time, meaning that you will soon not know for sure if the person at the other end of a video call is real or not.

3. Guardrails

Some—but not all—commercial generative-AI services place semantic guardrails on user-generated prompts, making it difficult to create obviously abusive content that is sexually explicit or depicts gore. On the other hand, dozens of websites market services with the explicit purpose of inserting a woman's likeness into sexually explicit material. Dozens of websites allow for the creation of lip-sync deep-fakes already being used to perpetrate financial fraud and push disinformation campaigns. And several commercial voice cloning services require nothing more than checking a box asserting permission to use the voice in the uploaded audio.

While some regulatory or liability pressure might help rein in this Wild West of generative AI, the plethora of open-source models will be nearly impossible to control.

4. Real or Fake?

If past trends continue, it is reasonable to predict that all forms of generative AI will eventually be perceptually indistinguishable from reality. The field of digital forensics[16] develops computational techniques to detect manipulated media. There are two broad categories for detecting manipulated or AI-generated content: reactive and proactive.

Reactive techniques analyze various aspects of an image or video for traces of implausible or inconsistent properties. For example, shown in Figure 1.5 is a simple AI-generated image created with the text prompt "three boxes on a sidewalk on a sunny day." As expected for a sunny day, the generative AI created an image with cast shadows. The shape and position of these shadows can be useful in a forensic analysis.

The geometry of cast shadows is dictated by the 3D shape and location of an object and the illuminating light. This relationship is wonderfully simple: a point on an object, its corresponding shadow, and the light source

responsible for the shadow all lie on a single line. Because straight lines in the 3D scene are imaged to straight lines in the 2D image, this constraint holds in the image. Locate any point on a shadow and its corresponding point on the object and draw a line through them. Repeat for as many clearly defined shadow and object points as possible, and for an authentic image, all the lines will intersect at one point—the location of the illuminating light. As shown in Figure 1.5, these object-shadow constraints do intersect at a single point for the central box (solid lines) but not the flanking boxes (dashed lines), revealing a physically implausible scene in this AI-generated image.

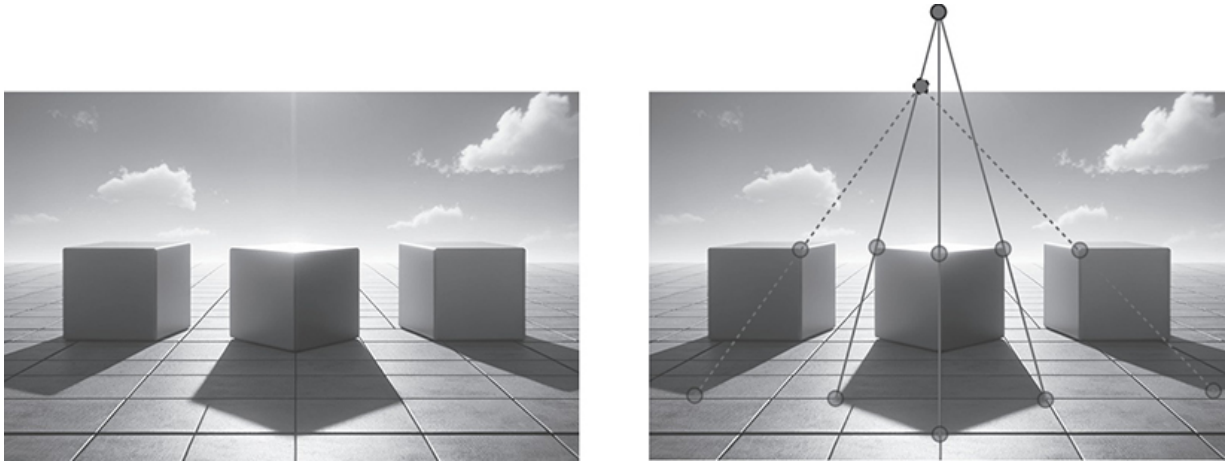

Figure 1.5: An AI-generated image and a forensic analysis of the physical plausibility of the shadows. The five shadow-object constraints do not intersect at a single point, revealing a physically implausible scene.

The benefit of these types of reactive forensic techniques is that they are applicable to a broad category of content. The drawback, however, is that they often require the assistance of an experienced analyst.

Proactive techniques, on the other hand, operate at the source of content creation, embedding into or extracting from an image or video an identifying digital watermark or signature.

