Northwestern Journal of Technology and Intellectual Property

Volume 19 | Issue 1 Article 2

12-2021

Artificial Intelligence as Evidence

Paul W. Grimm

Maura R. Grossman

Gordon V. Cormack

N O R T H W E S T E R N
JOURNAL OF TECHNOLOGY
AND
INTELLECTUAL PROPERTY

ARTIFICIAL INTELLIGENCE AS EVIDENCE

Paul W. Grimm, Maura R. Grossman & Gordon V. Cormack

December 2021 VOL. 19, NO. 1

Volume 19, Number 1 (2021)
Northwestern Journal of Technology and Intellectual Property

ARTIFICIAL INTELLIGENCE AS EVIDENCE[1]

Paul W. Grimm, Maura R. Grossman & Gordon V. Cormack

ABSTRACT— This article explores issues that govern the admissibility of Artificial Intelligence ("AI") applications in civil and criminal cases, from the perspective of a federal trial judge and two computer scientists, one of whom also is an experienced attorney. It provides a detailed yet intelligible discussion of what AI is and how it works, a history of its development, and a description of the wide variety of functions that it is designed to accomplish, stressing that AI applications are ubiquitous, both in the private and public sectors. Applications today include: health care, education, employment-related decision-making, finance, law enforcement, and the legal profession. The article underscores the importance of determining the *validity* of an AI application (*i.e.*, how accurately the AI measures, classifies, or predicts what it is designed to), as well as its *reliability* (*i.e.*, the consistency with which the AI produces accurate results when applied to the same or substantially similar circumstances), in deciding whether it should be admitted into evidence in civil and criminal cases. The article further discusses factors that can affect the validity and reliability of AI evidence, including bias of various types, "function creep," lack of transparency and explainability, and the sufficiency of the objective testing of AI applications before they are released for public use. The article next provides an in-depth discussion of the evidentiary principles that govern whether AI evidence should be admitted in court cases, a topic which, at present, is not the subject of comprehensive analysis in decisional law. The focus of this discussion is on providing a step-by-step analysis of the most important issues, and the factors that affect decisions on whether to admit AI evidence. Finally, the article concludes with a discussion of practical suggestions intended to assist lawyers and judges as they are called upon to introduce, object to, or decide on whether to admit AI evidence.

---

[1] Hon. Paul W. Grimm is a United States District Judge for the District of Maryland, and an adjunct professor at both the University of Maryland Carey School of Law and the University of Baltimore School of Law. Maura R. Grossman, J.D., Ph.D., is a Research Professor, and Gordon V. Cormack, Ph.D., is a Professor, in the David R. Cheriton School of Computer Science at the University of Waterloo. Professor Grossman is also an affiliate faculty member at the Vector Institute for Artificial Intelligence. Her work is funded, in part, by the National Sciences and Engineering Council of Canada ("NESERC"). The opinions expressed in this article are the authors' own, and do not necessarily reflect the views of the institutions or organizations with which they are affiliated.

INTRODUCTION 10
I. WHAT IS "ARTIFICIAL INTELLIGENCE"? 14
II. WHY AI HAS COME TO THE FOREFRONT TODAY 17
III. THE AI TECHNOLOGY LANDSCAPE 24
IV. USES OF AI IN BUSINESS AND LAW TODAY 32
V. ISSUES RAISED BY THE USE OF AI IN BUSINESS AND LAW TODAY 41
A. Bias 42
B. Lack of Robust Testing for Validity and Reliability 48
C. Failure to Monitor for Function Creep 51
D. Failure to Ensure Data Privacy and Data Protection 53
E. Lack of Transparency and Explainabilty 60
F. Lack of Accountability 65
G. Lack of Resilience 72
VI. ESTABLISHING VALIDITY AND RELIABILITY 79
A. Testimony, Expert Testimony, or Technology? 79
B. Benchmarks and Goodhart's Law 82
VII. EVIDENTIARY PRINCIPLES THAT SHOULD BE CONSIDERED IN EVALUATING THE ADMISSIBILITY OF AI EVIDENCE IN CIVIL AND CRIMINAL TRIALS 84
A. Adequacy of the Federal Rules of Evidence in Addressing the Admissibility of AI Evidence 84
B. Relevance 86
C. Authentication of AI Evidence 90
D. Usefulness of the Daubert Factors in Determining Whether to Admit AI Evidence 95
E. Practice Pointers for Lawyers and Judges 97
CONCLUSION 105

INTRODUCTION

We live in an increasingly automated world. We use search engines to find much of the information we need for work and leisure, navigate our way to work using Waze or Google Maps, bank electronically without even the thought of entering an actual bank, instruct voice-activated personal assistants like Alexa or Siri to help us in countless ways, and socialize online without the inconvenience of having to actually be social. Soon, we hear, our cars will be driving themselves, and it is only a matter of time before airplanes will be able to fly themselves from one place to another without the need for human pilots.

Software applications, powered by seemingly omniscient and omnipotent "artificial intelligence" algorithms,[2] are used to diagnose and treat patients,[3] evaluate applicants for employment or promotion,[4] determine who is a good risk for a bank loan or credit card,[5] determine where police departments should deploy officers to most effectively prevent and respond to crime,[6] recognize faces in a photograph or video and match them to a real person,[7] forecast which offenders will recidivate,[8] and even predict an

[2] An algorithm is defined as "a procedure for solving a mathematical problem . . . in a finite number of steps that frequently involves repetition of an operation . . . [and more broadly as] a step-by-step procedure for solving a problem or accomplishing some end." *Algorithm*, MERRIAM-WEBSTER.COM DICTIONARY, https://www.merriam-webster.com/dictionary/algorithm [https://perma.cc/93SR-MGM7].

[3] *See, e.g.*, Jonathan G. Richens, Clarán M. Lee & Saurabh Johri, *Improving the Accuracy of Medical Diagnosis with Causal Machine Learning*, 11 NATURE COMMUNICATIONS Article No. 3921 (2020), https://www.nature.com/articles/s41467-020-17419-7 [https://perma.cc/VU5Y-PNZQ]; Thomas Davenport & Ravi Kalakota, *The Potential for Artificial Intelligence in Health Care*, 6 FUTURE HEALTH J. 94-98 (2019), https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6616181 [https://perma.cc/42CM-JFVN].

[4] *See, e.g.*, Kumba Sennaar, *Machine Learning for Recruiting and Hiring – 6 Current Applications*, EMERJ (last updated May 20, 2019), https://emerj.com/ai-sector-overviews/machine-learning-for-recruiting-and-hiring [https://perma.cc/RY7R-WBMH]; Ann Fisher, *An Algorithm May Decide Your Next Pay Raise*, FORTUNE (July 14, 2019), https://fortune.com/2019/07/14/artificial-intelligence-workplace-ibm-annual-review [https://perma.cc/2QSV-DMBF].

[5] *See, e.g.*, Dinesh Bacham & Janet Zhao, Machine Learning: Challenges, Lessons, and Opportunities in Credit Risk Modeling, IX Moody's Analytics Risk Perspectives | Managing Disruption (July 2017), https://www.moodysanalytics.com/risk-perspectives-magazine/managing-disruption/spotlight/machine-learning-challenges-lessons-and-opportunities-in-credit-risk-modeling [https://perma.cc/2537-C7RJ]; Rahul Shukla, *Prediction of Loan Approval with Machine Learning* (Sept. 19, 2020), https://medium.com/@rahulshuklawork/prediction-of-loan-approval-with-machine-learning-539cbd2aad31 [https://perma.cc/ZQ6H-H5MR] (last visited Nov. 15, 2021).

[6] *See, e.g.*, Steven L. Ostrowski, *How Machine Learning Can be a Force Multiplier for Public Safety*, POLICE1 BY LEXIPOL (Apr. 2, 2020), https://www.police1.com/police-products/police-technology/articles/how-machine-learning-can-be-a-force-multiplier-for-public-safety-30AaqNplj9Hq95ap [https://perma.cc/G378-KL3K]; Jonathan Chase et al., *Improving Law Enforcement Daily Deployment Through Machine Learning-Informed Optimization Under Uncertainty*, PROC. OF THE 28TH INT'L JOIN CONF. ON AI (IJCAI-19) 1-7 (2019), https://www.ijcai.org/proceedings/2019/0806.pdf [https://perma.cc/B3UV-FUVL].

[7] *See, e.g.*, Ewan, *What is Image Recognition?*, DEEPOMATIC (January 8, 2019), https://deepomatic.com/what-is-image-recognition [https://perma.cc/U68V-SDCD]; James Vincent, *FBI Used Facial Recognition to Identify Capitol Rioter From His Girlfriend's Instagram Posts*, THE VERGE (Apr. 21, 2021), https://www.theverge.com/2021/4/21/22395323/fbi-facial-recognition-us-capital-riots-tracked-down-suspect [https://perma.cc/R58L-5L3N].

[8] *See, e.g.*, Mirilla Zhu, *An Algorithmic Jury: Using Artificial Intelligence to Predict Recidivism Rates*, YALE SCIENTIFIC (May 15, 2020), https://www.yalescientific.org/2020/05/an-algorithmic-jury-using-artificial-intelligence-to-predict-recidivism-rates/ [https://perma.cc/CGA4-MZ9Q]; Mehdi Ghasemi et al., *The Application of Machine Learning to a General Risk-Need Assessment Instrument in the Prediction of Criminal Recidivism*, 48 CRIM. JUSTICE & BEHAVIOR 518–38 (Apr. 2020), https://journals.sagepub.com/doi/full/10.1177/0093854820969753 [https://perma.cc/ZYX9-VWTG];

attorney's chance of winning a lawsuit by analyzing data gathered about the presiding judge and opposing counsel.[9]

References to Artificial Intelligence are now so ubiquitous that we no longer need to use more than the abbreviation "AI" to understand what is meant. But there is something inscrutable about AI. We understand it to involve software programs powered by complicated mathematical rules called "algorithms," but most of us have never met anyone who has ever created a computer algorithm, or who can tell us how they actually work. We hear references to "machine learning," by which we understand that software applications are either entirely self-taught or trained—initially by humans—but eventually are able to teach themselves, and perform tasks far more complex than humans can, in but a fraction of the time.

However mysterious this may be to most of us, AI algorithms are no longer the stuff of science fiction or the imagination of high-tech brainiacs. They are being used right now, in countless software applications, and in increasingly expansive ways, in our personal undertakings, and by businesses and governments. For many AI applications, however, very little is known about the data they are fed, how they are developed and trained, or whether they produce consistently accurate results. And despite the generic phrase "artificial intelligence," this technology is hardly monolithic; there are many variants. Some AI applications are "trained" using supervised machine learning; others are self-taught through unsupervised machine learning, and there are still others that use reinforcement learning.[10] Some can be differentiated by what they are programmed to do, such as classifying or ranking data by its value or relationship to other data, versus others, which do regression analysis, by attaching specific values or weight to data in a large data set.

---

[9] *See, e.g.*, LEX MACHINA.COM, https://lexmachina.com [https://perma.cc/F43A-LJDM](AI tool to "[p]redict the behavior of courts, judges, lawyers, and parties with Legal Analytics"); Masha Medvedeva, Michael Vol & Martijn Wieling, *Using Machine Learning to Predict Decisions of the European Court of Human Rights*, 8 AI AND LAW 237–266 (2020), https://link.springer.com/article/10.1007/s10506-019-09255-y [https://perma.cc/YS87-JBLJ].

[10] In reinforcement learning, an AI system "learns to achieve a goal in an uncertain and potentially complex environment. The AI faces a game-like situation. [It] employs trial and error [methods] to come up with a solution to the problem. To get the machine to do what the programmer wants, the [AI system] gets either rewards or penalties for the actions it performs. Its goal is to maximize the total reward [and to minimize the total penalties]. Although the designer sets the reward policy—[in other words, devises] the rules of the game—[the designer] gives the model no hints or suggestions about how to solve the game. It's up to the model to figure out how to perform the task to maximize the reward, starting from totally random trials" and learn as it goes. *See* Błażej Osiński & Konrad Budek, *What Is Reinforcement Learning? The Complete Guide*, DEEPSENSE.AI (July 5, 2018), https://deepsense.ai/what-is-reinforcement-learning-the-complete-guide [https://perma.cc/3USA-7ZGV].

And if AI applications now dominate our lives, it is unavoidable that the evidence that will be needed to resolve civil litigation and criminal trials will include facts that are generated by this enigmatic technology. Whether they want to or not, lawyers seeking to introduce or object to AI evidence, and judges who must rule on its admissibility, need to have a working knowledge of what AI is and how it works, what it does accurately and reliably, and what it does not. Yet, there are few, if any, published court opinions that consider the issues regarding AI admissibility in any depth. And while there are many articles that raise concerns about privacy, bias in data or algorithms, lack of transparency, and the absence of accepted governance standards[11] with regard to AI evidence, there is a need for a practical (*i.e.*, not overly technical or esoteric) overview of both the technical and evidentiary issues implicated by AI evidence that is understandable to lay persons, lawyers, and judges alike, describing (i) what AI is, (ii) the factors that should be considered in evaluating its validity and reliability, and (iii) setting forth a systematic framework for addressing the evidentiary issues that must be considered when AI evidence is used in court. We have written this article from the perspective of two computer scientists (one of whom also is an experienced lawyer) and a trial judge. It is our hope that it will serve as a useful primer and prove helpful to lawyers and judges who must tackle the challenges associated with admissibility of AI evidence.

We begin by discussing what AI is and provide an overview of its origins. We discuss the different types of AI applications and the different functions they are designed to accomplish. Next, we illustrate the various ways in which AI technology is already in use today and some of the concerns about how it is deployed, including the frequent lack of transparency in how it was developed and tested. We explain how concerns about how programmatic bias and inaccurate assumptions may undermine or

[11] *See generally* Melissa Hamilton, *The Biased Algorithm: Evidence of Disparate Impact on Hispanics*, 56 AM. CRIM. L. REV. 1553 (2019); Patrick W. Nutter, Comment, *Machine Learning Evidence: Admissibility and Weight*, 21 U. PA J. CONST. L. 919 (2019); Jeff Ward, *10 Things Judges Should Know About AI*, 103 JUDICATURE 12 (Spring 2019); Andrea Roth, *Machine Testimony*, 126 YALE L.J. 1972 (2017); David Lehr & Paul Ohm, *Playing with the Data: What Legal Scholars Should Learn About Machine Learning*, 51 U.C. DAVIS L. REV. 653 (2017); Michael L. Rich, *Machine Learning, Automated Suspicion Algorithms, and the Fourth Amendment*, 164 U. PA. L. REV. 871 (2016); Harry Surden, *Machine Learning and Law*, 89 WASH. L. REV. 87 (2014); Pamela S. Katz, *Expert Robot: Using Artificial Intelligence to Assist Judges in Admitting Scientific Expert Testimony*, 24 ALB. L.J. SCI. & TECH. 1 (2014); John Nawara, *Machine Learning: Face Recognition Technology Evidence in Criminal Trials*, 49 U. LOUISVILLE L. REV. 601 (2011). It should be noted that one of the authors of this article (Judge Grimm) previewed some of the ideas and discussion found in this paper in two pieces published in early 2021: The Sedona Conference, *Commentary on ESI Evidence & Admissibility, Second Ed.*, 22 SEDONA CONF. J. 83, 183–90 & n.237 (2021), and Paul W. Grimm, *Practical Considerations for the Admissibility of Artificial Intelligence Evidenc*e, 2 MD. B.J. 39 (2021). Both pieces reference this article, which was already in draft form, as the original source for the ideas and discussion herein.

taint the appropriateness of its use. In the process, we stress the importance of two related concepts: *validity* (or accuracy in performance of the functions the technology was programmed to undertake), and *reliability* (the consistency with which the technology produces similar results when used in similar circumstances). Next, we discuss the evidentiary rules that must be considered in assessing the admissibility of AI evidence in court proceedings, and, finally, we conclude with some practical suggestions for lawyers and judges.

I. WHAT IS "ARTIFICIAL INTELLIGENCE"?

Artificial Intelligence is the hypothetical ability of a computer to match or exceed a human's performance in tasks requiring cognitive abilities, such as perception, language understanding and synthesis, reasoning, creativity, and emotion.[12] For some specific tasks, such as playing games like chess, Jeopardy, or Go, purpose-built computer systems have achieved performance rivaling or bettering the world's best experts, [13] while free or consumer-priced commodity chess-playing systems are at least as good as the average player.[14] For other tasks, such as voice or facial recognition and language translation, commonly deployed systems today are arguably as good as most people, and possibly better.[15] Complex tasks, such as driving an automobile or flying an airplane, can now—or will in the near future—be accomplished as well by computers as by licensed drivers or pilots.[16]

---

[12] *See* A.M. Turing, *I.—Computing Machinery and Intelligence*, 59 MIND 433, 460 (1950); John McCarthy et al., *A Proposal for the Dartmouth Summer Research Project on Artificial Intelligence*, August 31, 1955, reprinted in 27 AI MAG. 12 (2006).

[13] *See Deep Blue versus Gary Kasparov*, WIKIPEDIA, https://en.wikipedia.org/w/index.php?title=Deep_Blue_versus_Garry_Kasparov&oldid=990729889 [https://perma.cc/CA39-K92E]; Jo Best, *IBM Watson: The Inside Story of How the Jeopardy-Winning Supercomputer Was Born, and What It Wants to Do Next*, TECHREPUBLIC (Sept. 9, 2013), https://www.techrepublic.com/article/ibm-watson-the-inside-story-of-how-the-jeopardy-winning-supercomputer-was-born-and-what-it-wants-to-do-next [https://perma.cc/YQC6-FZSC]; *AlphaGo*, DEEPMIND, https://deepmind.com/research/case-studies/alphago-the-story-so-far [https://perma.cc/DER7-NC5L].

[14] *See, e.g.*, *Top 6 Best Chess Engines in the World in 2021*, ICHESS.NET (June 3, 2021), https://www.ichess.net/blog/best-chess-engines [https://perma.cc/BLG2-LZQU].

[15] *See* Norberto Andrade, *Computers Are Getting Better Than Humans at Facial Recognition*, THE ATLANTIC (June 9, 2014), https://www.theatlantic.com/technology/archive/2014/06/bad-news-computers-are-getting-better-than-we-are-at-facial-recognition/372377 [https://perma.cc/88L7-GAJH]; Vanessa Bates Ramirez, *A Computer Can Now Translate Languages as Well as a Human*, SINGULARITYHUB (Oct. 4, 2016), https://singularityhub.com/2016/10/04/a-computer-can-now-translate-languages-as-well-as-a-human [https://perma.cc/L2U3-356Z].

[16] *See, e.g.*, Chris Isidore, *Self-Driving Cars Are Already Really Safe*, CNN BUS. (Mar. 21, 2018, 12:07 PM ET), https://money.cnn.com/2018/03/21/technology/self-driving-car-safety/index.html [https://perma.cc/ZA7W-E72U]; Eric R. Teoh & David G. Kidd, *Rage Against the Machine? Google's Self-Driving Cars Versus Human Drivers*, 63 J. SAFETY RSCH. 57, 59 (2017); Aaron Pressman, *An F-16*

Computers can generate original music that is pleasant to the ear,[17] as well as artificial or altered images, videos, social media personas, and even news articles that humans have difficulty distinguishing from ones that are real.[18] Computers can also predict the near future; in many instances better than humans.[19] What computers cannot *yet* do is autonomously mine the energy and resources they need to feed themselves and to reproduce.[20]

The term "artificial intelligence" or "AI" refers to an aspirational goal (or the dystopian outcome) of exploring the limits of computation. The examples above of what computers can now do are generally referred to as "narrow" or "weak" AI, because they use purpose-built hardware and/or software systems that seek to emulate (or better) human performance at a single, well-defined task.[21] "General" or "strong" AI refers to a computer's ability to rival or exceed human performance at a full complement of cognitive tasks, including but not limited to, the ability to sustain itself (*i.e.*, the task of *go forth and multiply*).[22] At the time of this writing, the domain of

---

Pilot Took on A.I. in a Dogfight. Here's Who Won, FORTUNE (Aug. 20, 2020, 4:40 PM CDT), https://fortune.com/2020/08/20/f-16-fighter-pilot-versus-artificial-intelligence-simulation-darpa [https://perma.cc/LK6N-WLXD]; Arash Heydarian Pashakhanlou, *AI, Autonomy, and Airpower: The End of Pilots?*, 19 DEF. STUD. 337 (Oct. 12, 2019).