A simple watermark can be added to a digital image by, for example, tweaking every tenth image pixel so that its color (typically encoded as a number in the range of 0 to 255) is even valued. Because this adjustment is so minor, the watermark is visually imperceptible. And, because this periodic pattern is unlikely to occur naturally, it can be used to verify an image's provenance.

The ideal watermark is one that is imperceptible and resilient to basic manipulations like cropping or resizing. Although the pixel manipulation just

described is not resilient to, for example, image resizing, many robust watermarking strategies have been proposed that are resilient—though not impervious—to attempts to remove them.

The benefit of watermarks is that identifying information is directly associated with a piece of content before it is released into the wild, making identification fast and easy. The drawback is that watermarks are vulnerable to attack, where an adversary can digitally remove the watermark while leaving the underlying content largely intact.

Therefore, in addition to embedding watermarks, a creator can extract an identifying fingerprint from the content and store it in a secure centralized ledger. This fingerprint is also referred to as a perceptual hash.[17] The provenance of a piece of content can then be determined by comparing the fingerprint of any image or video to the fingerprint stored in the ledger. Both watermarks and fingerprints can be made cryptographically secure, making them difficult to forge.

This type of watermarking and fingerprinting is equally effective for proactively tracking AI-generated and human-recorded content but will require integration into mobile devices and as many generative-AI services as possible (see https://contentauthenticity.org for such an industry-led effort).

Although not perfect, these combined reactive and proactive technologies will make it harder to create a compelling fake and easier to verify the integrity of real content. The creation and detection of manipulated media, however, are inherently adversarial and both sides will continually adapt, making distinguishing the real from the fake an ongoing challenge.

IV. Discussion

Writing a book chapter on technology is often fraught with concerns that the content will soon be outdated. This chapter is no exception, particularly at a moment when AI appears to be experiencing a real renaissance. We may soon see entirely new methodologies for predictive- and generative-AI methodologies as laid out in Sections II and III, and we may see entirely new applications of AI. However, the basic methodologies outlined here—linear regression, logistic regression, ANNs, GANs, and diffusion—are likely to persist in some form in the coming years. Even if these methodologies

change, the basic structure of AI techniques, and the cautionary notes, most likely will not.

Moving forward, and mostly independent of the specific underlying computational machinery powering AI, I contend that the legal profession should both embrace and be cautious of the use of AI. I next enumerate key takeaways regarding the use (or not) of AI in the legal profession:

1. Predictive AI can, in theory, be a powerful tool to remove bias in everything from policing to sentencing. As described in Section II.C, however, today's AI-powered tools are trained on historical data that itself contains various forms of bias. Any historical inequities will, therefore, be mirrored in the output of predictive AI. The deployment of predictive AI requires a careful examination of accuracy and potential bias against different demographic groups.
2. While ANNs can extract complex patterns in data (Section II.B), these mostly black-box predictive models offer little in terms of explainability. That is, it can be difficult to understand why an ANN predicts a certain outcome. Turning over life-altering decisions to opaque machines should be of concern to everyone. To this end, nascent efforts in explainable AI (xAI)[18] hold some promise, but significant breakthroughs are needed before we have a complete understanding of large ANNs and LLMs.
3. While some predictive AI models are open source or otherwise examinable, other commercial models are not. The number of bugs per 1,000 lines of code is estimated to be between 0.5 and 25, and it is estimated that even in widely scrutinized code, bugs can remain for years.[19] AI models that are not made open for scrutiny should, therefore, be of concern to the courts.
4. Today's incarnation of large language models like ChatGPT can easily make up facts, details, and citations (Section III.A). Several highly publicized cases have already exposed attorneys using an LLM to write court briefs littered with nonexistent citations. While it is likely that LLMs will eventually work out this hallucination problem, in the interim, the courts will need to establish guidelines for how and when LLMs can be used in court filings.

5. Generative AI will complicate the process of authenticating audio/visual evidence. While 20 years ago photo-editing tools like Photoshop made it possible for photographic evidence to be manipulated, skill and time were required to do so convincingly. Today, however, nearly anyone has access to powerful generative-AI tools to create and manipulate audio/visual content (Section III.B.1 and III.B.2). In addition, it is now easy for anyone to claim that audio/video evidence is fake—the so-called liar's dividend.[20] That is, the mere existence of deepfakes casts a cloud over all photographic evidence. We will, therefore, need to carefully rethink how the rules of evidence should be updated to contend with this new era of generative AI and deepfakes.
6. More broadly, AI is not a panacea for all that ails us. On the one hand, AI-powered face recognition can be quite accurate, surpassing humans in some situations.[21] On the other hand, AI is unlikely to quickly alleviate significant failings in the field of forensic science.[22] A recent study, for example, found that a state-of-the-art AI system for estimating a person's height and weight from a photograph is no more accurate than a visual assessment made by a nonexpert.[23] The archetypal Silicon Valley motto of "move fast and break things" should not be a model for the courts; we must innovate, move carefully, and not break things.

---

1. Alan M. Turing. Computing machinery and intelligence. *Mind*, 49:433–460, 1950.

2. John McCarthy, Marvin Minsky, Nathan Rochester & Claude Shannon, *A Proposal for the Dartmouth Summer Research Project on Artificial Intelligence* (1955), https://web.archive.org/web/20070826230310/http:/www-formal.stanford.edu/jmc/history/dartmouth/dartmouth.html.

3. Frank Rosenblatt. The perceptron: A probabilistic model for information storage and organization in the brain. *Psychological Review*, 65(6):386, 1958.

4. Joy Buolamwini and Timnit Gebru. Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Conference on Fairness, Accountability and Transparency*, pages 77–91. PMLR, 2018.

5. Rishi Bommasani, Drew A. Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S. Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, et al. On the opportunities and risks of foundation models. arXiv:2108.07258, 2021.

6. Surya Mattu, Julia Angwin, Jeff Larson, and Lauren Kirchner. Machine bias: There's software used across the country to predict future criminals. And it's biased against blacks. ProPublica, 2016.

7. Julia Dressel and Hany Farid. The accuracy, fairness, and limits of predicting recidivism. *Science Advances*, 4(1):eaao5580, 2018.

8. Ashley Nellis. The color of justice: Racial and ethnic disparity in state prisons. https://www.sentencingproject.org/reports/the-color-of-justice-racial-and-ethnic-disparity-in-state-prisons-the-sentencing-project/, 2016.

9. Dressel and Farid, *supra* note 7.

10. *Id*.

11. Hany Farid. Creating, using, misusing, and detecting deep fakes. *Journal of Online Trust and Safety*, 1(4), 2022.

12. Sébastien Bubeck, Varun Chandrasekaran, Ronen Eldan, Johannes Gehrke, Eric Horvitz, Ece Kamar, Peter Lee, Yin Tat Lee, Yuanzhi Li, Scott Lundberg, et al. Sparks of artificial general intelligence: Early experiments with GPT-4. arXiv:2303.12712, 2023.

13. Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Wade-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in Neural Information Processing Systems*, pages 2672–2680, 2014.

14. Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and improving the image quality of StyleGAN. In *International Conference on Computer Vision and Pattern Recognition*, pages 8110–8119, 2020.

15. Sophie Nightingale and Hany Farid. AI-synthesized faces are indistinguishable from real faces and more trustworthy. *Proceedings of the National Academy of Sciences*, 119(8), 2022.

16. Hany Farid. *Photo Forensics*. MIT Press, 2016.

17. Hany Farid. An overview of perceptual hashing. *Journal of Online Trust and Safety*, 1(1), 2021.

18. Robert R. Hoffman, Shane T. Mueller, Gary Klein, and Jordan Litman. Metrics for explainable AI: Challenges and prospects. arXiv:1812.04608, 2018.

19. Andrew Habib and Michael Pradel. How many of all bugs do we find? A study of static bug detectors. In *ACM/IEEE International Conference on Automated Software Engineering*, pages 317–328, 2018.

20. Bobby Chesney and Danielle Citron. Deep fakes: A looming challenge for privacy, democracy, and national security. *California Law Review*, 107:1753, 2019.

21. Alice J. O'Toole and Carlos D. Castillo. Face recognition by humans and machines: Three fundamental advances from deep learning. *Annual Review of Vision Science*, 7:543–570, 2021.

22. National Research Council Committee on Identifying the Needs of the Forensic Sciences Community. *Strengthening forensic science in the United States: A path forward*. National Academies Press, 2009; Harry T. Edwards. Ten years after the National Academy of Sciences' landmark report on strengthening forensic science in the United States: A path forward—where are we? Available at SSRN 3379373, 2019.

23. Sarah Barrington and Hany Farid. A comparative analysis of human and AI performance in forensic estimation of physical attributes. *Scientific Reports*, 13(1):4784, 2023.