[17] Listen to some of the musical creations of AIVA at https://www.aiva.ai/creations [https://perma.cc/Y7FB-Y9VC].

[18] *See, e.g.*, Sophie J. Nightingale et al., *Can People Identify Original and Manipulated Photos of Real-World Scenes?*, 2 COGNITIVE RSCH. 30 (2017); Oscar Schwartz, *You Thought Fake News Was Bad? Deep Fakes Are Where Truth Goes to Die*, GUARDIAN (Nov. 12, 2018, 05.00 EST), https://www.theguardian.com/technology/2018/nov/12/deep-fakes-fake-news-truth [https://perma.cc/9KZY-EQY3]; Camila Domonoske, *Students Have 'Dismaying' Inability to Tell Fake News from Real, Study Finds*, NPR (Nov. 23, 2016, 2:44 PM ET), https://www.npr.org/sections/thetwo-way/2016/11/23/503129818/study-finds-students-have-dismaying-inability-to-tell-fake-news-from-real [https://perma.cc/GG5J-HEMN].

[19] *See* Berkeley J. Dietvorst et al., *Algorithm Aversion: People Erroneously Avoid Algorithms After Seeing Them Err*, 144 J. EXPER. PSYCH. 114 (2015), at 1 ("Research comparing the effectiveness of algorithmic and human forecasts shows that algorithms consistently outperform humans. In his book *Clinical Versus Statistical Prediction: A Theoretical Analysis and Review of the Evidence*, Paul Meehl (1954) reviewed results from 20 forecasting studies across diverse domains, including academic performance and parole violations, and showed that algorithms outperformed their human counterparts"; citing additional studies and meta-analyses and concluding that "across the vast majority of forecasting tasks, algorithmic forecasts are more accurate than human forecasts").

[20] *See* Kenneth Chang, *Can Robots Rule the World? Not Yet*, N.Y. TIMES (Sept. 12, 2000), https://www.nytimes.com/2000/09/12/science/can-robots-rule-the-world-not-yet.html [https://perma.cc/N6MP-S6XT]. *But see* Big Think, *AI Can Now Self-Reproduce—Should Humans Be Worried? | Eric Weinstein*, YOUTUBE (May 22, 2017), https://www.youtube.com/watch?v=Wu8s0tp9yzY [https://perma.cc/G6FY-KRHY].

[21] *See* Jake Frankenfield, *Weak AI*, INVESTOPEDIA (Feb. 25, 2021), https://www.investopedia.com/terms/w/weak-ai.asp [https://perma.cc/87LF-3RVD].

[22] *See Strong AI*, IBM Cloud Education (Aug. 31, 2020), https://www.ibm.com/cloud/learn/strong-ai [https://perma.cc/ZNQ4-RUTM]. Some futurists recognize a category of AI that exceeds strong AI, referred to as "artificial superintelligence" or "super AI," which "surpasses human intelligence and ability

tasks to which computers have been successfully applied—weak AI—along with their effectiveness at those tasks, has grown and continues to grow apace. Whether or when strong AI will be achieved in the future, and its possible consequences, is the subject of vigorous debate among experts,[23] a subject which is beyond the scope of this paper. Here, we are concerned with how the law should analyze and treat (i) the use of computers to perform or to assist in specific tasks that were heretofore the purview of human intellect, and (ii) the evidence derived from those computer systems.

As a term of art in computer science, "artificial intelligence" is an umbrella term for a number of research topics and underlying technologies aimed at furthering the application of computers to intellectual tasks, as well as the tasks themselves. It is not a single technology or function. "Rule-bases," "language models," and "machine learning" are common underlying technologies, while "chess playing," "question answering," and "automobile driving" are common applications. Various related applications are often considered together as fields of study, such as game playing, natural language processing ("NLP"),[24] computer vision,[25] information retrieval ("IR"), and robotics.

In common parlance, "artificial intelligence" is often little more than a synonym for either the latest, greatest technology, the technology of science fiction, or simply, a reference to a computer system that can somehow learn.

---

in all respects. . . . It's the best at everything – maths, science, medicine, hobbies, you name it. Even the brightest minds cannot come close to [its] abilities. . . ." *Types of AI: Distinguishing Weak, Strong, and Super AI*, THINKAUTOMATION, https://www.thinkautomation.com/bots-and-ai/types-of-ai-distinguishing-between-weak-strong-and-super-ai [https://perma.cc/S9TM-BZ8C]. At least for now, this type of AI remains in the realm of science fiction. *Id.* Nonetheless, for a dystopian view on what may be coming our way in the future, *see* Maureen Dowd, *A.I. Is Not A-OK*, NEW YORK TIMES (Oct. 30, 2021), https://www.nytimes.com/2021/10/30/opinion/eric-schmidt-ai.html [https://perma.cc/574T-74SQ].

[23] *See, e.g.*, Ragnar Fjelland, *Why General Artificial Intelligence Will Not Be Realized*, 7 HUMAN. & SOC. SCI. COMM. 10 (2020). *But see* VINCENT C. MÜLLER & NICK BOSTROM, *Future Progress in Artificial Intelligence: A Survey of Expert Opinion*, *in* FUNDAMENTAL ISSUES OF ARTIFICIAL INTELLIGENCE (Vincent C. Müller ed., Springer 2014). For an early take on this subject, *see* IRVING JOHN GOOD, *Speculations Concerning the First Ultraintelligent Machine*, in* 6 ADVANCES IN COMPUTER 31, 31–33 (1966).

[24] *See* Michael J. Garbade, *A Simple Introduction to Natural Language Processing*, BECOMING HUMAN: A.I. MAG. (Oct. 15, 2018), https://becominghuman.ai/a-simple-introduction-to-natural-language-processing-ea66a1747b32 [https://perma.cc/45GN-S9KB] ("Natural Language Processing, usually shortened as NLP, is a branch of [AI] that deals with the interaction between computers and humans using the natural language. The ultimate objective of NLP is to read, decipher, understand, and make sense of the human language. . . ."); *see also Natural Language Processing*, WIKIPEDIA, https://en.wikipedia.org/w/index.php?title=Natural_language_processing&oldid=1001740510 [https://perma.cc/ 8GJ6-WDEU].

[25] *See Computer Vision*, WIKIPEDIA, https://en.wikipedia.org/w/index.php?title=Computer_vision&oldid=1000754216 [https://perma.cc/VZH4-N2JN] ("Computer vision is an interdisciplinary scientific field that deals with how computers can gain high-level understanding from digital images or videos.").

Arguably, once an application of technology becomes well established, it becomes engineering,[26] rather than AI. For example, spam filters and computerized systems that can compare two documents and identify their differences were both once considered AI, but today are simply referred to as "software." This has led some commentators to conclude that AI is "whatever computers cannot do . . . until they can."[27] Thus, part of the challenge in defining AI is that its goal posts are constantly changing.

For our purpose, it is useful to outline the common technologies and tasks of AI, but not to be overly concerned with whether any particular technology—or any particular combination of technologies—constitutes AI, or merely reflects the products of engineering.

II. WHY AI HAS COME TO THE FOREFRONT TODAY

Although the term "artificial intelligence" appears to have been coined in 1956 by the organizers of the Dartmouth Summer Research Project on Artificial Intelligence,[28] the idea coincides with the invention of the modern computer. In 1948, Alan Turing, who had previously described mathematical problems that no computer could solve, wrote the manuscript "Intelligent Machinery,"[29] outlining the prospect that digital computers could "show intelligent behavior." In 1950, Turing proposed "The Imitation Game,"[30] now commonly known as the "Turing Test," to illustrate the question: "Can machines think?" The Imitation Game was somewhat more complicated than it is commonly paraphrased today. It involved three players: a woman ("A"), a man or a computer disguising itself as a woman ("B"), and a human interrogator of either sex ("C"), who could ask written questions and receive written answers from A and B, anonymized as X and Y. The interrogator would then guess which of X or Y was A, and which was B. If the computer

---

[26] Engineering is defined as "the application of science and mathematics by which the properties of matter and the sources of energy in nature are made useful to people [such as through] the design and manufacture of complex products." *Engineering*, MERRIAM-WEBSTER.COM DICTIONARY, https://www.merriam-webster.com/dictionary/engineering [https://perma.cc/V3FR-Q4ZT].

[27] Kathryn Hume, *Five Distractions in Thinking About AI*, QUAM PROXIME | AS NEAR AS MAY BE (Mar. 25, 2017), https://quamproxime.com/2017/03/25/five-distractions-in-thinking-about-ai [https://perma.cc/7Y2Q-TF6N]. *Cf. Artificial Intelligence is What We Can Do That Computers Can't . . . Yet*, SELFAWAREPATTERNS.COM (Feb. 27, 2014), https://selfawarepatterns.com/2014/02/27/artificial-intelligence-is-what-we-can-do-that-computers-cant-yet [https://perma.cc/7GA8-KHY2].

[28] *See* McCarthy et al., *supra* note 12.

[29] A.M. TURING, INTELLIGENT MACHINERY, NAT'L PHYSICAL LAB. (1948), *reprinted in* THE ESSENTIAL TURING: SEMINAL WRITINGS IN COMPUTING, LOGIC, PHILOSOPHY, ARTIFICIAL INTELLIGENCE, AND ARTIFICIAL LIFE: PLUS THE SECRETS OF ENIGMA 395–432 (B. Jack Copeland ed., 2004).

[30] Turing, *supra* note 12.

could fool the interrogator as often as the man, it could be said to display intelligent behavior.

Arguably, state-of-the-art technology today could be mustered to pass this test of weak AI, which would illustrate not only the computer's ability to emulate one human, but also to fool another. To be reasonably convincing, however, the test would need to be conducted according to a valid scientific protocol; most likely a randomized, controlled, double-blind trial.

In 1951, Claude Shannon, who had shared ideas with Turing since 1943, demonstrated a robotic mouse named Theseus that could find its way out of a maze, learning the layout of the maze in the process.[31] Theseus could remember the maze and find its way out a second time without making a wrong turn, but could also adapt its understanding if it discovered the maze had been changed. Theseus' logic was implemented by a large computer built from switching circuits, which communicated with the mouse using magnetic and electrical signals. Theseus illustrates many aspects of modern AI systems: perception, memory, problem solving, and active interaction with its environment.

Turing died tragically in 1954; Shannon, in collaboration with Marvin Minsky, John McCarthy, and Nathaniel Rochester organized the Dartmouth Project in 1956.[32] The Project identified several aspects of the "artificial intelligence problem," including the speed and memory capacities of computers, efficient and effective algorithms, programming a computer to use language, employing neural nets to represent concepts, abstraction from raw data, and harnessing randomness and creativity.[33]

The research community has made steady progress on these foundational technologies, as well as their application to particular narrow AI tasks. Arguably, we are just beginning to round the "Peak of Inflated Expectations,"[34] but this should not obscure the explosive progress that has

---

[31] *See* Robert G. Gallager, *Claude E. Shannon: A Retrospective on His Life, Work, and Impact*, 47 IEEE TRANSAC. ON INFO. THEORY 2681 (2001); *see also* Nokia Bell Labs Archives and the AT&T Archives and History Center, *Where Did Digital Communication Begin? Curated Highlights of "Theseus," Circa 1950s*, YOUTUBE (June 10, 2015), https://www.youtube.com/watch?v=nS0luYZd4fs [https://perma.cc/M47U-S6E7].

[32] *See* McCarthy et al., *supra* note 12.

[33] *See id.*

[34] The "Peak of Inflated Expectations" is the phase of the Gartner technology hype lifecycle where "[e]arly publicity produces a number of success stories—often accompanied by scores of failures. Some companies take action; many do not." *Gartner Hype Cycle*, GARTNER, https://www.gartner.com/en/research/methodologies/gartner-hype-cycle [https://perma.cc/Z2EA-JKTF]. *See also* Laurence Goasduff, *2 Megatrends Dominate the Gartner Hype Cycle for Artificial Intelligence, 2020* (Sept. 28, 2020), https://www.gartner.com/smarterwithgartner/2-megatrends-dominate-the-gartner-hype-cycle-for-artificial-intelligence-2020 [https://perma.cc/6RE8-QCB5] ("If AI as a general concept was positioned on this year's Gartner Hype Cycle, it would be rolling off the Peak of Inflated

been made and will continue to be made in this century, notwithstanding and throughout the ensuing "Trough of Disillusionment."[35]

Progress in AI can, in large part, be attributed to advances in the ability to gather and store vast amounts of raw data.[36] Where computers of Turing's and Shannon's day were severely limited by their memory capacity, today's computer systems are limited, not so much by their ability to gather or store data, but by their ability to make sense of it.[37] The transition from scarcity to glut has occasioned the use of machine-learning algorithms—both old and new—to achieve remarkable progress in many AI tasks.

The speed of computer processors has increased dramatically to the point that a typical processor at the turn of the century was about a million times faster than the processors available at the time of the Dartmouth Project.[38] Since that time, the speed of individual processors has plateaued due to the limitations of physics, and increased computational power has come by placing several processors ("cores") into a common device, or by connecting many discrete computer systems together in a communication network to form a cluster. Graphics processing units ("GPUs")[39] contain hundreds or thousands of cores; the clusters maintained by cloud service providers contain thousands of interconnected discrete computer systems. To harness the computing power afforded by multiple processors, algorithms

---

Expectations," meaning that "AI is starting to deliver on its potential and its benefits for businesses are becoming a reality.").

[35] The "Trough of Disillusionment" is the phase of the Gartner technology hype lifecycle where "[i]nterest wanes as [technological] experiments and implementations fail to deliver. Producers of the technology shake out or fail. Investments continue only if the surviving providers improve their products to the satisfaction of early adopters." *Gartner Hype Cycle*, GARTNER, https://www.gartner.com/en/research/methodologies/gartner-hype-cycle [https://perma.cc/Z2EA-JKTF].

[36] For example, the average consumer today carries more computing power in their pocket than that which landed a satellite on the moon. *See* Tibi Puiu, *Your Smartphone Is Millions of Times More Powerful than the Apollo 11 Guidance Computers*, ZME SCI. (May 13, 2021).

[37] *See* F.J. BURKOWSKI ET AL., A GLOBAL SEARCH ARCHITECTURE, Technical Report CS-95-12 (Dep't of Computer Sci., Univ. Waterloo, Mar. 15, 1995), https://cs.uwaterloo.ca/research/tr/1995/12/mt.pdf [https://perma.cc/EC7P-8VVJ].

[38] *See* Jonathan G. Koomey et al., *Implications of Historical Trends in the Electrical Efficiency of Computing*, 33 IEEE ANNALS OF THE HISTORY OF COMPUTING 46 (2011).

[39] A graphics processing unit ("GPU") is a "specialized, [programmable,] electronic circuit designed to rapidly . . . accelerate the creation [and rendering] of images" on a computer screen or other display device. "GPUs are used in embedded systems, mobile phones, personal computers, workstations, and game consoles. Modern GPUs are very efficient at manipulating computer graphics and image processing. Their highly parallel structure makes them more efficient than general-purpose central processing units (CPUs) for algorithms that process large blocks of data in parallel," as used for example, in the smoot decoding and rendering of 3D animations and video. The more sophisticated the GPU, the higher the resolution and the faster and smoother the motion. *See Graphics Processing Unit*, WIKIPEDIA, https://en.wikipedia.org/w/index.php?title=Graphics_processing_unit&oldid=1000546516 [https://perma.cc/KSE2-H4P7]; *GPU*, PCMAG ENCYCLOPEDIA, https://www.pcmag.com/encyclopedia/term/gpu [https://perma.cc/CQ8D-YHAC].

must split the problem up into pieces, each of which is solved by a separate processor. Specialized software tools like Google's TensorFlow[40] facilitate the implementation of machine-learning algorithms on GPUs, while tools like Apache Hadoop®[41] and Apache Spark™[42] facilitate the use of clusters.

The ready availability of commodity computers, Internet access, and open-source software has spawned a plethora of high-quality tools like TensorFlow, Hadoop, and Spark, as well as the Linux® operating system,[43] the Android mobile operating system,[44] and implementations of state-of-the-art learning algorithms like logistic regression, support vector machines ("SVM"), random forests, and artificial neural networks ("ANN"). Commercial enterprises like Google, Amazon, Microsoft, Oracle, Yandex, Baidu, and Huawei, as well as professionals, hobbyists, and hackers throughout the world are members of the open-source ecosystem, using and contributing to a global body of software, often stored in freely accessible repositories like Github.[45] This low barrier to entry allows almost anyone to build AI. Much, if not most commercial software relies, at least in part, on open-source software, even if it is not itself open-source.

Crowd-sourcing platforms, gamification, and instrumentation of search engines, application software, and "smart" appliances provide vast amounts of raw data for use as input to machine-learning systems. Perhaps the largest source is the Web itself, and other data sources, private and public, available through the Internet. Yet access to some data—including medical data, certain personal information (*e.g.*, bank records[46]), and government

---

[40] *See An End-To-End Open Source Machine Learning Platform*, TENSORFLOW, https://www.tensorflow.org [https://perma.cc/72ST-ZZRQ]; s*ee also* Martín Abadi, *TensorFlow: Learning Functions at Scale*, 51 PROC. OF THE 21ST ACM SIGPLAN INT'L CONF. ON FUNCTIONAL PROGRAMMING 1 (2016).

[41] *See* APACHE HADOOP, https://hadoop.apache.org [https://perma.cc/TU3S-AVM9]; s*ee also* Konstantin Shvachko et al., *The Hadoop Distributed File System*, 2010 IEEE SYMP. ON MASS STORAGE SYST. & TECH. 1 (2010), https://storageconference.us/2010/Papers/MSST/Shvachko.pdf [https://perma.cc/HG8F-GT65].

[42] *See Unified Engine for Large-Scale Data Analytics: What is Apache Spark™?*, APACHE HADOOP, https://spark.apache.org [https://perma.cc/AY24-E5P5]; *see also* MATEI ZAHARIA ET AL., SPARK: CLUSTER COMPUTING WITH WORKING SETS 1 (EECS Dep't, Univ. of Cal., Berkeley 2010), https://www.usenix.org/legacy/event/hotcloud10/tech/full_papers/Zaharia.pdf [https://perma.cc/NC7E-F92J].

[43] *Understanding Linux*, REDHAT (March 19, 2018), https://www.redhat.com/en/topics/linux [https://perma.cc/5QWU-BYBA].

[44] *See Introducing Android 11.*, ANDROID, https://www.android.com/intl/en_ca [https://perma.cc/TU9Y-QZPE]; *see also Android (Operating System)*, WIKIPEDIA, https://en.wikipedia.org/w/index.php?title=Android_(operating_system)&oldid=1001663336 [https://perma.cc/TDF3-LX45].

[45] *See* GITHUB, https://github.com [https://perma.cc/UZ69-C9CS].

[46] "Financial privacy laws regulate the manner in which financial institutions handle the nonpublic financial information of consumers. In the United States, financial privacy is regulated through laws

records—remains heavily restricted, especially to impartial observers. Corporations that collect data, particularly in the United States, are subject to less onerous restrictions than university researchers subject to ethics oversight; hackers who acquire or deduce unauthorized data are essentially unconstrained in their use of it for nefarious purposes.

Organized evaluation efforts with multiple participants have been instrumental in advancing the state of the art in AI. The National Institute of Technology's (NIST's) Text REtrieval Conference (TREC),[47] for example, poses annual information-retrieval tasks which are undertaken by academic and non-academic teams throughout the world. At TREC's inception in 1992, the challenge was to find relevant information in a corpus of one-half million documents, which was distributed on two compact discs.[48] At that time—the dawn of information abundance—that was the largest controlled evaluation of information-retrieval systems ever undertaken, by more than an order of magnitude.[49] 1992 also saw explosive growth of the World Wide Web, originally conceived in 1989,[50] followed a few years later by the first

enacted at the federal and state level. Federal regulations [include] the Bank Secrecy Act, Right to Financial Privacy Act, the Gramm-Leach-Bliley Act, and the Fair Credit Reporting Act. Provisions within other laws like the Credit and Debit Card Receipt Clarification Act of 2007, as well as the Electronic Funds Transfer Act also contribute to financial privacy in the United States. State regulations vary from state to state. While each state approaches financial privacy differently, they mostly draw from federal laws and provide more stringent outlines and definitions. Government agencies like the Consumer Financial Protection Bureau and the Federal Trade Commission provide enforcement for financial privacy regulations." *Financial Privacy Laws in the United States*, WIKIPEDIA, https://en.wikipedia.org/w/index.php?title=Financial_privacy_laws_in_the_United_States&oldid=994468039 [https://perma.cc/9JMD-UKPD].

[47] NIST was founded in 1901 and is a part of the United States Department of Commerce. Its website describes it as "one of the nation's oldest physical science laboratories. Congress established the agency to remove a major challenge to U.S. industrial competitiveness at the time—a second-rate measurement infrastructure that lagged behind the capabilities of the United Kingdom, Germany, and other economic rivals." *About NIST*, NIST, https://www.nist.gov/about-nist [https://perma.cc/3RU7-ASNX]. "The Text REtrieval Conference (TREC), co-sponsored by [NIST] and [the] U.S. Department of Defense, was started in 1992 as part of the TIPSTER Text program. Its purpose was to support research within the information retrieval community by providing the infrastructure necessary for large-scale evaluation of text retrieval methodologies." *Overview*, TREC, https://trec.nist.gov/overview.html [https://perma.cc/2Y9P-GQWL].

[48] *See* Donna Harman, *Overview of the First Text REtrieval Conference (TREC-1)*, NIST SPECIAL PUB. 500-207 1 (1993), https://trec.nist.gov/pubs/trec1/papers/01.txt [https://perma.cc/G4LS-YTJL]; Donna K. Harman, *Overview of the First TREC Conference*, PROC. OF THE 16TH ANN. INT'L ACM SIGIR CONF. ON RSCH. AND DEV. IN IR 36–47 (1993), https://dl.acm.org/doi/10.1145/160688.160692 [https://perma.cc/MC2Z-K4PQ].

[49] *See* Donna K. Harman, *The TREC Test Collections*, in TREC: EXPERIMENT and EVALUATION IN INFORMATION RETRIEVAL 21–52 (Ellen M. Voorhees and Donna K. Harman eds., MIT Press 2005).

[50] *A Short History of the Web*, CERN, https://home.cern/science/computing/birth-web/short-history-web [https://perma.cc/2P93-KKE6].

Web search engines, arguably influenced by TREC.[51] By 2009, TREC used a corpus of 500 million documents.[52] Meanwhile, Google eclipsed this total, announcing in 2008 that it could search 1,000 billion (*i.e.*, one trillion) documents.[53]

The annual TREC challenges continue to this day, focused on more sophisticated tasks rather than sheer volume. Notable tracks have included "Question Answering," which arguably spawned IBM's Watson;[54] "Legal,"[55] which demonstrated the efficacy of technology-assisted review ("TAR")[56] in electronic discovery; and "Total Recall,"[57] which demonstrated the efficacy of Continuous Active Learning® ("CAL®")[58] on sensitive clinical and

[51] *See The History of Search Engines*, WORDSTREAM, https://www.wordstream.com/articles/internet-search-engines-history [https://perma.cc/L8ZT-8ZC3]; *see generally* BRENT R. ROWE ET AL., ECONOMIC IMPACT ASSESSMENT OF NIST'S TEXT RETRIEVAL CONFERENCE (TREC) PROGRAM: FINAL REPORT, RTI PROJ. NO. 0211875 (2010), https://trec.nist.gov/pubs/2010.economic.impact.pdf [https://perma.cc/X8K6-RAXL].

[52] *See The ClueWeb09 Dataset*, THE LEMUR PROJECT, https://lemurproject.org/clueweb09 [https://perma.cc/8GYS-JB8F].

[53] Jesse Alpert & Nissan Hajaj, *We Knew the Web Was Big . . .* , GOOGLE OFFICIAL BLOG (July 25, 2008), https://googleblog.blogspot.com/2008/07/we-knew-web-was-big.html [https://perma.cc/CBX7-8KS9].

[54] *See Question Answering Track*, NAT. INST. STANDARDS TECH., https://trec.nist.gov/data/qamain.html [https://perma.cc/PGP2-JBJW]; *see also* John Prager, *The TREC Question Answering Track and IBM Watson*, Celebrating 25 Years of TREC, Webcast Part 3 at mins. 24:00 to 50:00, NAT. INST. STANDARDS TECH. (Nov. 18, 2016), https://www.nist.gov/news-events/events/2016/11/webcast-text-retrieval-conference [https://perma.cc/52HE-UZ8Q].

[55] *See About the Legal Track*, TREC LEGAL TRACK, https://trec-legal.umiacs.umd.edu [https://perma.cc/8K44-FGR4].

[56] "Technology-Assisted Review (TAR) [is a] process for Prioritizing or Coding a Collection of Documents using a computerized system that harnesses human judgments of one or more Subject Matter Expert(s) on a smaller set of Documents and then extrapolates those judgments to the remaining Document Collection. Some TAR methods use Machine Learning Algorithms to distinguish Relevant from Non-Relevant Documents, based on Training Examples Coded as Relevant or Non-Relevant by the Subject Matter Experts(s), while other TAR methods derive systematic Rules that emulate the expert(s)' decision-making process. TAR processes generally incorporate Statistical Models and/or Sampling techniques to guide the process and to measure overall system effectiveness." Maura R. Grossman & Gordon V. Cormack, *The Grossman-Cormack Glossary of Technology-Assisted Review*, 7 FED. CTS. L. REV. 1, 32 (2013).

[57] *See TREC 2016 Total Recall Track*, UNIV. OF WATERLOO (May 23, 2016), https://plg.uwaterloo.ca/~gvcormac/total-recall [https://perma.cc/GK6X-6YBC].

[58] Continuous Active Learning® and CAL® refer to a particular TAR protocol. *See* Maura R. Grossman & Gordon V. Cormack, *Continuous Active Learning for TAR*, PRAC. L.J. (2016). For a more technical discussion of CAL®, *see* Gordon V. Cormack & Maura R. Grossman, *Evaluation of Machine-Learning Protocols for Technology-Assisted Review in Electronic Discovery*, PROC. 37TH INT'L ACM SIGIR CONFR. ON RSCH. & DEV. INFO. RETRIEVAL, 153, 153–62 (2014), https://dl.acm.org/doi/10.1145/2600428.2609601 [https://perma.cc/3MAE-PVD4]. Continuous Active Learning® and CAL® are registered trademarks of Maura R. Grossman and Gordon V. Cormack. *See CONTINUOUS ACTIVE LEARNING – Trademark Details*, JUSTIA TRADEMARKS, https://trademarks.justia.com/866/34/continuous-active-86634255.html [https://perma.cc/N3Z4-4JM9];

government data. The datasets and evaluation tools for the various TREC tracks remain available for the purpose of evaluating new approaches as they are invented.[59]

TREC is but one of many evaluation forums. The Defense Advanced Research Projects Agency ("DARPA") Grand Challenge[60] kick-started progress in the development of autonomous vehicles. In 2004, no participant was able to complete the specified route.[61] By 2005, five teams completed the route, deploying an impressive array of innovative combinations of technology.[62]

The Knowledge Discovery and Data Mining competition ("KDD Cup") is an on-line competition that has run since 1997.[63] Since then, hundreds, if not thousands, of similar competitions have been held in which participants are given a task and submit their results or their software to a server that evaluates their submissions.[64] Netflix offered a $1M prize to participants who could build the best recommender system for movies.[65] Kaggle[66] runs a commercial platform that clients can use to host similar competitions.

The confluence of increased data capacity, processing power, the Internet, low barriers to entry, innovation, and community evaluation have undoubtedly spurred the progress of AI. So, too, has advertising had a significant influence on it. The primary impetus for the providers of search engine or social media platforms is to entice users to click on ads; a secondary goal may be to collect information about them, so as to use that information to entice users, along the way, to click on more ads, or to sell the

---

CAL – Trademark Details, JUSTIA TRADEMARKS, https://trademarks.justia.com/866/34/cal-86634265.html [https://perma.cc/TAR8-LTZV].

[59] TREC Research Collections Volumes 1–5 (English-language data) can be found at *Data – English Documents*, NIST, https://trec.nist.gov/data/docs_eng.html [https://perma.cc/AT76-UHS4]. Other collections are available through websites devoted to particular TREC Tracks. S*ee, e.g.*, *Data*, NIST, https://trec.nist.gov/data.html [https://perma.cc/R6DL-C8PZ]; s*ee generally* Harman, *supra* note 49.

[60] *The Grand Challenge*, DEFENSE ADVANCED RESEARCH PROJECTS AGENCY, https://www.darpa.mil/about-us/timeline/-grand-challenge-for-autonomous-vehicles [https://perma.cc/27HB-D2EY].

[61] *Id.*

[62] *See id.*

[63] *KDD Cup Archives*, KDD, https://www.kdd.org/kdd-cup [https://perma.cc/G87N-XWVX].

[64] *See, e.g.*, *Analytics, Data Science, Data Mining Competitions*, KDNUGGETS™, https://www.kdnuggets.com/competitions [https://perma.cc/6X8A-U45C]; Benedict Neo, *11 Data Science Competitions for You to Hone Your Skills for 2020*, TOWARDS DATA SCI. (Dec. 2, 2019), https://towardsdatascience.com/10-data-science-competitions-for-you-to-hone-your-skills-for-2020-32d87ee19cc9 [https://perma.cc/M6CQ-YM56]; Parul Pandey, *Top Competitive Data Science Platforms Other Than Kaggle*, TOWARDS DATA SCI. (Apr. 7, 2019), https://towardsdatascience.com/top-competitive-data-science-platforms-other-than-kaggle-2995e9dad93c [https://perma.cc/82YJ-NN5J].

[65] *Netflix Prize*, NETFLIX, https://www.netflixprize.com [https://perma.cc/2N44-UZW2].

[66] *Competitions*, KAGGLE, https://www.kaggle.com/competitions [https://perma.cc/6RFM-SA7F].

users' information to other enterprises wishing to get such users to purchase their wares, to vote for their candidate, to write a product review, to participate in an opinion poll, or to otherwise influence the users' behavior. Fulfilling the users' explicit needs is an incentive for the service provider only insofar as it furthers their own ends. Even Uber—the ride-sharing service—was developed, in part, to generate data to be used for the autonomous vehicles that the company was developing, as well as for other uses.[67]

III. THE AI TECHNOLOGY LANDSCAPE

Foundational AI technologies may be classified according to a set of abstract problems they are designed to solve, and the methods they employ to solve those problems. One of the most fundamental abstract problems is that of *classification*: determining whether a plant is edible or inedible, whether evidence is relevant or not, whether a potential juror will vote to convict or acquit, and so on. A related problem is one of *ranking*: ordering plants according to their food value, or evidence according to its weight, or jurors according to how likely they are to vote to convict. A third related problem is one of *regression*: rendering a quantitative estimate of a specific value, such as the caloric value of a plant, the probative value of a particular piece of evidence, or the probability that an individual juror will vote to convict. The solutions to all three problems can be used to summarize existing data and/or to predict future outcomes.

The problems of classification, ranking, and regression are commonly addressed by supervised machine-learning algorithms, such as Naïve Bayes, Nearest Neighbor, Perceptron, Random Forests, Logistic Regression, Support Vector Machines ("SVM"), and Artificial Neural Networks ("ANN"), including Convolutional Neural Networks ("CNN") and

[67] *See* Prableen Bajpai, *How Uber Uses Your Ride Data*, INVESTOPEDIA (Sept. 20, 2021), https://www.investopedia.com/articles/investing/030916/how-uber-uses-its-data-bank.asp [https://perma.cc/T3NU-MNWM]; Neil Patel, *How Uber Uses Data to Improve Their Service and Create the New Wave of Mobility*, NEILPATEL BLOG, https://neilpatel.com/blog/how-uber-uses-data [https://perma.cc/DM8H-AU9W]. The Uber example epitomizes the umbrella concept of "data monetization," *i.e.*, "the process of using data to obtain quantifiable economic benefit. Internal or indirect methods include using data to make measurable business performance improvements and inform decisions. External or direct methods include data sharing to gain beneficial terms or conditions from business partners, information bartering, selling data outright (via a data broker or independently), or offering information products and services (for example, including information as a value-added component of an existing offering)." *Data Monetization*, GARTNER, https://www.gartner.com/en/information-technology/glossary/data-monetization [https://perma.cc/6B8H-W4CP]; s*ee also Data Monetization*, WIKIPEDIA, https://en.wikipedia.org/w/index.php?title=Data_monetization&oldid=984813795 [https://perma.cc/K4BJ-GG5D].

Recurrent Neural Networks ("RNN"). The latter three algorithms are often referred to as "Deep Learning."[68]

A supervised machine-learning algorithm is trained to make distinctions in the same way that a child is taught to learn by showing the child examples, along with the correct answer: This is a cat, and that is a dog. Essentially, a supervised machine-learning algorithm infers mathematical functions from old, labeled data to make guesses about new, unlabeled data. So, for example, a classification algorithm might be shown examples of foods and poisons, of relevant and non-relevant evidence, or of jurors who have voted in the past to convict or to acquit. Based on this training, the learning algorithm builds a model, which is used to classify new examples for which it has not been given the correct answer. Many models, rather than yielding a categorical answer, in fact perform regression, estimating the likelihood, the probability, or a confidence score that the new example belongs to a particular category. This score is transformed into a categorical result by setting a threshold and deeming all examples above the threshold to be, for example, edible, and all others to be inedible. The scores can similarly be used for ranking or ordering a list of examples by their scores.

Some AI applications are fairly straightforward instances of the abstract problems of classification, ranking, and regression. A spam filter quarantines or deletes email that it classifies as inappropriate or malevolent. A Web Search engine ranks Web pages according to the likelihood they will satisfy the user's request, yielding a results page containing the 10-best hits, in order, from billions of potential candidates. Regression methods—some of which have been in use since before the invention of modern computers—can be used to estimate the probability of disease given certain risk factors, the maximum safe speed for maneuvering a vehicle over a particular terrain, the value of a particular property, or the grade to assign to an essay. Other AI applications, such as speech recognition, language translation, and autonomous vehicles, must address a complex web of interdependent AI problems.

Active learning and reinforcement learning are supervised machine-learning strategies in which the machine-learning algorithm selects its own training examples from which to best learn. In so doing, the algorithm must balance two objectives: *exploration*, in which it learns as much as it can, and

[68] Aravind Pai, *CNN vs. RNN vs. ANN – Analyzing 3 Types of Neural Networks in Deep Learning*, ANALYTICS VIDHYA (Feb. 17, 2020), https://www.analyticsvidhya.com/blog/2020/02/cnn-vs-rnn-vs-mlp-analyzing-3-types-of-neural-networks-in-deep-learning [https://perma.cc/V3FP-USDX]; *see* Abhishek Gupta, *Difference Between ANN, CNN and RNN*, GEEKSFORGEEKS (July 17, 2020), https://www.geeksforgeeks.org/difference-between-ann-cnn-and-rnn [https://perma.cc/Q78G-MKS2].

exploitation, in which it employs what it has learned thus far to address the problem at hand.[69]

Unsupervised machine-learning algorithms, in contrast, are not given the correct answer for any of their training examples. Instead, they look for patterns, groupings, or anomalies that might be of interest—either to the end user or as fodder for a supervised machine-learning algorithm.[70] The most common abstract problems to which unsupervised learning are applied are clustering and latent feature analysis. Clustering groups together things that the algorithm considers to be similar.[71] For example, given a deck of playing cards, it might consider the red cards to be one cluster, and the black to be another. Or it might consider the face cards to be one cluster, the numbered cards to be a second cluster, and the aces to be a third. Or it might consider spades, hearts, and clubs to be a cluster because their suit icons are curvy, and diamonds to be a separate cluster, because they are not. Clustering can be a useful aid in exploration of new or unknown data sets, either by a human or by a supervised machine-learning algorithm.

Feature analysis decomposes the input for classification, ranking, or regression systems into components ("features") for analysis by a supervised machine-learning algorithm.[72] In many cases, features are identified by a manual process known as feature engineering.[73] The features of a document written in English, for example, may be the words or phrases that it contains.

---

[69] *See, e.g.*, THOMAS OSUGI ET AL., BALANCING EXPLORATION AND EXPLOITATION: A NEW ALGORITHM FOR ACTIVE MACHINE LEARNING, (CSE Conference and Workshop Papers 2005), https://digitalcommons.unl.edu/cgi/viewcontent.cgi?article=1141&context=cseconfwork [https://perma.cc/628P-ZM9N].

[70] *See Unsupervised Learning*, WIKIPEDIA, https://en.wikipedia.org/w/index.php?title=Unsupervised_learning&oldid=1001697007 [https://perma.cc/4H8U-3PGV].

[71] *Cluster Analysis*, WIKIPEDIA, https://en.wikipedia.org/w/index.php?title=Cluster_analysis&oldid=1001573349 [https://perma.cc/6UZY-88CQ].

[72] "In machine learning . . . a feature is an individual measurable property or characteristic of a phenomenon being observed. Choosing informative, discriminating and independent features is a crucial step for effective algorithms in pattern recognition, classification and regression." *Feature (Machine Learning)*, WIKIPEDIA, https://en.wikipedia.org/w/index.php?title=Feature_(machine_learning)&oldid=993569874 [https://perma.cc/5TM9-FNAW].

[73] "Feature engineering is the process of using domain knowledge to extract features from raw data via data mining techniques. These features can be used to improve the performance of machine learning algorithms." *Feature Engineering*, WIKIPEDIA, https://en.wikipedia.org/w/index.php?title=Feature_engineering&oldid=996982436 [https://perma.cc/N2L2-GZTM]; *see also* Will Koehrsen, *Feature Engineering: What Powers Machine Learning - How to Extract Features from Raw Data for Machine Learning*, TOWARDS DATA SCI. (Nov. 12, 2018), https://towardsdatascience.com/feature-engineering-what-powers-machine-learning-93ab191bcc2d [https://perma.cc/3WKF-ZTB2].

But many words have similar underlying meanings, and unsupervised methods like latent semantic indexing ("LSI") or latent semantic analysis ("LSA"),[74] probabilistic latent semantic indexing ("PLSI") or probabilistic latent semantic analysis ("PLSA"),[75] and latent Dirichlet analysis ("LDA")[76] identify combinations of words that are used in similar contexts, under the theory that they are likely to represent similar concepts. For example, the terms "bat," "baseball," "pitcher," and "glove," might be grouped together to represent one concept, as might "bat," "Halloween," vampires," and "blood" to represent another. The words in the document are transformed into a list of concept weights, denoting the extent to which each is represented in the document. It is important to note that latent feature analysis does not itself do classification, ranking, or regression, but may be used to create features that are used as input to a supervised machine-learning algorithm that performs those tasks.

Feature engineering for English text is relatively easy because it can be split into words using simple lexical rules. But languages like Chinese, Japanese, and Korean have no lexical cues that split the text into "words." An even more challenging issue arises for images, as well as audio and video

[74] "Latent semantic analysis (LSA) is a technique in natural language processing . . . of analyzing relationships between a set of documents and the terms they contain by producing a set of concepts related to the documents and terms. . . . In the context of its application to information retrieval, it is sometimes called latent semantic indexing (LSI)." *Latent Semantic Analysis*, WIKIPEDIA, https://en.wikipedia.org/w/index.php?title=Latent_semantic_analysis&oldid=1001352831 [https://perma.cc/K768-GTAQ]. For a more technical discussion of LSI/LSA, see Susan T. Dumais, *Latent Semantic Analysis*, 38 ANN. REV. INFO. SCI. & TECH 188 (2005); Scott Deerwester et al., *Indexing by Latent Semantic Analysis*, 41 J. AM. SOC. INFO. SCI. 391 (1990).

[75] "Probabilistic latent semantic analysis (PLSA), also known as probabilistic latent semantic indexing (PLSI, especially in information retrieval circles) is a statistical technique for the analysis of two-mode and co-occurrence data. In effect, one can derive a low-dimensional representation of the observed variables in terms of their affinity to certain hidden variables, just as in latent semantic analysis, from which PLSA evolved." *Probabilistic Latent Semantic Analysis*, WIKIPEDIA, https://en.wikipedia.org/w/index.php?title=Probabilistic_latent_semantic_analysis&oldid=993310631 [https://perma.cc/6C5G-JD4R]. For a more technical discussion of PLSA/PLSI, see Thomas Hofmann, *Probabilistic Latent Semantic Indexing*, PROC. 22ND ANN. INT'L ACM SIGIR CONF. ON RSCH. & DEV. IN IR, 50–57 (1999), http://cis.csuohio.edu/~sschung/CIS660/PLSIHoffman.pdf [https://perma.cc/A49W-Q6X8].

[76] "In natural language processing, the latent Dirichlet allocation (LDA) is a generative statistical model that allows sets of observations to be explained by unobserved groups that explain why some parts of the data are similar. For example, if observations are words collected into documents, it posits that each document is a mixture of a small number of topics and that each word's presence is attributable to one of the document's topics. LDA is an example of a topic model and belongs to the machine learning toolbox and in wider sense to the artificial intelligence toolbox." *Latent Dirichlet Allocation*, WIKIPEDIA, https://en.wikipedia.org/w/index.php?title=Latent_Dirichlet_allocation&oldid=1000686922 [https://perma.cc/VVK5-SHWM]. For a more technical discussion of LDA, see generally David M. Blei et al., *Latent Dirichlet Allocation*, 3 J. MACH. LEARNING. RSCH. 993 (2003).

recordings, where there are no easily identified features that can be used for classification, ranking, or regression.

"Deep Learning" refers to the combination of two or more learning algorithms which, in combination, perform feature analysis as well as an abstract task such as classification, ranking, or regression.[77] Typically, these algorithms are implemented as multi-layered neural networks, which, given enough training examples, can perform remarkably well. The first layer takes raw data as input, and combines it in various ways, passing the result on to another layer, and so on. Each layer analyzes different features and adjusts its model in response to training data so as to improve the overall effectiveness. Eventually, the combined models yield superior results for the task at hand.[78]

Deep Learning has led to breakthroughs in speech and image recognition, as well as fact-based question answering.[79] What these problems have in common is the availability of a vast number of training examples from which to derive models.

While machine learning represents the current state of the art for the three abstract AI problems outlined above, other approaches have been used—and continue to be used and promoted—as AI. In particular, a *rule base* is simply a set of rules or patterns designed to specify the outcome for all possible inputs.[80] A rule base may take the form of a decision tree or a flowchart working through the possibilities in a systematic fashion. The possible outcomes at each level are enumerated by subject-matter experts ("SMEs") in collaboration with rule-base experts, often statisticians or linguists. A rule base may take the form of a number of "patterns" designed—again by SMEs in collaboration with technical experts—to recognize the features that distinguish one class from another.[81] Flowcharts or patterns may be augmented with scores—again manually determined—that may be used for ranking or regression.[82]

---

[77] For a more technical discussion of deep learning, see Yann LeCun et al., *Deep Learning*, 521 NATURE 436, 436–42 (2015).

[78] *See* LeCun et al., *supra* note 77.

[79] *See id*. *See also* Yashvardhan Sharma & Sahil Gupta, *Deep Learning Approaches for Question Answering System*, 132 PROCEDIA COMPUT. SCI. 785, 786 (2018).

[80] *See* Frederick Hayes-Roth, *Rule-Based Systems*, 28 COMMC'N OF THE ACM 921, 921–22 (1985). *See generally* Randall Davis & Jonathan J. King, *The Origin of Rule-Based Systems in AI*, *reprinted in* RULE-BASED EXPERT SYSTEMS: THE MYCIN EXPERIMENTS OF THE STANFORD HEURISTIC PROGRAMMING PROJECT (Addison-Wesley Pub. Co. 1984), http://digilib.stmik-banjarbaru.ac.id/data.bc/2.%20AI/2.%20AI/1984%20Rule-Based%20Expert%20Systems.pdf [https://perma.cc/AP2A-SM94].

[81] *See* Hayes-Roth, *supr*a note 80.

[82] *See* Penka Georgieva, *Fuzzy Rule-Based Systems for Decision-Making*, 53 J. BULGARIAN ACAD. SCI. 5, 10–14 (2016).

Familiar examples of rule bases include the flow charts used by a call center, the scoring systems used by Consumer Reports, or the complex Boolean searches used to identify potentially relevant documents in a responding party's email during the process of electronic discovery. Because they are familiar, they typically offer comfort—often undeserved—from the sense that we can understand how they operate. But rule bases, and all attempts to codify human behavior, have unintended consequences, and absent formal evaluation, are of questionable effectiveness.

In a seminal 1985 study,[83] Blair and Maron had lawyers and paralegals construct Boolean queries and then review the resulting documents until they believed they had found at least 75% of those that were relevant to each of 51 different aspects (*i.e.*, essentially topics or requests for production) related to a San Francisco Bay Area Rapid Transit ("BART") train accident. These search and retrieval efforts found, on average, only 20% of the documents relevant to each aspect. This result indicates that humans are not nearly as good at constructing Boolean queries—or any other sort of rules—as they may think they are. That is why rule-base approaches are typically time consuming and require experts and validation processes.

But manually constructed rules offer transparency that machine-learned models do not; particularly the models that result from Deep Learning, which are generally not well understood by their developers, if at all. Users can observe and understand the mechanics of how rules work, from which it is all too easy for them to draw specious conclusions regarding how *effectively* they achieve their intended purposes. Often, however, it is the less transparent algorithms that have better predictive power.[84]

Some automated learned models are more transparent than others, for example, if the feature engineering is straightforward and the method of combining evidence from the features is not too complicated. We can easily comprehend the process of dividing text into words, and even without understanding the formula, we can understand that each word might have a score indicating the weight of evidence that it conveys. We can display the top-scoring words that contribute to the classification or ranking of a document. Some learned models are in essence decision trees or Boolean queries. These models closely resemble rule bases that might be constructed manually, offering a measure of transparency. But they are typically more

---

[83] David C. Blair & M.E. Maron, *An Evaluation of Retrieval Effectiveness for a Full-Text Document-Retrieval System*, 28 COMM'NS ACM 289, 289 (1985).

[84] *See, e.g.*, Grant Duwe and Kim KiDeuk, *Sacrificing Accuracy for Transparency in Recidivism Risk Assessment: The Impact of Classification Method on Predictive Performance*, 1 CORRECTIONS 155, 155–76 (2016), https://www.tandfonline.com/doi/pdf/10.1080/23774657.2016.1178083 (last visited Nov. 15, 2021).

complex and so, notwithstanding their apparent transparency, they are not as easily understood as more straightforward manually constructed rule bases.

Other learned models, including those underlying state-of-the-art image recognition, voice recognition, and translation methods, are inscrutable to humans. In face recognition, for example, several layers will group together increasingly abstract features, which only vaguely correspond to features a human would recognize, such as eyes, ears, and hair color.[85] The issues that arise are illustrated by those that arise in DNA analysis—a high-profile classification problem.[86] No human can work through the operation of the classification algorithm, but after many years, the results have been shown to be much more reliable than more "transparent" alternatives.

Closely associated with the current wave of AI enthusiasm are the notions of "big data," "data analytics," "data mining," and "data science." "Big data" refers to the algorithms and techniques used to harness a massive glut of raw data, as opposed to the carefully curated information stored in a structured database.[87] "Data analytics"[88] and "data mining"[89] refer to processes for harvesting previously unknown information from a vast sea of raw data, while "data science" refers to the practice of performing data analytics or data mining.[90] Arguably, public-health researchers and meteorologists have been doing "data science" for years without labeling their efforts as such, but as for DNA testing, whether or not these pursuits are AI is a distinction without a difference. Data analytics in law is typically used to respond to questions facing lawyers that ought to have data-driven

---

[85] *See* OMAR M. PARKHI ET AL., DEEP FACE RECOGNITION 1, 2, 5–8 (Xianghua Xie et al. eds., BMVA Press 2015).

[86] DNA analysis is not commonly referred to as AI, but it addresses a classification problem that at one time was considered the exclusive domain of human perception and intellect. The same can be said for weather forecasting.

[87] *See* Troy Segal, *Big Dat*a, INVESTOPEDIA (Jan. 1, 2021), https://www.investopedia.com/terms/b/big-data.asp [https://perma.cc/EV97-NVC2]; s*ee also Big Data: What It Is and Why It Matters*, SAS, https://www.sas.com/en_ca/insights/big-data/what-is-big-data.html [https://perma.cc/46RN-6TCG]; *see also What is Big Data?*, ORACLE CANADA, https://www.oracle.com/ca-en/big-data/what-is-big-data.html [https://perma.cc/WSH3-4PXT].

[88] *See* Jake Frankenfield, *Data Analytics*, INVESTOPEDIA (Sept. 4, 2021), https://www.investopedia.com/terms/d/data-analytics.asp [https://perma.cc/SFQ5-NV29]; *Big Data Analytics: What It Is and Why It Matters*, SAS, https://www.sas.com/en_ca/insights/analytics/big-data-analytics.html [https://perma.cc/K8AE-DJ8T].

[89] *See* Alexandra Twin, *Data Mining*, INVESTOPEDIA (Sept. 17, 2021), https://www.investopedia.com/terms/d/datamining.asp [https://perma.cc/Z225-6DM2]; *see also Data Mining: What It Is & Why It Matters*, SAS, https://www.sas.com/en_ca/insights/analytics/data-mining.html [https://perma.cc/Q3J9-V6GX].

[90] *See* Caroline Banton, *Data Science*, INVESTOPEDIA (Sept. 12, 2021), https://www.investopedia.com/terms/d/data-science.asp [https://perma.cc/N2BH-KWSG]*; see also What Is Data Science?*, ORACLE CANADA, https://www.oracle.com/ca-en/data-science/what-is-data-science.html [https://perma.cc/6598-8UYU].

answers, such as: "What is the market for this [product or service]?"; "How long is this going to take and what will it cost?"; "Which [jurisdiction/court/judge/argument] is most likely to result in a favorable outcome?"; "What has [our firm/opposing counsel/the judge] done in the past?; "How big is the risk?"

To the extent that AI techniques are used to do classification, ranking, or regression, their effectiveness can be measured and compared to current best practice, given enough examples representative of best practice.[91] If, on the other hand, the techniques are used to cluster data, to detect anomalies, or to predict exceedingly rare events, it is quite difficult to establish their efficacy and reliability.

One of the authors of this paper (Cormack) had an unfortunate interaction with two anomaly detection algorithms. After using his credit card at New York's JFK airport in the afternoon, at Los Angeles' LAX airport in the evening, and at Brisbane's BNE airport the following morning, his credit card ceased working, because the issuing bank's fraud-detection software flagged it. At the same time, the bank left a phone message on the author's voicemail, but the email notification of that message was flagged as spam and was not delivered by the email provider (who was one and the same as the voicemail provider). As a result, the credit card could not be used for the duration of the author's trip to Australia. After contacting the bank and learning of their attempt to call him, the author also discovered several voicemail messages from the Canada Revenue Agency ("CRA"), which were also flagged as spam. Arguably, neither blocking the credit card nor blocking the messages from the bank and the government would be considered reasonable human errors. Whether or not the fraud-detection and email-filtering AI methods would be considered reasonable would depend on their overall accuracy: How often do they make such errors versus how often do they not? It would also depend on their reliability with respect to similar situations: A fraud-detection method that flagged every trip to Australia would not be considered reasonable, even though trips to Australia for the author are rare events; a spam filter than blocked all voicemail messages from CRA would not be considered reasonable, even if CRA rarely calls the author.

[91] *See generally* ALICE ZHENG, EVALUATING MACHINE LEARNING MODELS: A BEGINNERS GUIDE TO KEY CONCEPTS AND PITFALLS (O'Reilly Media 2015), https://www.scribd.com/document/465392869/Evaluating-Machine-Learning-Models [https://perma.cc/E9G3-DPLY].

While issues concerning the validity and reliability[92] of AI methods for their intended purposes are addressed in more detail later this paper, they are often eschewed in the rush to market,[93] even though they are critical to assessing AI technologies and, in particular, the value of the output as evidence in litigation. We will discuss this issue further in section VII below.

IV. USES OF AI IN BUSINESS AND LAW TODAY

In recent years, AI has made major inroads in many fields, including health care, education, employment, banking and finance, policing, and the criminal justice system, to name but a few. Before the COVID 19 pandemic hit, a Canadian-based company, BlueDot, used AI to identify an emerging health risk in China on December 31, 2019, and subsequently, to predict the global spread of the disease.[94] Two of the authors of this paper (Grossman and Cormack) used supervised machine learning to assist medical researchers at St. Michael's Hospital in Toronto, working in conjunction with the Canadian Frailty Network, Health Canada, and the World Health Organization ("WHO") to perform rapid systematic reviews to identify scientific studies related to methods for preventing the transmission of Coronavirus in older adults living in long-term care, and to determine the effectiveness and safety of therapeutic options for COVID-19 and other Coronaviruses that cause serious respiratory infections, by searching massive medical publication and pre-print services that were constantly updating.[95] Using AI, they were able to hasten a task that normally can take

[92] *Validity* refers to the degree to which an AI tool measures what it purports to measure. *Reliability* refers to the consistency with which it does so. Valid algorithms are accurate predictors; reliable algorithms reach the similar conclusions in similar circumstances over time. One useful classification scheme further subdivides validity into "construct validity," *i.e.*, whether the measurements derived from the data measure what we think they measure, "internal validity," *i.e.*, whether the analysis correctly leads from the measurements to the conclusions reached, and "external validity," *i.e.*, whether and the extent to which findings from the measurements can be generalized to other situations. *See* Alexandra Olteanu et al., *Social Data: Biases, Methodological Pitfalls, and Ethical Boundaries*, FRONTIERS BIG DATA, July 11, 2019, at 1, 4–5.

[93] Indeed, one commentator has even connected the appearance of big data with the demise of the scientific method. "[F]aced with massive data, [the scientific approach]—hypothesize, model, test—is becoming obsolete. . . . Petabytes allow us to say: 'Correlation is enough.' We can stop looking for models. We can analyze the data without hypotheses about what it might show. We can throw the numbers into the biggest computing clusters the world has ever seen and let statistical algorithms find patterns where science cannot." Chris Anderson, *The End of Theory: The Data Deluge Makes the Scientific Method Obsolete*, WIRED (June 23, 2008, 12:00 PM), https://www.wired.com/2008/06/pb-theory [https://perma.cc/XC8G-T54X].

[94] *See* Bill Whitaker, *The Computer Algorithm that Was Among the First to Detect the Coronavirus Outbreak*, 60 MINUTES (April 27, 2020), https://www.cbsnews.com/news/coronavirus-outbreak-computer-algorithm-artificial-intelligence [https://perma.cc/8NBN-RMHD].

[95] "A systematic review attempts to identify, appraise and synthesize all the empirical evidence that meets pre-specified eligibility criteria to answer a specific research question. Researchers conducting

a year or more to less than two weeks. For a number of years, dermatologists have used AI to help predict skin cancers,[96] and radiologists have used AI to help determine whether patients have breast cancer, often more accurately than they can do on their own, unaided by such technology.[97]

AI has also been used to evaluate the performance of teachers,[98] to determine who gets job interviews,[99] and in credit forecasting for loans,[100]

systematic reviews use explicit, systematic methods that are selected with a view aimed at minimizing bias, to produce more reliable findings to inform decision making. . . . A Cochrane Review is a systematic review of research in health care and health policy that is published in the *Cochrane Database of Systematic Reviews*." *About*, COCHRANE LIBR., https://www.cochranelibrary.com/about/about-cochrane-reviews [https://perma.cc/6DX7-P97M] (last visited Jan. 22, 2021). A rapid (systematic) review is one that is typically completed in five weeks or less, although the time frame can vary. *See Systematic Reviews and Other Review Types*, TEMP. U. LIBR., https://guides.temple.edu/c.php?g=78618&p=4156608 [https://perma.cc/6364-75VJ]. For authors Grossman and Cormack's systematic reviews related to COVID-19, see Patricia Rios et al., *Preventing the Transmission of COVID-19 and Other Coronaviruses in Older Adults Aged 60 Years and Above Living in Long-Term Care: A Rapid Review*, 9 SYS. REV. 118 (2020); Patricia Rios et al., *Effectiveness and Safety of Pharmacological Treatments for COVID-19: A Rapid Scoping Review*, BR. MED. J. (forthcoming 2022).

[96] *See generally* Andre Esteva et al, *Dermatologist-Level Classification of Skin Cancers with Deep Neural Networks*, 542 NATURE 115 (2017); Kara Mayer Robinson, *How Artificial Intelligence Helps Diagnose Skin Cancer*, WEBMD,https://www.webmd.com/melanoma-skin-cancer/features/ai-skin-cancer#1 [https://perma.cc/K3GZ-CMHL].

[97] *See generally* Scott May McKinney et al., *International Evaluation of an AI System for Breast Cancer Screening*, 577 NATURE 89 (2020); Hannah Slater, *AI Assisted Radiologists See Improved Performance in Detection of Breast Cancer*, CANCER NETWORK (Feb. 29, 2020), https://www.cancernetwork.com/view/ai-assisted-radiologists-see-improved-performance-detection-breast-cancer [https://perma.cc/N7Y8-4M2C]; Fergus Walsh, *AI 'Outperforms' Doctors Diagnosing Breast Cancer*, BBC NEWS (Jan. 2, 2020), https://www.bbc.com/news/health-50857759 [https://perma.cc/W2YJ-GEPQ].

[98] *See generally Hous. Fed'n of Teachers, Loc. 2415 v. Hous. Indep. Sch. Dist.*, 251 F. Supp. 3d 1168 (S.D. Tex. 2017) (lawsuit challenging use of AI to evaluate teacher performance); CATHY O'NEIL, *Sweating Bullets: On the Job*, *in* WEAPONS OF MATH DESTRUCTION: HOW BIG DATA INCREASES INEQUALITY AND THREATENS DEMOCRACY (Crown Publishers 2016). For a discussion of the use of AI tools for the purposes of law firm recruitment see Victoria Hudgins, *Diversity, Metrics Demands Are Pushing Firms to Embrace AI Hiring Tools*, LEGALTECH NEWS (Jan. 13, 2021, 12:15 PM), https://www.law.com/legaltechnews/2021/01/13/diversity-metrics-demands-are-pushing-firms-to-embrace-ai-hiring-tools [https://perma.cc/3FYZ-TXJ5].

[99] Rebecca Heilweil, *Artificial Intelligence Will Help Determine If You Will Get Your Next* Job, VOX (Dec. 12, 2019, 8:00 AM EST), https://www.vox.com/recode/2019/12/12/20993665/artificial-intelligence-ai-job-screen [https://perma.cc/M9WL-T87J].

[100] Zoran Ereiz, *Predicting Default Loans Using Machine Learning (OptiML)*, 2019 27TH TELECOMMS. F. (TELFOR) 1 (2019); Daniel Faggella, *Artificial Intelligence Applications for Lending and Loan Management*, EMERJ: THE AI RESEARCH AND ADVISORY COMPANY (Apr. 3, 2020), https://emerj.com/ai-sector-overviews/artificial-intelligence-applications-lending-loan-management [https://perma.cc/R9RD-73QY].

mortgages,[101] and credit cards[102]—sometimes resulting in high-profile scandals involving the potential for bias in such systems.[103] AI has long been used in Fintech for high-speed securities trading,[104] where the advantage of a few milliseconds can result in huge financial gains. The number of new applications of AI that emerge each week is staggering.[105]

AI has also entered the legal realm in numerous ways,[106] some more risky and harmful than others. In addition to the use of data analytics and technology-assisted review in electronic discovery, ever since TAR was first approved by the courts in 2012,[107] machine-learning technologies have also been used for contract management and for due-diligence reviews in mergers and acquisitions,[108] for public disclosure analytics,[109] for natural-language

---

[101] Lin Zhu et al., *A Study on Predicting Loan Default Based on the Random Forest Algorithm*, 162 PROCEDIA COMPUT. SCI. 503, 508–09 (2019), https://www.sciencedirect.com/science/article/pii/S1877050919320277 [https://perma.cc/KV84-VQT4]; Michael J. Cooper, A Deep Learning Prediction Model for Mortgage Default (May 2018) (Master's thesis, University of Bristol) (ResearchGate).

[102] Scott Zoldi, *How to Build Credit Risk Models Using Artificial Intelligence and Machine Learning*, FICO: BLOG (Apr. 6, 2017), https://www.fico.com/blogs/how-build-credit-risk-models-using-ai-and-machine-learning [https://perma.cc/H893-CJ94]*; Risk and Reward: The Role of AI in Acquiring Credit Card Prospects*, APPIER (Oct. 17, 2019), https://www.appier.com/blog/risk-and-reward-the-role-of-ai-in-acquiring-credit-card-prospects [https://perma.cc/72B5-95V2].

[103] *See, e.g.*, Jeffrey Dastin, *Amazon Scraps Secret AI Recruiting Tool that Showed Bias Against Women*, Reuters (Oct. 10, 2018, 6:04 PM), https://www.reuters.com/article/us-amazon-com-jobs-automation-insight-idUSKCN1MK08G [https://perma.cc/A2ZE-64J4]; Neil Vigor, *Apple Card Investigated After Gender Discrimination Complaints*, N.Y. TIMES (Nov. 10, 2019), https://www.nytimes.com/2019/11/10/business/Apple-credit-card-investigation.html [https://perma.cc/5JES-6ZUW].

[104] *See* JASMINA ARIFOVIC ET AL., HIGH FREQUENCY TRADING IN FINTECH AGE: AI WITH SPEED (SSRN 2019), https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2771153 [https://perma.cc/L9JV-UNCM].

[105] For a compendium of AI applications across all domains, the reader can register for the weekly *Cognitive RoundUp* from *SwissCognitive – The Global AI Hub*, SWISSCOGNITIVE, https://swisscognitive.ch [https://perma.cc/H5KB-YCDX].

[106] For a compendium of AI applications in law and legal practice, the reader is referred to Daniel Faggella, *AI in Law and Legal Practice – A Comprehensive View of 35 Current Applications*, EMERJ: THE AI RESEARCH AND ADVISORY COMPANY (Mar. 14, 2020), https://emerj.com/ai-sector-overviews/ai-in-law-legal-practice-current-applications [https://perma.cc/GLS7-8X2R].

[107] *See* Da Silva Moore v. Publicis Groupe, 287 F.R.D. 182 (S.D.N.Y 2012), adopted sub nom. Moore v. Publicis Groupe SA, No. 11 CIV. 1279 ALC AJP, 2012 WL 1446534 (S.D.N.Y. Apr. 26, 2012), (first federal case); Glob. Aerospace Inc. v. Landow Aviation, L.P., No. CL 61040 (Vir. Cir. Ct. Apr. 23, 2012) (first state case).

[108] Brittainy Boessel, *The Role of AI in Contract Management*, KIRA (June 15, 2020), https://kirasystems.com/learn/role-of-ai-in-contract-management [https://perma.cc/7F26-R9UG]; Ellie Nikolova, *AI's Role in Mergers and Acquisitions*, CORP. L.J. (Sept. 25, 2019), https://www.thecorporatelawjournal.com/technology/ais-role-in-mergers-and-acquisitions [https://perma.cc/B738-CUCP].

[109] For an example of an AI tool that analyzes SEC filings and associated exhibits, see LEXISNEXIS' *Intelligize*, https://www.intelligize.com/products/intelligize [https://perma.cc/3482-C77E].

legal research inquiries,[110] for legal brief analytics,[111] for drafting of legal memoranda and pleadings,[112] for litigation forecasting for the purposes of litigation funding,[113] for review of legal billing,[114] and even in bots employed to analyze claims and to complete forms to improve access to justice.[115]

[110] Nicole Black, *Lawyers Have a Bevy of Advanced and AI-Enhanced Legal Research Tools at Their Fingertips*, ABA J. (Nov. 22, 2019), https://www.abajournal.com/web/article/lawyers-have-a-bevy-of-advanced-and-ai-enhanced-legal-research-tools-at-their-fingertips [https://perma.cc/9QA3-TSW3].

[111] For examples of AI tools that can analyze briefs to find and recommend the most on-point authorities or to uncover cases that opposing counsel has failed to cite, see *CARA A.I.*, https://casetext.com/cara-ai [https://perma.cc/84BY-YEB3] or *Brief Analyzer*, https://pro.bloomberglaw.com/brief-analyzer [https://perma.cc/LS9M-5HPP]. There is even an AI brief-checking tool designed specifically for judges: *Quick Check Judicial*, https://legal.thomsonreuters.com/en/c/quick-check-judicial-on-westlaw-edge?cid=9023855&sfdccampaignid=7014O000001iorNQAQ&chl=pr [https://perma.cc/FH8S-3MEJ].

[112] For an example of an AI tool that can provide responses to legal questions in a memo form, see Alexsei, https://www.alexsei.com [https://perma.cc/2QVU-E3D9]. For examples of AI tools that automate the preparation of the first draft of legal pleadings or briefs, respectively, see LegalMation®, https://www.legalmation.com [https://perma.cc/XMP6-JC5Z], and see *Compose*, https://compose.law [https://perma.cc/PM85-WWK2] .

[113] "Legalist, a legal startup backed by PayPal co-founder Peter Thiel, bills itself as 'the first AI-powered litigation finance firm.'" *AI-Powered Litigation Finance Firm Offer Bounty to Sexual Harassment Victims*, LEGAL TECH BLOG (Oct. 19, 2017), https://legal-tech-blog.de/ai-powered-litigation-finance-firm-offers-bounty-to-sexual-harassment-victims [https://perma.cc/T3DF-WH5M]; *see also* Bob Ambrogi, *Litigation Finance Startup Legalist Raises $100 Million to Fund Lawsuits*, LAWSITES (Sept. 19, 2019), https://www.lawsitesblog.com/2019/09/litigation-finance-startup-legalist-raises-100-million-to-fund-lawsuits.html [https://perma.cc/YB5X-DRB9] ("Legalist leads the new wave of technologists using artificial intelligence and machine learning to streamline and underwrite litigation investments.").

[114] For examples of AI tools used for automated review of legal bills, see Bilr, https://www.getbilr.com/legal-invoice-review [https://perma.cc/ZR5N-RMXL], and see *Brightflag*, https://brightflag.com [https://perma.cc/W332-WQCG] .

[115] Luke Dormehl, *Meet the British Whiz Kid Who Fights Justice with a Robo-Lawyer Sidekick*, DIGITALTRENDS (March 25, 2018), https://www.digitaltrends.com/cool-tech/robot-lawyer-free-acess-justice [https://perma.cc/BV2E-LPKS] (discussing Joshua Browder's DoNotPay chatbot that has helped to successfully appeal millions of dollars' worth of parking tickets); Luis Millán, *AI Initiative Seeks to Improve Access to Justice, Law in Quebec* (Jan. 13, 2020),https://lawinquebec.com/ai-initiative-seeks-to-improve-access-to-justice [https://perma.cc/Y7GZ-VEBP]. For a comprehensive discussion of the pros and cons of the use of AI to address "the justice gap," see Katherine L.W. Norton, *The Middle Ground: A Meaningful Balance Between the Benefits and Limitations of Artificial Intelligence to Assist with the Justice Gap*, 75 U. MIA. L. REV. 190 (2020).

Perhaps of greater interest (and concern) to the present audience is software used to analyze opposing counsel or judges,[116] and for online adjudication.[117]

More controversial uses lie in the area of law enforcement and the criminal justice system, including algorithms used for predictive policing, facial recognition, bail setting, and sentencing decisions. These contexts pose higher risk of harm than many of the aforementioned uses and are more likely to come to the attention of judicial officers.

Predictive Policing has been around for some time. In or about 2010 or 2011, UCLA scientists working with the Los Angeles Police Department ("LAPD") developed a software program called PredPol, designed to analyze crime data to spot patterns of criminal behavior, so that police could intervene in predicted high-crime areas to prevent crimes from happening.[118] The software is now used by more than 60 police departments around the country to identify neighborhoods where serious crimes are more likely to occur during particular periods of time.[119] The company that designed the software claims that its research has shown that it is "twice as accurate as human analysts" in predicting where crimes will take place, but these self-

[116] For examples of AI tools that may be used for legal analytics involving opposing counsel, judges, or courts, see *Lex Machina Legal Analytics Platform*, https://lexmachina.com/legal-analytics [https://perma.cc/A56Z-F49H]; Premonition, https://premonition.ai [https://perma.cc/2S4L-QGBN]; and *Context*, *Ravel*, https://home.ravellaw.com [https://perma.cc/L2EB-QEQW]. Note that France banned the use of judicial analytics in Article 33 of the Justice Reform Act of Mar. 23, 2019. A violation of the law can result in a criminal penalty of up to five years in prison. Jason Tashea, *France Bans Publishing of Judicial Analytics and Prompts Criminal Penalties*, ABA J. (June 7, 2019), https://www.abajournal.com/news/article/france-bans-and-creates-criminal-penalty-for-judicial-analytics [https://perma.cc/C2LH-C4PH].

[117] Carole Piovesan & Vivian Ntiri, *Adjudication by Algorithm: The Risks and Benefits of Artificial Intelligence in Judicial Decision-Making*, ADVOCS.' J. 42 (2018), https://marcomm.mccarthy.ca/pubs/Spring-2018-Journal_Piovesan-and-Ntiri-article.pdf (discussing use of AI technology and online dispute resolution for low-value claims). *See also* Eric Niiler, *Can AI Be a Fair Judge in Court? Estonia Thinks So*, WIRED (Mar. 25, 2019), https://www.wired.com/story/can-ai-be-fair-judge-court-estonia-thinks-so [https://perma.cc/BJM5-3JU5] (last visited Jan. 22, 2021).

[118] 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 [https://perma.cc/RVK9-R44G]. There is considerable debate over whether the LAPD's predictive tool is effective. *Compare* Stuart Wolpert, *Predictive Policing Substantially Reduces Crime in Los Angeles During Months-Long Test*, UCLA NEWSROOM (Oct. 7, 2015), https://newsroom.ucla.edu/releases/predictive-policing-substantially-reduces-crime-in-los-angeles-during-months-long-test [https://perma.cc/2E93-WDGA] *with* Mark Puente, *LAPD Pioneered Predicting Crime with Data. Many Police Don't Think It Works*, L.A. TIMES (July 3, 2019),https://www.latimes.com/local/lanow/la-me-lapd-precision-policing-data-20190703-story.html [https://perma.cc/2PPC-LB2T]. With more and more critics—particularly with respect to its potentially discriminatory impact on minority populations—predictive policing "may be falling out of fashion." Eva Ruth Moravec, *Do Algorithms Have a Place in Policing?*, ATLANTIC (Sept. 5, 2019), https://www.theatlantic.com/politics/archive/2019/09/do-algorithms-have-place-policing/596851 [https://perma.cc/8VXK-H95J].

[119] Rieland, *supra* note 118.

reported results have not been independently verified.[120] The City of Chicago took crime projection a step further by building a "Strategic Subject List" of individuals "most likely to be involved in future shootings," either as perpetrators or victims.[121] The American Civil Liberties Union ("ACLU"), the Brennan Center for Justice, and other civil rights organizations have sounded the alarm about the risk of bias inherent in such prediction software because historical data from police practices is used to train the algorithm, leading to a feedback loop through which the software makes forward-looking decisions that both reflect and reinforce past beliefs about which neighborhoods (or which people) are "safe" or "dangerous."[122] Software that relies on arrest data carries an even higher degree of risk of bias than software based on, for example, convictions, because it is more reflective of police practices than actual crime.[123] After all, police only arrest people for crimes where they look for them.

Facial recognition by police has recently come under greater scrutiny. In June 2020, the *New York Times* reported on the first-known case where a faulty facial recognition match led to the arrest of a Michigan man for a crime he did not commit.[124] The man was handcuffed on his front lawn, in front of his wife and two young daughters, and subsequently booked and held overnight for allegedly shoplifting five watches worth $3,800 from an upscale Detroit boutique, based on a grainy still image retrieved from a surveillance video that was incorrectly matched to the man's driver's license photo by a facial recognition algorithm used to search a police database of 49 million photos.[125] Apparently, without much further investigation, the detectives simply included the large Black man's picture in a six-pack photo lineup that they showed to the store's loss-prevention coordinator—who had previously reviewed the store's surveillance video and sent a copy to the Detroit police—and she subsequently identified the man as the perpetrator.[126]

---

[120] *Id.*

[121] *Id.* The controversial eight-year program was quietly retired in early 2020. Jeremy Gorner & Annie Sweeney, *For Years Chicago Police Rated the Risk of Tens of Thousands Being Caught Up in Violence. That Controversial Effort Has Quietly Been Ended.*, CHI. TRIB. (Jan. 24, 2020), https://www.chicagotribune.com/news/criminal-justice/ct-chicago-police-strategic-subject-list-ended-20200125-spn4kjmrxrh4tmktdjckhtox4i-story.html [https://perma.cc/U73T-3CZC].

[122] Rieland, *supra* note 118.

[123] *Id.*

[124] Kashmir Hill, *Wrongfully Accused by an Algorithm*, N.Y. TIMES (June 24, 2020), https://www.nytimes.com/2020/06/24/technology/facial-recognition-arrest.html [https://perma.cc/B6VA-2HS8].

[125] *Id.*

[126] *Id.* Apparently, this did not turn out to be the first such event. An earlier misidentification occurred in May 2019 when the Detroit Police wrongly charged a 25-year-old Black man of felony larceny for allegedly reaching into a teacher's vehicle, grabbing a cellphone, and throwing it, resulting in a cracked

Facial recognition systems have been used by police for more than two decades.[127] Recent studies conducted by researchers at the Massachusetts Institute of technology ("MIT") and Microsoft Research, as well as at NIST, have found that while the technology works relatively well on White men, the results are less accurate for other demographics, in part, because they are less well represented in the sources of the images used to train the algorithms.[128] These AI tools are reported to falsely identify African American and Asian faces between 10 and 100 times more often than Caucasian faces.[129] In the same month as the *New York Times* reported on the Michigan misidentification case, Amazon, Microsoft, and IBM announced that they planned to cease—or at least pause—their facial recognition offerings for law enforcement.[130] But these are not the big players in this industry,[131] so the use of these technologies by police departments continues

screen and broken case. Elisha Anderson, *Controversial Detroit Facial Recognition Got Him Arrested for a Crime He Didn't Commit*, DETROIT FREE PRESS (July 10, 2020), https://www.freep.com/story/news/local/michigan/detroit/2020/07/10/facial-recognition-detroit-michael-oliver-robert-williams/5392166002 [https://perma.cc/87FH-K9FR]. Since the publication of these two articles, a third misidentification of a Black man using faulty facial recognition has occurred. *See* Kashmir Hill, *Another Arrest, and Jail Time, Due to a Bad Facial Recognition Match*, N.Y. TIMES (Dec. 29, 2020), https://www.nytimes.com/2020/12/29/technology/facial-recognition-misidentify-jail.html. [https://perma.cc/7ZNP-85K6]. In this instance, a New Jersey man was accused of "shoplifting candy and trying to hit a police officer with a car . . . "The man turned out to have been 30 miles away at the time of the incident. He spent 10 days in jail and paid approximately $5,000 to defend himself. *Id.*

[127] Jennifer Valentino-DeVries, *How the Police Use Facial Recognition, and Where It Falls Short*, N.Y. TIMES (Jan. 12, 2020), https://www.nytimes.com/2020/01/12/technology/facial-recognition-police.html [https://perma.cc/6QS7-7HH4].

[128] Kyle Wiggers, *NIST Benchmarks Show Facial Recognition Technology Still Struggles to Identify Black Faces*, VENTUREBEAT (Sept. 9, 2020), https://venturebeat.com/2020/09/09/nist-benchmarks-show-facial-recognition-technology-still-struggles-to-identify-black-faces [https://perma.cc/3ANZ-FQGB]; Larry Hardesty, *Study Finds Gender and Skin-Type Bias in Commercial Artificial-Intelligence Systems*, MIT NEWS (Feb. 11, 2018), https://news.mit.edu/2018/study-finds-gender-skin-type-bias-artificial-intelligence-systems-0212 [https://perma.cc/N4T9-UKF4]; Steve Lohr, *Facial Recognition is Accurate if You're a White Guy*, N.Y. TIMES (Feb. 9, 2018), https://www.nytimes.com/2018/02/09/technology/facial-recognition-race-artificial-intelligence.html [https://perma.cc/TD6Z-RPB2]. One of the authors of the MIT/MS Research study (Timnit Gebru) claimed that she was later fired by Google because she refused to retract a subsequent paper also on responsible AI and algorithmic accountability. *See* Nitasha Tiku, *Google Hired Timnit Gebru to Be an Outspoken Critic of Unethical AI. Then She Was Fired for It.*, WASH. POST (Dec. 23, 2020), https://www.washingtonpost.com/technology/2020/12/23/google-timnit-gebru-ai-ethics [https://perma.cc/GUA5-X7HM]; Alex Hanna & Meredith Whittaker, *Timnit Gebru's Exit From Google Exposes a Crisis in AI*, WIRED (Dec. 31, 2020), https://www.wired.com/story/timnit-gebru-exit-google-exposes-crisis-in-ai [https://perma.cc/ZH4R-D6LG].

[129] Natasha Singer & Cade Metz, *Many Facial-Recognition Systems Are Biased, Says U.S. Study*, N.Y. TIMES (Dec. 19, 2019), https://www.nytimes.com/2019/12/19/technology/facial-recognition-bias.html [https://perma.cc/48NZ-MKVT].

[130] Hill, *supra* note 124.

[131] The technology that police departments use is supplied by Vigilant Solutions, Cognitec, NEC, Rank One Computing, and Clearview AI, and NTech Labs. Hill, *supra* note 124; Tate Ryan-Mosley,

largely unabated.[132] Since the Michigan misidentification case was reported, at least two more arrests of Black men using faulty facial identification have been divulged.[133]

Perhaps of even more concern than predictive policing and the use of facial recognition by law enforcement is the following. In March, 2016, an article published in *Pro Publica* reported on risk-assessment software called the Correctional Offender Management Profiling for Alternative Sanctions ("COMPAS") that was being used, with increasing frequency, across the United States to inform—and sometimes to make—decisions about a criminal defendant's or convict's risk of reoffending during various points in the criminal justice system, from pre-trial release, to criminal sentencing and probation.[134] These tools are vaguely reminiscent of "Minority Report," "1984," "Black Mirror," and other dystopian science fiction.[135] COMPAS is not the only proprietary risk and needs assessment ("RNA") tool available—there are over 100 general and specialty tools that have been developed by private entities, non-profit organizations, universities, and even states.[136] While most of the tools are computerized to some degree, not all of them rely on AI to make predictions. COMPAS does. At the time of the *Pro Publica* article, while dozens of criminal RNA tools were in use, few had been independently tested.[137] In a 2013 study, researchers Sarah Desmarais and Jay Singh examined 19 such tools used across the United States and found that "in most cases, validity had been examined in one or two studies,"

There Is a Crisis of Face Recognition and Policing in the US, MIT TECH. REV. (Aug. 14, 2020), https://www.technologyreview.com/2020/08/14/1006904/there-is-a-crisis-of-face-recognition-and-policing-in-the-us [https://perma.cc/D7CX-PDZP].

[132] Hill, *supra* note 124. We do not actually know how often U.S. police departments use facial recognition because in most jurisdictions they are not required to report it. The most recent numbers come from 2016 and are speculative, but they suggest that at that time, at least half of Americans' photos were contained in a facial recognition system and that one county in Florida ran 8,000 searches each month. Ryan-Mosley, *supra* note 131.

[133] *See supra* note 126; Ryan-Mosley *supra* note 131.

[134] Julia Angwin et al., *Machine Bias*, PROPUBLICA (May 23, 2016), https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing [https://perma.cc/H3S4-G7PQ].

[135] *See* Rhys Dipshan et al., *The United States of Risk Assessment: The Machines Influencing Criminal Justice Decisions*, LEGALTECH NEWS (July 13, 2020), https://www.law.com/legaltechnews/2020/07/13/the-united-states-of-risk-assessment-the-machines-influencing-criminal-justice-decisions [https://perma.cc/Q5DV-WN2W].

[136] Specialized tools include those used for women or juvenile offenders, and those that assess a defendant's or convict's likelihood of committing domestic or sexual violence.

[137] Angwin et al., *supra* note 134. For a comprehensive critique of "the serious shortcomings of risk assessment tools in the U.S. criminal justice system," including "[c]oncerns about the validity, accuracy, and bias in the tools themselves," see PARTNERSHIP ON AI, *Report on Algorithmic Risk Assessment Tools in the U.S. Justice System*, 2 (2019), https://www.partnershiponai.org/report-on-machine-learning-in-risk-assessment-tools-in-the-u-s-criminal-justice-system [https://perma.cc/LQR7-3RH7].

and that "frequently, those investigations were completed by the same people who developed the instrument."[138] They concluded that the tools "were moderate at best in terms of predictive validity."[139] *Pro Publica's* own study was even more troubling. Their reporters collected the risk scores of more than 7,000 people arrested in Broward County, Florida, in 2013 and 2014, and followed them to see how many were charged with another crime over the following two years, the same benchmark used by COMPAS.[140] "The score proved remarkably unreliable in forecasting violent crime: Only 20% of the people predicted to commit violent crimes actually went on to do so."[141] When a full range of crimes was taken into account, "[o]f those deemed likely to reoffend, 61% were arrested for any subsequent crimes within two years."[142] What *Pro Publica* found next was even more problematic—significant racial disparities: Black offenders were twice as likely as White offenders to be incorrectly labeled as high risk (44.85% versus 23.45%), while White offenders were twice as likely as Black offenders to be incorrectly labeled as low risk (47.72% versus 27.99%).[143] COMPAS' developer admitted that it was difficult to construct a score that did not include items that could be correlated with race—such as poverty, joblessness, and social marginalization. "If those are omitted from your risk assessment, accuracy goes down."[144]

Defendants rarely have an opportunity to challenge the results of their risk and need assessments. While the overall score may be shared with their attorney, the algorithm that produced the score, and the underlying data on which it relied, are typically not disclosed; they are almost always withheld as proprietary trade secrets. This problem was raised in a Wisconsin criminal case involving defendant, Eric Loomis, who was a repeat offender labeled by COMPAS as high risk to the community.[145] Loomis was charged with

---

[138] Angwin et al, *supra* note.134.

[139] *Id.*

[140] *Id.*

[141] *Id.*

[142] *Id.*

[143] *See id.; see also* Jeff Larson et al., *How We Analyzed the COMPAS Recidivism Algorithm*, PRO PUBLICA (May 23, 2016), https://www.propublica.org/article/how-we-analyzed-the-compas-recidivism-algorithm [https://perma.cc/DXW4-ME4E]. However, Pro Publica's analysis of the COMPAS data is not without its critics. *See, e.g.,* Anthony W. Flores et al., *False Positives, False Negatives, and False Analyses: A Rejoinder to "*Machine Bias: There's Software Used Across the Country to Predict Future Criminals. And it's Biased Against Blacks.*"*, 80 FED. PROB. 1 (Sept. 2016), https://www.researchgate.net/publication/306032039_False_Positives_False_Negatives_and_False_Analyses_A_Rejoinder_to_Machine_Bias_There%27s_Software_Used_Across_the_Country_to_Predict_Future_Criminals_And_it%27s_Biased_Against_Blacks [https://perma.cc/D4DS-Z3ZV].

[144] Angwin et al., *supra* note 134.

[145] Wisconsin v. Loomis, 881 N.W.2d 749, 755 (2016).

driving a stolen vehicle away from the scene of a drive-by shooting and fleeing the police.[146] The judge in the case imposed a sentence of eleven years.[147] Loomis challenged the use of the COMPAS score at his sentencing as a violation of his due process rights because the proprietary nature of the tool prevented him from challenging the scientific validity of the assessment (*e.g.*, how COMPAS weighed various factors, how the algorithm calculated risk, the impact of the comparator data—which was based on a national not a local (*i.e.,* Wisconsin) sample, the fact that some studies of COMPAS' RNA scores had raised questions about whether they disproportionally classified minorities as having a higher risk of recidivism, and thus, the accuracy of the scores).[148]

The case went to the Wisconsin Supreme Court, which pointed out that the Presentence Investigation (i) warned that "the COMPAS risk assessment does not predict the specific likelihood that an individual will reoffend. Instead, it provides a prediction based on a comparison of information about the individual to a similar data group," and (ii) cautioned that "risk scores are not intended to determine the severity of a sentence or whether an offender is incarcerated."[149] Because the COMPAS risk score was accompanied by such admonitions, and was not the *sole* determinant of the Court's sentencing decision—it was ostensibly used only to corroborate the Court's findings—its use did not violate a Mr. Loomis' right to due process.[150]

These examples are only the tip of the iceberg with respect to how lawyers and judges can expect AI to arise in the cases they handle, and how AI increasingly may be applied in the justice system.

V. ISSUES RAISED BY THE USE OF AI IN BUSINESS AND LAW TODAY

While AI offers great promise for the advancement of social good in many domains—including access to justice—it also poses significant risks and challenges, some of which are likely apparent from the examples provided above. Unfortunately, the benefits and burdens of AI are often not equally distributed across society, and we risk losing the benefits if we cannot find solutions to the challenges raised by AI. Some of these challenges are discussed below.

---

[146] *Id.* at 754.

[147] *See Id.* at 756 note 18.

[148] *Id.* at 756, 760–63.

[149] *Id.* at 754, 770.

[150] *See id.* at 755, 771–72.

A. Bias

Bias leading to sometimes intended—but more often unintended—discriminatory outcomes is a serious problem with AI. There are multiple places where bias can impact AI systems, from the inputs to the outputs of such systems, and even in the ways in which the outputs are interpreted and used by humans.[151]

Because machine-learning algorithms are trained using historical data, they can serve to perpetuate the very biases they are often intended to prevent. Bias in data can occur because the training data is not representative of a target population to which the AI system will later be applied. Two high-profile examples of this problem include Google Photo's mistaken identification of two Black people as gorillas,[152] and Amazon's failed experiment with a hiring algorithm that merely replicated the company's existing disproportionately male workforce.[153] We see this same problem with facial recognition software that has difficulty correctly identifying Black women's faces because they are not adequately reflected in the training set.[154] Data can also be differentially noisy for different groups, meaning that errors are not evenly distributed across the different groups, or data may simply be missing for certain groups as compared to others, for example, when the data is either unavailable or the collection process is incomplete because the techniques used to capture data fail to capture all data equally. This is particularly the case when the law prohibits collecting, labeling, or using the data of certain protected groups. This can cause other problems, for example, when a treatment actually works better for one gender or race than another, but the beneficial effect is masked by an overall (*i.e.*, combined) accuracy rate that is low, or because the protected data is either not collected or not considered by the algorithm.[155] Defendant Loomis

---

[151] For a useful discussion of some of the different types of bias that can impact AI systems, see Selena Silva & Martin Kenney, *Viewpoint: Algorithms, Platforms, and Ethnic Bias*, 62 COMMC'N ACM 37 (2019), https://cacm.acm.org/magazines/2019/11/240361-algorithms-platforms-and-ethnic-bias/fulltext [https://perma.cc/JF4P-APNW].

[152] Maggie Zhang, *Google Photos Tags Two African-Americans As Gorillas Through Facial Recognition Software*, FORBES (July 1, 2015), https://www.forbes.com/sites/mzhang/2015/07/01/google-photos-tags-two-african-americans-as-gorillas-through-facial-recognition-software/?sh=23e9c0a3713d [https://perma.cc/G5ZW-NDDZ]; Pete Pachal, *Google Photos Identified Two Black People As 'Gorillas,'* MASHABLE (July 1, 2015), https://mashable.com/2015/07/01/google-photos-black-people-gorillas [https://perma.cc/5RGQ-K5NB].

[153] Dastin, *supra* note 103.

[154] *E.g.*, Lohr, *supra* note 128.

[155] *Cf.* Heather P. Whitley & Wesley Lindsey, *Sex-Based Differences in Drug Activity*, 80 AM. FAM. PHYSICIAN 1254 (2009), https://www.aafp.org/afp/2009/1201/p1254.html [https://perma.cc/TA2Z-3WC2]; Valentine J. Burroughset al., *Racial and Ethnic Differences in Response to Medicines: Towards Individualized Pharmaceutical Treatment*, 94 J. NAT'L MED. ASS'N 1 (2002), https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2594139 [https://perma.cc/L8XW-ZAE3]. A similar

asserted that the COMPAS tool discriminated on the basis of gender because the tool assessed male and female offenders separately due to the fact that research has shown that female offenders are different from male offenders.[156] Thus, it is not always clear when information about protected classes should and should not be used by AI.

Data can also be biased for the reason that while an AI system may not take a protected class label or feature such as race directly into account, the data includes proxies for that label or feature that the algorithm does consider. For example, the COMPAS tool asks for information about arrests for drug possession and use.[157] It is well known that Black people are arrested for drug possession and use many times more often than White people,[158] so this question is a ready proxy for race, as are many other features like zip code, education, employment, and incarceration. When arrest records for drug use are used as a predictor in RNA algorithms, they may be more reflective of police activity than recidivism risk and can therefore lead to

problem can occur when an algorithm fails to take racial differences into account when it should. In one prominent example, a health-care algorithm used health-care costs as a proxy for health-care needs, without taking into account the fact that unequal access to health care meant that less money was spent caring for Black patients than White patients. Thus, at the same score on the predictive measure, Black patients were considerably sicker than White patients, but were systematically offered less care. Ziad Obermeyer et al., *Dissecting Racial Bias in an Algorithm Used to Manage the Health of Populations*, 366 SCI. 447, 453 (2019), https://science.sciencemag.org/content/366/6464/447/tab-pdf [https://perma.cc/NB9D-XN4Q]. Tom Simonite, *A Health Care Algorithm Offered Less Care to Black Patients*, WIRED (Oct. 24, 2019), https://www.wired.com/story/how-algorithm-favored-whites-over-blacks-health-care [https://perma.cc/ZPG7-ND4U].

[156] *See Loomis*, 881 N.W.2d at 765–66; *see also* Rhys Dipshan, *Constitutional Brawl Looms Over How Risk Assessment Tools Account for Gender*, LEGALTECH NEWS (July 20, 2020), https://www.law.com/legaltechnews/2020/07/20/constitutional-brawl-looms-over-how-risk-assessment-tools-account-for-gender [https://perma.cc/YN6J-3F3S].

[157] See *Risk Assessment*, Northpointe Suite v. 8.1.18.12 (Northpointe, Inc. 2011) ("19. How many prior possession/use offense arrests as an adult?") (copy on file with author Grossman).

[158] *See* Rhys Dipshan & Victoria Hudgens, *Risk Assessment Tools Aren't Immune From Systemic Bias. So Why Use Them?*, LEGALTECH NEWS (July 17, 2020), https://www.law.com/legaltechnews/2020/07/17/risk-assessment-tools-arent-immune-from-systemic-bias-so-why-use-them [https://perma.cc/QST5-LTQA]. Dr. Jennifer Skeem, Professor of Public Policy at the University of California, Berkeley "notes that, where possible, tools should avoid criteria that [are] impacted by the differential treatment African Americans receive in the criminal justice system. 'A really good example is arrest for drug offense. We know that policing patterns make it such that Blacks are much more likely to be arrested for drug offenses than whites, even though there isn't much difference at the behavioral level and in terms of rates of drug use, etc.'" *See also* Peter Walker, *Black People Twice as Likely to Be Charged with Drug Possession – Report*, THE GUARDIAN (Aug. 21, 2013), https://www.theguardian.com/world/2013/aug/21/ethnic-minorities-likely-charged-drug-possession [https://perma.cc/FHJ8-HYXZ]; PARTNERSHIP ON AI, *supra* note 137, at 16, n.15 ("Statistical validation of recidivism in particular suffers from a fundamental problem: the ground truth of whether an individual committed a crime is generally unavailable, and can only be estimated via imperfect proxies such as crime reports or arrests. . . . One problem with using such imperfect proxies is that different demographic groups are stopped, searched, arrested, charged, and are wrongfully convicted at very different rates in the current US criminal justice system." (citations omitted)).

biased outcomes. Another example would be an AI tool that uses health-care costs as a measure of health-care needs. It is well known that minority communities have less access to health care and pay less into the health-care system, thus their needs may be improperly reflected when the algorithm considers health-care costs as a measure of health-care needs. In these examples, what we observe, and measure does not line up with what we actually care about.

Finally, bias in data also can obviously occur because the data reflects the systematic race and gender discrimination that exists in society. This can be seen with tools that assess resumes for interviews,[159] or applications for Apple credit cards.[160]

While the line between the data and the model that is derived from it can be fuzzy, bias can also come into play with respect to the algorithm itself. Most machine-learning algorithms are premised on "bias" in the sense that their entire purpose is to discriminate between X and Y, because that is what helps the tool to make predictions. The choice of tool itself imposes assumptions on the data, and norms and values are built into the models the tools generate. AI developers make certain decisions about problem specification, what the system is trying to model or predict and how best to do so, including methods for data cleansing and processing, the features the system will consider, the weights the system will assign to those features, how data (particularly outliers) are to be treated, outcome variables, and so on, often without much consideration of the potential harms or unintended consequences that can flow from these hidden choices. For example, if an AI system is trying to predict the quality of employees, and it takes the number of promotions, raises, and highest-attained salary into account, the output will necessarily be biased because those features typically are not evenly distributed across race and gender. Another example might be an algorithm designed to determine where street repairs are needed based on reports of potholes reported by phone or on a website. But if the algorithm does not take into account the fact that not all people in all neighborhoods have access to cell phones or computers, and undocumented residents may be unwilling to contact a public agency, such an algorithm would only serve to increase disparities in road conditions in poor versus wealthy neighborhoods.

AI developers may not be the best qualified or the best equipped to make such algorithmic design choices in light of the fact that that they typically do not reflect the diversity of the populations to which the

[159] *See* Dastin, *supra* note 103.

[160] *See* Vigor, *supra* note 103.

algorithms will be applied,[161] and have little to no training in ethics or the law, and therefore may be insensitive to the unintended consequences of their decisions. Lawyers, ethicists, policy makers, and regulators are brought into the process, if at all, long after these decisions have been made and when they are no longer transparent or easily altered. This oversight results in silent failures that often go undetected until they result in public relations nightmares.

Most AI tools place a great emphasis on achieving predictive accuracy and efficiency, but do not always consider statistical or demographic parity,[162] the distribution of false positives and false negatives,[163] or other measures of fairness and bias. Even if society were able to come to consensus on a definition of "fairness" in AI,[164] fairness would still be incredibly hard

---

[161] *See* Sarah Myers West et al., *Discriminating Systems: Gender, Race and Power*, in AI, AI NOW INSTITUTE (Apr. 2019), https://ainowinstitute.org/discriminatingsystems.pdf [https://perma.cc/8AXB-46RV]; *see also* Kari Paul, *'Disastrous' Lack of Diversity in AI Industry Perpetuates Bias, Study Finds*, THE GUARDIAN (Apr. 17, 2019), https://www.theguardian.com/technology/2019/apr/16/artificial-intelligence-lack-diversity-new-york-university-study [https://perma.cc/KB5C-MGKY].

[162] An unknown author once defined *statistical parity* as "the statistical equivalent of the legal doctrine of adverse impact. It measures the difference that the majority and protected classes get a particular outcome. When that difference is small, the classifier is said to have 'statistical parity,' *i.e.*, to conform to this notion of fairness." *Cf.* Gal Yona, *A Gentle Introduction to the Discussion on Algorithmic Fairness*, TOWARDS DATA SCI. (Oct. 5, 2017), https://towardsdatascience.com/a-gentle-introduction-to-the-discussion-on-algorithmic-fairness-740bbb469b6 [https://perma.cc/AWT2-HMSH] ("US legal theory uses the 'disparate impact theory' principle: a practice is considered *illegal discrimination* if it has a 'disproportionately adverse' effect on members of a protected group . . . The mathematical equivalence of the disparate impact principle at its most extreme version (allowing no adverse effect on members of the protected group) for binary classification tasks is the Statistical Parity condition: it essentially equalizes the outcomes across the protected and non-protected groups."). For more technical discussions of statistical or demographic parity, and fairness of algorithms, see Jeremy Kun, *One Definition of Algorithmic Fairness: Statistical Parity*, MATH ∩ PROGRAMMING (Oct. 19, 2015), https://jeremykun.com/2015/10/19/one-definition-of-algorithmic-fairness-statistical-parity [https://perma.cc/9Y9V-DVK6]; Simon Prince, *Tutorial #1: Bias and Fairness in AI*, BOREALIS AI (Aug. 19, 2019), https://www.borealisai.com/en/blog/tutorial1-bias-and-fairness-ai [https://perma.cc/P8PV-UFMR].

[163] *See, e.g.*, PARTNERSHIP ON AI, *supra* note 137, at n.6 ("[E]valuation of machine learning models is a complicated and subtle topic which is the subject of active research. In particular, note that inaccuracy can and should be divided into errors of 'Type I' (false positive) and 'Type II' (false negative) – one of which may be more acceptable than the other, depending on the context.").

[164] *See, e.g.*, Kenn So, *A Primer on Fairness*, TOWARDS DATA SCI., https://towardsdatascience.com/artificial-intelligence-fairness-and-tradeoffs-ce11ac284b63 [https://perma.cc/VGN8-NUM6] ("There is no one definition of what is fair. What is considered fair depends on the context."); Louise Mastakis, *What Does a Fair Algorithm Actually Look Like?*, WIRED, https://www.wired.com/story/what-does-a-fair-algorithm-look-like [https://perma.cc/G32Z-MGVT] ("The question of '[w]hat it means for an algorithm to be fair?' does not have a technical answer alone. . . . It matters what social processes are in place around that algorithm."); Jeremy Kun, *What Does It Mean for an Algorithm to Be Fair?*, MATH ∩ PROGRAMMING, https://jeremykun.com/2015/07/13/what-does-it-mean-for-an-algorithm-to-be-fair [https://perma.cc/E2CK-9X6U] ("[T]here is no accepted definition of what it *means* for an algorithm to be fair.") (emphasis in original); Alexandra Ebert, *We Want Fair Algorithms – But How to Define Fairness? (Fairness Series Part 3)*, MOSTLY • AI,

to operationalize and highly context-dependent.[165] Many commentators have noted that it may not be possible to achieve both good predictive accuracy and fairness at the same time,[166] and lawyers and judges may be forced to decide which of these competing values is more important under any given set of circumstances. While many high-level aspirational principles and guidelines have been promulgated for trustworthy or ethical AI, and while they are admirable, many simply cannot be implemented in any practical way.[167] And, it is questionable in the first place whether we want developers making "de-biasing" decisions in the dark. This leaves it up to lawyers and judges to make sure that the correct questions are being asked—for example, whether impact assessments have been performed, how the tool was assessed for bias and by whom, and whether the correct metrics were collected and reported.

Finally, bias arises as a result of the human interpretation of the output of AI systems. All humans have unconscious or implicit biases,[168] such as confirmation bias. *Confirmation bias* is the tendency for humans to search for, interpret, favor, and recall information that confirms their prior beliefs and values;[169] It has a tendency to distort evidence-based decision-making.

---

https://mostly.ai/2020/05/06/we-want-fair-ai-algorithms-but-how-to-define-fairness [https://perma.cc/4FCD-VNMV] ("Fairness is a vastly complex concept and as people tend to have different values their interpretations of fairness differ as well.").

[165] *See* So, *supra* note 164.

[166] Indeed, the Practitioner's Guide to COMPAS Core itself cites to a 2018 study that concluded from "a thorough examination of risk assessment fairness in criminal justice settings" that "[e]xcept in trivial cases, it is impossible to maximize accuracy and fairness at the same time and impossible simultaneously to satisfy all kinds of fairness." *Practitioner's Guide to COMPAS Core*, NORTHPOINTE INC. D/B/A EQUIVANT 1, 19 (2019), https://www.equivant.com/wp-content/uploads/Practitioners-Guide-to-COMPAS-Core-040419.pdf [https://perma.cc/CMG4-R2QA] (quoting Richard Berk et al., *Fairness in Criminal Justice Risk Assessments: The State of the Art*, 50 SOC. METHODS & RES. 1, 1 (2018)); *see also* Katherine B. Forrest, *When AI Tools Are Designed for Accuracy Over Fairness*, N.Y. L.J., https://www.law.com/newyorklawjournal/2020/10/06/when-ai-tools-are-designed-for-accuracy-over-fairness [https://perma.cc/K4KU-VSCL].

[167] For a global inventory of AI Ethics Guidelines, see *AI Ethics Guidelines Global Inventory*, ALGORITHM WATCH, https://inventory.algorithmwatch.org [https://perma.cc/6XEG-W3DX].

[168] *See* Karen Steinhauser, *Everyone Is a Little Bit Biased*, ABA, https://www.americanbar.org/groups/business_law/publications/blt/2020/04/everyone-is-biased [https://perma.cc/P5WR-2KF7]; Keith Payne et al., *How to Think About 'Implicit Bias,'* SCI. AM., https://www.scientificamerican.com/article/how-to-think-about-implicit-bias [https://perma.cc/Y8DH-PQY8]; Perry Hinton, *Implicit Stereotypes and the Predictive Brain: Cognition and Culture in 'Biased' Person Perception*, 3 PALGRAVE COMM., Art. No. 17086 (2017); the interested reader can test their own implicit biases using the Harvard Implicit Association ("HIA") Test. *See* Project Implicit, HARVARD, https://implicit.harvard.edu/implicit [https://perma.cc/LZ7S-RNX7].

[169] *See Confirmation Bias*, WIKIPEDIA, https://en.wikipedia.org/w/index.php?title=Confirmation_bias&oldid=1001137946 [https://perma.cc/2MRU-Z77M]; Bettina J. Casad, *Confirmation Bias*, BRITANNICA, https://www.britannica.com/science/confirmation-bias [https://perma.cc/M5T8-BVZJ].

There are other biases that are more specific to algorithms and their outputs. Berkeley J. Dietvorst and his colleagues at the University of Pennsylvania wrote a seminal paper on *algorithm aversion* showing that, even though in many circumstances automated decision-making systems can more accurately predict the future than human forecasters,[170] when forecasters are given the choice of whether to use a human prediction or an algorithmic one, they tend to favor the former even when they have observed the algorithmic predictor repeatedly outperform the human forecaster.[171] Dietvorst et al. posit that this is because people more quickly lose confidence in algorithms than in humans when they make the same mistakes, holding the algorithms to a higher standard.[172] This phenomenon can be observed, for example, with autonomous vehicles. Even though evidence shows that these vehicles are likely to reduce car accidents by 94%, people continue to fear them because what they remember is Google's relatively limited number of accidents.[173]

On the other side of the coin is the problem of *automation bias*, the tendency for humans to favor results from automated decision-making systems and to ignore or discount contradictory evidence generated separately from such systems, even if it is correct, because they believe that the automated decision-making system is somehow more "trustworthy" or "objective."[174] A classic example of this is the case of three foreign tourists vacationing in Australia who followed the instructions of their GPS system and drove straight into Moreton Bay so far that they were forced to abandon their vehicle in the water.[175] We see both of these tendencies at work with the

---

[170] *See* Dietvorst et al., *supra* note 19, at 123

[171] *See id.*

[172] *See id.*

[173] *See* Teena Maddox, *How Autonomous Vehicles Could Save Over 350k Lives in the US and Millions Worldwide*, ZDNET, https://www.zdnet.com/article/how-autonomous-vehicles-could-save-over-350k-lives-in-the-us-and-millions-worldwide [https://perma.cc/9JLA-EMBX] ("[Department of Transportation ("DOT")] researchers estimate that fully autonomous vehicles, also known as self-driving cars, could reduce fatalities by up to 94% by eliminating those accidents that are due to human error."). *But see* Matthew Hutson, *People Don't Trust Driverless Cars. Researchers Are Trying To Change That*, SCI., https://www.sciencemag.org/news/2017/12/people-don-t-trust-driverless-cars-researchers-are-trying-change [https://perma.cc/YQ6Y-ML3U] ("Unnerved by the idea of not being in control—and by news of semi-AVs that have crashed, in one case killing the owner—many consumers are apprehensive.").

[174] *See Automation Bias*, WIKIPEDIA, https://en.wikipedia.org/w/index.php?title=Automation_bias&oldid=1001875428 [https://perma.cc/UAE5-EYCK]; Mary L. Cummings, *Automation Bias in Intelligent Time Critical Decision-Making Support Systems*, PROC. AM. INST. OF AERONAUTICS & ASTRONAUTICS ("AIAA") 1ST INTELLIGENT SYS. TECH. CONF. (2014).

[175] *See* Hillary Hanson, *GPS Leads Japanese Tourists to Drive into Australian Bay*, HUFFPOST US, https://www.huffpost.com/entry/gps-tourists-australia_n_1363823 [https://perma.cc/D575-HG68]. *See also What Is Automation Bias and How Can You Prevent It*, PA CONSULTING, https://www.paconsulting.com/insights/what-is-automation-bias-how-to-prevent

use of RNA tools: States and judges both under- and over-rely on them.[176] They ignore RNAs in determining treatments for offenders—the very purpose for which they were designed—and rely on them for sentencing—a use for which their own developer expressed concerns.[177]

B. Lack of Robust Testing for Validity and Reliability

A second serious concern with algorithms and their outputs is the lack of proper evaluation of many AI systems commonly used today. Unlike drugs, which must undergo a rigorous testing and approval process under the auspices of the U.S. Food and Drug Administration ("FDA"), algorithms—even those that can have a significant impact on legal and human rights—do not need to undergo any evaluation at all prior to the time that their output is offered into evidence in a civil or criminal trial. And even when testing is performed, it is rarely independent, peer-reviewed, or sufficiently transparent to be properly assessed by those competent to do so. There are no standards for the conduct of AI product testing and many tools that are in use today would not pass muster if they were subjected to the scientific method.

Validity is the quality of being correct or true, in other words, whether and how *accurately* an AI system measures (*i.e.*, classifies or predicts) what it is intended to measure.[178] *Reliability* refers to the *consistency* of the output of an AI system; that is, whether the same (or a highly correlated) result is obtained under the same set of circumstances.[179] Both need to be measured and both need to exist for an AI system to be trustworthy. As mentioned with respect to COMPAS, focus on overall "accuracy,"[180] at the expense of

[https://perma.cc/S5Z7-3CLX] ("This sort of thing happens so often in Death Valley, California, that the local rangers have coined the term 'death by GPS.'").

[176] *See* Rhys Dipshan, *Judges May Be Using Risk Assessments Too Much—and Too Little*, LEGALTECH NEWS, https://www.law.com/legaltechnews/2020/07/16/judges-may-be-using-risk-assessments-too-much-and-too-little [https://perma.cc/CUT7-HMTF].

[177] *See id.*; *See also* Angwin et al, *supra* note 134 ("I didn't design this software to be used in sentencing. . . . But as time went on, I started realizing that so many decisions are made, you know, in the courts. So I gradually softened on whether this could be used in the courts or not.").

[178] *See* Roberta Heale & Alison Twycross, *Validity and Reliability in Quantitative Studies*, 18 EVID.-BASED NURS. 66 (July 15, 2015).

[179] *See id.*

[180] According to *Pro Publica's* analysis, COMPAS' predictive validity is at best moderate. The score has proved remarkably unreliable in forecasting violent crime: Only 20% of the people predicted to commit violent crimes in next two years went on to do so. When a full range of crimes were considered—including misdemeanors and driving with an expired license—of those deemed likely to re-offend, only 61% were arrested for a subsequent crime within the next two years. *See* Angwin et al., *supra* note 134. There are others, however, who have criticized *Pro Publica's* findings. *See, e.g.*, Flores et al., *supra* note 143

measures that illuminate false-positive and false-negative errors,[181] and other metrics, can mislead users about the quality of the classifications or predictions made by an AI system.[182] As of 2016, when the *Pro Publica* piece was written, even though COMPAS was being used in connection with sentencing, it had never been tested by the U.S. Sentencing Commission.[183] While the tool was developed using a nation-wide training sample, it was not always tested using a sample of local offenders before it was applied such that there was a reason to believe that the training set was reflective of the population on which the algorithm would be used.[184] Since the publication of

[181] "A false positive is an error in binary classification in which a test result incorrectly indicates the presence of a condition such as a disease when the disease is not present, while a false negative is the opposite error where the test result incorrectly fails to indicate the presence of a condition when it is present. These are the two kinds of errors in a binary test, in contrast to the two kinds of correct result (a true positive and a true negative)." *False Positives and False Negatives*, WIKIPEDIA, https://en.wikipedia.org/w/index.php?title=False_positives_and_false_negatives&oldid=1001661831 [https://perma.cc/J2F3-AQT9]. In statistical hypothesis testing, these are typically referred to as "Type I" and "Type II" errors, respectively. *See id*.; *See also supra* note 162.

[182] *See* Jason Brownlee, *Classification Accuracy Is Not Enough: More Performance Measures You Can Use*, MACHINE LEARNING MASTERY (Mar. 21, 2014), https://machinelearningmastery.com/classification-accuracy-is-not-enough-more-performance-measures-you-can-use [https://perma.cc/GM8H-BER5]. This problem is sometimes referred to as the "accuracy paradox." *See Accuracy Paradox*, WIKIPEDIA, https://en.wikipedia.org/w/index.php?title=Accuracy_paradox&oldid=979551882 [https://perma.cc/Q7XA-XGJB] ("The accuracy paradox is the paradoxical finding that accuracy is not a good metric for predictive models when classifying in predictive analytics. This is because a simple model may have a high level of accuracy but be too crude to be useful. For example, if the incidence of category A is dominant, being found in 99% of cases, then predicting that *every* case is category A will have an accuracy of 99%. Precision [*i.e.*, the proportion of cases predicted to be in category A that are actually in category A] and recall [*i.e.*, the proportion of actual cases in category A that are correctly predicted to be in category A] are better measures in such cases. The underlying issue is that there is a class imbalance between the positive class and the negative class," which causes accuracy to be a misleading measure) (emphasis in original). For definitions of "precision" and "recall" in the context of information retrieval, see Maura R. Grossman & Gordon V. Cormack, *The Grossman-Cormack Glossary of Technology-Assisted Review*, 7 FED. CTS. L. REV. 1, 25, 27 (2013) (Precision is "[t]he fraction of Documents identified as Relevant by a search or review effort, that are in fact Relevant;" Recall is "[t]he Fraction of Relevant Documents that are identified as Relevant by a search or review effort."). *See also Precision and Recall*, WIKIPEDIA, https://en.wikipedia.org/w/index.php?title=Precision_and_recall&oldid=1001750137 [https://perma.cc/R59H-DTYB]. With respect to COMPAS specifically, Dr. Jennifer Skeem, Professor of Public Policy at the University of California at Berkeley notes "If you try to equalize false positive rates [between Black and White people], you may find that your calibration suffers, you're going to misclassify people in terms of their likelihood of reoffending. But if you have really good calibration, you're going to have unbalanced error rates, and that's really the conundrum." *See* Rhys Dipshan & Victoria Hudgens, *Risk Assessment Tools Aren't Immune from Systematic Bias. So Why Use Them?*, LEGALTECH NEWS (July 17, 2021), https://www.law.com/legaltechnews/2020/07/17/risk-assessment-tools-arent-immune-from-systemic-bias-so-why-use-them [https://perma.cc/BKK4-64QT].

[183] *See* Angwin et al., *supra* note 134.

[184] *See* Wisconsin v. Loomis, 371 Wis. 2d 235 (2016), *cert. denied*, 137 S. Ct. 2290 (2017) ¶ 27 (citing to expert testimony opining that "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. . .

the *Pro Publica* article, there has been greater testing and evaluation of RNAs, but some still question whether there has been enough, and whether they should ever have been deployed without stringent prior validation.[185]

.”). *See also* Rhys Dipshan, *Same Score, Different Impact: States Can Decide Who Assessment Tech Deems ‘High Risk,’* LEGALTECH NEWS (July 15, 2020), https://www.law.com/legaltechnews/2020/07/15/same-score-different-impact-states-can-decide-whom-assessment-tech-deems-high-risk [https://perma.cc/3CBT-D9WY] (“Decisions about risk thresholds . . . have to be made for each criminal justice population. After all, risk scores and their related failure rates are specific to particular populations, not just within a jurisdiction, but within different parts of the criminal justice system as well. . . . Risk factors are changed to ensure a tool accounts for a locality’s specific characteristics.”); Rhys Dipshan et al., *States vs. Vendors: Are Some Risk Assessment Tools Better Than Others?*, LEGALTECH NEWS (July 14, 2020), https://www.law.com/legaltechnews/2020/07/14/states-vs-vendors-are-some-risk-assessment-tools-better-than-others [https://perma.cc/J43F-E4AM] (“some states and jurisdictions choose to build their own tool[s] . . . because of the notion that developing and validating an instrument for their specific population will be more accurate than validating one originally built for another population. . . . ‘You get better results if you develop your instrument and test it on your own population.’”); PARTNERSHIP ON AI, *supra* note 137 (“[V]alidating a tool in one context says little about whether that tool is valid in another context. . . . [A] risk assessment might predict future arrests quite well . . . in one jurisdiction, but not another.”; “Given that validity often depends on local context to ensure a tool’s utility, where possible, the data . . . should be collected on a jurisdiction-by-jurisdiction basis in order to capture significant differences in geography, transportation, and local procedure[s]. . . .”).

[185] *See* Alex Chohlas-Wood, *Understanding Risk Assessment Instruments in Criminal Justice*, BROOKINGS INST. (June 19, 2020), https://www.brookings.edu/research/understanding-risk-assessment-instruments-in-criminal-justice [https://perma.cc/XBE6-2QG7] (“Though many studies have simulated the impact of RAIs [risk assessment instruments], research on their real-world use is limited.”; “Finally—and perhaps most important—algorithms should be evaluated as they are implemented. It is possible that participants in any complicated system will react in unexpected ways to a new policy (e.g., by selectively using RAI predictions to penalize communities of color). Given this risk, policymakers should carefully monitor behavior and outcomes as each new algorithm is introduced and should continue routine monitoring once a program is established to understand longer-term effects. These studies will ultimately be key in assessing whether algorithmic innovations generate the impacts they aspire to achieve.”); PARTNERSHIP ON AI, *supra* note 137, at 3, 11, 15, 33 (“[The Partnership] has outlined ten largely unfulfilled requirements that jurisdictions should weigh heavily and address before further use of risk assessment tools in the criminal justice system. . . . Challenges in using these tools [include] . . . [c]oncerns about the validity, accuracy, and bias in the tools themselves. . . .”; “An overwhelming majority of the Partnership’s consulted experts agreed that current risk assessment tools are not ready for use in helping to make decisions to detain criminal defendants without the use of an individualized hearing.”; “In combination with concerns about accuracy and validity, [challenges with bias] present significant concern for the use of risk assessment tools in criminal justice domains.”; “One approach is for jurisdictions to cease using the tools in decisions to detain individuals until they can be shown to have overcome the numerous validity, bias, transparency, procedural, and governance problems that currently beset them.”); Alexander Babuta & Marion Oswald, *Data Analytics and Algorithmic Bias in Policing* 1, 7 (RUSI 2019) (“Independent, methodologically robust evaluation of trials is essential to demonstrate the accuracy and effectiveness of a particular tool or method. If such evaluation does not demonstrate the tool’s effectiveness and proportionality, continued use would raise significant legal concerns regarding whether use of the tool was justified to fulfil a particular policing function, requiring the police force to review its design and operational use.”). Babuta and Oswald’s report focuses on both predictive crime mapping, as well as risk assessment, both of which are referred to as forms of “predicting policing.” *Id.* at 4. *See also Dipshan*, *supra* note 184 (“[V]alidations don’t always happen as expected. Some jurisdictions that lack criminal justice outcome data, for instance, will implement a third-party tool without first testing it on their own population. . . . States will also differ in how often they revalidate

While the issue of evaluation is addressed in more detail in section VIII below, discussing the factors that ought to be considered by lawyers and judges when the results of an AI analysis is being offered into evidence in a civil or criminal trial, it is imperative that both groups understand the scientific method and statistical measurement so they can properly assess the validity, reliability, and error rates of AI systems. Often, they lack the training to do so.

C. Failure to Monitor for Function Creep

Closely related to the problem of inadequate testing and evaluation is the problem of *function creep*, which refers to the gradual widening of the use of a technology or system beyond the use for which it was originally intended, often, but not always, without validation and/or leading to an invasion of privacy.[186] COMPAS, again, provides a good example of this. As explained above, COMPAS was originally designed for assessing the

tools to confirm the instruments still work as intended, a necessity given demographic changes and new research findings. While some revalidations are required every few years by law in some states, in others, their timing can depend as much on available resources as need."); Stephanie LaCambra et al., *Recidivism Risk Assessments Won't Fix the Criminal Justice System*, ELECTRONIC FRONTIER FOUND., (Dec. 21, 2018), https://www.eff.org/deeplinks/2018/12/recidivism-risk-assessments-wont-fix-criminal-justice-system [https://perma.cc/Y2FC-KTNL] ("Risk assessment tools are often built using incomplete or inaccurate data because the representative dataset needed to correctly predict recidivism simply doesn't exist. There is no reason to believe that the crime data we do have is sufficiently accurate to make reliable predictions."; "Risk assessment tools must be *evaluated by independent scientific researchers*—not the DOJ itself or a private vendor. To the extent Congress intends the law to reduce disparate impacts on protected classes, independent research must verify that the system can accomplish that and not make the problem worse. Those evaluations should be made public.") (emphasis in original); Thomas Douglas et al., *Risk assessment Tools in Criminal Justice and Forensic Psychiatry: The Need for Better Data*, 42 EUR. PSYCHIATRY 134, 134 (May 2017) ("Violence risk assessment tools are increasingly used within criminal justice and forensic psychiatry, however there is little relevant, reliable and unbiased data regarding their predictive accuracy."). The Law Commission of Ontario ("LCO") recently raised the question of whether Canada should impose "a moratorium on algorithmic risk assessments or similar tools in the Canadian criminal justice system," noting that "Many advocates in the United States would answer . . . affirmatively. This belief is based on the many significant and legitimate criticisms of these systems as presently deployed." *See* LCO, *The Rise and Fall of AI and Algorithms in American Criminal Justice*: LESSONS FOR CANADA 1, 41 (2020). For a recent paper discussing three guiding principles—auditability, transparency, and consistency—that should govern the use of RNA tools to help ensure due process for defendants, see John Villasenor & Virginia Foggo, *Artificial Intelligence, Due Process, and Criminal Sentencing*, 2020 MICH. ST. L. REV. 295 (2020).

[186] *See Function Creep*, DICTIONARY.COM, https://www.dictionary.com/browse/function-creep [https://perma.cc/5W77-9W5S]. *See also Function Creep: The Frankenstein of Privacy*, VICTORIA MCINTOSH (Oct. 1, 2018), https://victoriamcintosh.com/function-creep-the-frankenstein-of-privacy [https://perma.cc/KE63-AYJE]. Oddly, while the term is used in hundreds of articles every year, the phenomenon is largely unresearched and there are few, if any, papers written on the phenomenon itself. Bert-Jaap Koops, *The Concept of Function Creep*, 13 LAW, INNOVATION, & TECH. 29, 30 (2021). "What distinguishes function-creep from . . . innovat[ion] . . . [is that it] denotes some qualitative change [in functionality] . . . that causes concern not only . . . because of the change itself, but also because the change is insufficiently acknowledged as transformative and in need of discussion." *Id.* at 53–55.

treatment needs of offenders, but its use morphed from that to pre-trial release and bail decisions, and from there to sentencing, despite its lack of validation for the additional purposes.[187]

Another concern relates to full-body security scanners used at airports and court houses. While the U.S. Transportation Security Administration ("TSA") claims that its equipment is configured so that images cannot be recorded, it nonetheless requires that all airport body scanners that it purchases have a hard drive and Internet connectivity so that they are able to store and transmit images for the purposes of "testing, training, and evaluation."[188] In 2010, the U.S. Marshals Service acknowledged that it surreptitiously recorded tens of thousands of images at a single Florida checkpoint and that the machine it used could even be operated remotely.[189] The purposes for which this data was collected remains unclear.

A recent example of function creep that implicates AI is Services Australia's use of the country's national facial biometrics database—developed for a different purpose—to confirm the identities of people who had their self-identifying documents ("IDs") destroyed as a result of catastrophic summer bushfires and were in need of disaster relief because of displacement.[190] While arguably a laudable application, and while the individuals involved were asked to provide their consent to the process, the Department of Home Affairs provided little detail about how the service was deployed and how it might be used in the future.[191] Apparently, in this case, a webcam setup was used to capture the facial images of those who lost their

---

[187] "Most modern risk tools were originally designed to provide judges with insight into the types of treatment that an individual might need—from drug treatment to mental health counseling." Angwin et al., *supra* note 134. COMPAS's developer himself "testified that he didn't design his software to be used in sentencing. 'I wanted to stay away from the courts . . . [b]ut as time went on I started realizing that so many decisions are made, you know, in the courts. So, I gradually softened on whether this could be used in the courts or not'. . . . Still, . . . 'I don't like the idea myself of COMPAS being the sole evidence that a decision would be based upon.'" *Id*.; *See also* PARTNERSHIP ON AI, *supra* note 137, at 22 note 42 ("Notably, part of the holding in Loomis, mandated a disclosure in any Presentence Investigation Report that COMPAS risk assessment information 'was not developed for use at sentencing, but was intended for use by the Department of Corrections in making determinations regarding treatment, supervision, and parole.'").

[188] Declan McCullagh, *Feds Found Storing Checkpoint Body Image Scan Images*, CBS NEWS (Aug. 4, 2010, 10:35 AM), https://www.cbsnews.com/news/feds-found-storing-checkpoint-body-scan-images [https://perma.cc/5HM7-UANJ].

[189] *See id.*

[190] *See* Justin Hendry, *Services Australia Put Face Matching to Work for Bushfire Relief Payments*, ITNEWS (June 5, 2020, 11:50 AM), https://www.itnews.com.au/news/services-australia-put-face-matching-to-work-for-bushfire-relief-payments-548978 [https://perma.cc/96T4-VFMP]; *see also* Marie Johnson, *Face Recognition, Function Creep and Democracy*, INNOVATIONAUS (June 9, 2020), https://www.innovationaus.com/face-recognition-function-creep-and-democracy [https://perma.cc/MMS8-UWFR].

[191] *See* Hendry, *supra* note 190.

IDs and sought disaster relief, and those photos were then matched to photos from passports, visas, and driver's licenses.[192] Even when the repurposing appears to be benign, lawyers and judges need to ensure that AI tools are being used for their intended purpose and that any expansion in their use is lawful and supported by empirical evidence.

As seen from the examples above, function creep can easily bleed into invasions of privacy, our next topic.

D. Failure to Ensure Data Privacy and Data Protection

It has been said that data is the new oil.[193] Supervised machine-learning algorithms, particularly those that employ deep learning, require massive amounts of labeled data to function. Where does this data come from? Sources include Internet searches and clicks, buying habits, and lifestyle and behavioral data gathered from public records, social network usage, mobile phones, video surveillance systems, sensors, and, more recently, the Internet of Things ("IoT"). Organizations analyze this information to classify individuals into different groups, often by using algorithms to identify correlations between different characteristics or behaviors taken from different data sets to create profiles about individuals. But, as most of us learned in grade school, "correlation does not imply causation."[194] That adage is often forgotten when it comes to AI applications.

"Big data" refers to the ways that organizations, including both private business and government, combine diverse datasets and then use statistics and other data-mining techniques to extract otherwise hidden information.

---

[192] *Id.*

[193] *See, e.g.*, *The World's Most Valuable Resource Is No Longer Oil, But Data*, THE ECONOMIST (May 6, 2017), https://www.economist.com/leaders/2017/05/06/the-worlds-most-valuable-resource-is-no-longer-oil-but-data [https://perma.cc/2K64-CK7T]; *see also* Mitt Rosebrough, *Is Data Really 'The New Oil'?*, KENWAY CONSULTING (Apr. 27, 2020), https://www.kenwayconsulting.com/blog/data-is-the-new-oil [https://perma.cc/3DUR-QWKK]; *see also* Kiran Bhageshpur, *Data Is The New Oil - - And That's A Good Thing*, FORBES (Nov. 15, 2019, 8:15 AM), https://www.forbes.com/sites/forbestechcouncil/2019/11/15/data-is-the-new-oil-and-thats-a-good-thing/?sh=3a287d6c7304 [https://perma.cc/HKK9-XMJ5]; Joris Toonders, *Data Is the New Oil of the Digital Economy*, WIRED (July 2014), https://www.wired.com/insights/2014/07/data-new-oil-digital-economy [https://perma.cc/AN44-ZKZB].

[194] *See, e.g.*, Seema Singh, *Why Correlation Does Not Imply Causation?*, TOWARDS DATA SCI. (Aug. 24, 2018), https://towardsdatascience.com/why-correlation-does-not-imply-causation-5b99790df07e [https://perma.cc/HZ8W-M9HQ]; Nathan Green, *Correlation Is Not Causation*, THE GUARDIAN (Jan. 6, 2012), https://www.theguardian.com/science/blog/2012/jan/06/correlation-causation [https://perma.cc/7MP5-H4ZZ]; *Correlation Does Not Imply Causation*, WIKIPEDIA, https://en.wikipedia.org/w/index.php?title=Correlation_does_not_imply_causation&oldid=1001822743 [https://perma.cc/5HH6-35PY]. At least one commentator believes that correlation is really all that matters in the age of big data. *See* Anderson, *supra* note 93 ("Petabytes allow us to say: 'Correlation is enough.'"; "Correlation supersedes causation, and science can advance even without coherent models, unified theories, or really any mechanistic explanation at all.").

These approaches raise serious privacy and fairness concerns. The profiles that result from these methods are then used to discover information about an individual's characteristics or preferences, to predict their future behavior, and/or to make decisions about them, often without appropriate disclosure. A good example of the danger of these systems is the Chinese Communist Party's (and similar private Chinese organizations') use of social credit scores for judging citizens' trustworthiness.[195] The data used by such rating systems can include anything from not paying a loan or a fine on time, to spending "frivolously," to misbehaving on a train by playing music too loud, to lighting up in a smoke-free zone, to walking a dog off-leash, to standing up a taxi, to driving through a red light, to spreading "fake news," to losing a defamation case against someone, to spending too much time playing video games.[196] One city, Rongcheng, gives all of its residents 1,000 points to start and deducts from these for "bad" behavior, such as traffic violations or stealing electricity, or adds points for "good behavior," such as donating to charity.[197] Even dating sites like Baihe allow potential partners to not only assess each other's looks, but also their social credit scores.[198] The consequences of low scores can be more serious than simply losing a date; they include the loss of educational and employment opportunities, as well as transportation restrictions (*e.g.*, the inability to purchase business class train tickets or to lodge at certain hotels).[199] Those with high scores get perks such as discounts on utility bills, the ability to book hotel rooms without deposits, and faster application processing to travel abroad.[200] There is little, if any, opportunity to challenge one's score.

Corporations in the United States are increasingly using AI to divide consumers along class lines. In fact, financial services institutions have used

---

[195] *See* Amanda Lee, *What Is China's Social Credit System and Why Is It Controversial?*, SOUTH CHINA MORNING POST (Aug. 9, 2020, 12:00 PM), https://www.scmp.com/economy/china-economy/article/3096090/what-chinas-social-credit-system-and-why-it-controversial [https://perma.cc/4AF5-MLMW]; Nicole Kobie, *The Complicated Truth about China's Social Credit System*, WIRED (July 6, 2019, 12:00 PM), https://www.wired.co.uk/article/china-social-credit-system-explained [https://perma.cc/7NAX-Q9SF].

[196] Nadre Nittle, *Spend 'Frivolously' and Be Penalized under China's New Social Credit System*, VOX (Nov. 2, 2018, 6:50 PM), https://www.vox.com/the-goods/2018/11/2/18057450/china-social-credit-score-spend-frivolously-video-games [https://perma.cc/B9AW-P48C].

[197] *See* Kobie, *supra* note 195.

[198] *See id.*; *see also* Celia Hatton, *China 'Social Credit': Beijing Sets Up Huge System*, BBC NEWS (Oct. 26, 2015), https://www.bbc.com/news/world-asia-china-34592186 [https://perma.cc/R5UN-2LLE] ("China's biggest matchmaking service, Baihe, has teamed up with Sesame [Credit, the financial wing of Alibaba] to promote clients with good credit scores, giving them prominent spots on the company's website. 'A person's appearance is very important,' explains Baihe's vice-president Zhuan Yirong. 'But it's more important to be able to make a living. Your partner's fortune guarantees a comfortable life.'").

[199] *See* Kobie, *supra* note 195; Nittle, *supra* note 196.

[200] *See* Nittle, *supra* note 196.

algorithms for these purposes for decades. The idea that a person's financial (*i.e.*, debt and credit) history and other characteristics reflect trustworthiness and reliability has long influenced employment and other decisions and can increasingly be expected to do so as AI continues to proliferate.

The collection of consumer data is often accomplished without meaningful informed consent. In circumstances where consent has been given, the subsequent sale of data to others may be inconsistent with reasonable expectations about its use, especially when it is being repurposed in unexpected ways to draw conclusions about individuals, with potentially harmful effects. Fairness dictates transparency in how data will be collected and used, and by whom; how long it will be retained; and the potential negative impact of the intended use of the data on the individual. Concerns about these issues have severely impeded the acceptance of contact tracing applications developed for COVID-19.[201]

Along with the collection of vast amounts of data for AI algorithms come the increasing risks of privacy violations and data breach. There is a tension between more accurate predictions based on larger, more representative data sets, and encroachment on privacy. Many commentators have scoffed that privacy is a dead letter.[202] They may be right. An early example of the illusion of anonymity occurred in 2006 when AOL released a large amount of data to the public showing user search requests. It turned out that some users could be identified by name based on their search queries.[203] This was followed by a scandal in 2008, in which two computer

---

[201] *See* Kayla Hui, *Privacy Concerns Continue to Prevent Contract Tracing App Use*, VERYWELLHEALTH (Nov. 28, 2020), https://www.verywellhealth.com/family-tension-privacy-contact-tracing-app-covid-19-5088798 [https://perma.cc/6ZB6-4WCS]; Alejandro De La Garza, *Contract Tracing Apps Were Big Tech's Best Idea for Fighting COVID-19. Why Haven't They Helped?*, TIME MAG. (Nov. 10, 2020, 7:00 AM), https://time.com/5905772/covid-19-contact-tracing-apps [https://perma.cc/2QP9-PY5T]; Sarah Kreps et al., *Contract-tracing Apps Face Serious Adoption Obstacles*, BROOKINGS INST. TECHSTREAM (May 20, 2020), https://www.brookings.edu/techstream/contact-tracing-apps-face-serious-adoption-obstacles [https://perma.cc/QQB4-VZ6H].

[202] Sun Microsystems' CEO is claimed to have said in an interview with reporters and industry analysts "You have zero privacy anyway. Get over it!" Polly Sprenger, *Sun on Privacy: 'Get over it,'* WIRED (Jan. 26, 1999, 12:00 AM), https://www.wired.com/1999/01/sun-on-privacy-get-over-it [https://perma.cc/K3SZ-PTKY]. *See also*, *e.g.*, Summer Lewis, *Is Privacy a Dead Letter?*, IP OSGOODE (Oct. 30, 2019), https://www.iposgoode.ca/2019/10/is-privacy-a-dead-letter [https://perma.cc/9TPC-BEGX]; Henry Mance, *Is Privacy Dead?*, FIN. TIMES (July 19, 2019), https://www.ft.com/content/c4288d72-a7d0-11e9-984c-fac8325aaa04 [https://perma.cc/F32L-QBN3]; Judith Rauhofer, *Privacy Is Dead, Get Over It! Information Privacy and the Dream of Risk-Free Society*, 17 INFO. & COMM. TECH. L. 185 (2008).

[203] *See AOL Search Data Leak*, WIKIPEDIA, https://en.wikipedia.org/w/index.php?title=AOL_search_data_leak&oldid=970872440 [https://perma.cc/RR8X-MVPB]. *See also* Michael Barbaro & Tom Zeller Jr., *A Face Is Exposed for AOL*

scientists were able to re-identify Netflix users in a database of customer records that Netflix had made available to researchers in a competition intended to improve the company's recommender system,[204] and another in 2013, by a study in which a computer scientist at Harvard was able to re-identify patients by name in a supposedly anonymized data set made publicly available by Washington State.[205] In a 2015 study entitled *Unique in the Shopping Mall: On the Reidentifiability of Credit Card Metadata*, researchers analyzed credit card transactions made by 1.1 million people in 10,000 stores over a three-month period.[206] The data contained basic information about the date of each transaction, the amount charged, and the name of the store.[207] Although the data had been anonymized by removing personal information such as names and account numbers, the uniqueness of people's behavior made it easy to single them out.[208] It turned out that by knowing just four pieces of information, the researchers were able to re-identify 90 % of the shoppers as unique individuals, and to uncover their records.[209] By combining their "unicity" with publicly available information, such as posts on social media, it was possible to re-identify many of the individuals by name.[210] Since then, a reporter at Gawker was able to re-identify celebrities by name in an anonymized database of taxi records made public by New York City's taxi and Limousine Commission.[211] These examples call into question the standard approaches many companies, hospitals, government agencies, and other organizations use to anonymize

Searcher No. 4417749, N.Y. TIMES (Aug. 9, 2006), https://www.nytimes.com/2006/08/09/technology/09aol.html [https://perma.cc/X9TT-3LHT].

[204] Arvind Narayanan & Vitaly Shmatikov, *Robust De-anonymization of Large Sparse Data Sets*, PROC. OF THE IEEE SYMP. ON SECURITY AND PRIV. PROC. 111–25 (2008).

[205] LATANYA SWEENEY, MATCHING KNOWN PATIENTS TO HEALTH RECORDS IN WASHINGTON STATE DATA (SSRN 2013), https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2289850 [https://perma.cc/T4D6-HWVS].

[206] Natasha Singer, *With a Few Bits of Data, Researchers Identify 'Anonymous' People*, N. Y. TIMES: BITS (Jan. 29, 2015, 2:01 PM), https://bits.blogs.nytimes.com/2015/01/29/with-a-few-bits-of-data-researchers-identify-anonymous-people [https://perma.cc/3Q2Z-7EMJ]; Yves-Alexandre de Montjoye et al., *Unique in the Shopping Mall: On the Reidentifiability of Credit Card Metadata*, 347 SCI. 536 (2015). *But see* David Sánchez et al., *Comment on* "*Unique in the Shopping Mall: On the Reidentifiability of Credit Card Metadata,"* 351 SCI. 1274 (2016) (arguing that "anonymization can be performed by techniques well established in the literature").

[207] *See* Montjoye et al., *supra* note 206, at 537–38.

[208] *See id.* at 538–39.

[209] *See* Singer, *supra* note 206.

[210] *Id.* "Unicity" refers to "the quality or state of being unique of its kind." *Unicity*, MERRIAM WEBSTER.COM DICTIONARY, https://www.merriam-webster.com/dictionary/unicity [https://perma.cc/ZE2C-Z9P2].

[211] *See* J.K. Trotter, *Public NYC Taxicab Database Lets You See How Celebrities Tip*, GAWKER (Oct. 23, 2014, 12:00 PM), https://gawker.com/the-public-nyc-taxicab-database-that-accidentally-track-1646724546 [https://perma.cc/BHX2-DCY9].

sensitive information. This problem will only increase as AI gets better at crunching information from disparate data sources.

There is presently very little law in the United States about how aggregated data and profiling may be used. This is not the case in the European Union ("EU"), where in 2018, the General Data Protection Regulation ("GDPR") was enacted.[212] The GDPR provides certain protections for the "processing" of personal data of data subjects in the EU.[213] While an extended discussion of the GDPR is beyond the scope of this paper, we will briefly mention a few protections that relate to "big data" and the use of AI.

Article 7 of the GDPR address the provisions relating to consent, which must be voluntary, freely given, informed, and unambiguous,[214] more so than those terms are typically understood in the United States. Consent must be obtained for the specific purpose for which the data will be used, so there cannot be undisclosed repurposing of the data.[215] Relatedly, Article 5(1)(b), which addresses how personal data may be processed (*i.e.*, used), requires that personal data must be "collected for specified, explicit and legitimate purposes and not further processed in a manner that is incompatible with those purposes."[216] This is referred to as the "purpose limitation."[217] Article 5(1)(c) further states that personal data must be "limited to what is necessary in relation to the purposes for which they are processed."[218] This is referred to as "data minimization."[219]

---

[212] *The General Data Protection Regulation Applies in All Member States from 25 May 2018*, EUR-LEX (May 24, 2018), https://eur-lex.europa.eu/content/news/general-data-protection-regulation-GDPR-applies-from-25-May-2018.html [https://perma.cc/L479-VS62]. Canada has recently proposed similar legislation. *See* News Release, Innovation, Sci. and Econ. Dev. Canada, *New Proposed Law to Better Protect Canadians' Privacy and Increase Their Control Over Their Data and Personal Information*, CANADA.CA (Nov. 17, 2020), https://www.canada.ca/en/innovation-science-economic-development/news/2020/11/new-proposed-law-to-better-protect-canadians-privacy-and-increase-their-control-over-their-data-and-personal-information.html [https://perma.cc/RR8W-MDVD].

[213] *See* Regulation 2016/679 of the European Parliament and of the Council of 27 April 2016, art. 2(1), 2016 O.J. (L 119) 1, 32 [hereinafter "GDPR"] ("This Regulation applies to the processing of personal data wholly or partly by automated means and to the processing other than by automated means of personal data which form part of a filing system or are intended to form part of a filing system."). The GDPR defines "processing" as "any operation or set of operations which is performed on personal data or on sets of personal data, whether or not by automated means, such as collection, recording, organisation, structuring, storage, adaptation or alteration, retrieval, consultation, use, disclosure by transmission, dissemination or otherwise making available, alignment or combination, restriction, erasure or destruction." *Id.* at art. 4(2).

[214] See *id.* at art. 7; *see also id.* at 6 (Recital 32).

[215] *See id.* at art. 7(2); *see also id.* at 6 (Recital 32).

[216] *Id.* at art. 5(1)(b); *see also id. at* 7, 9–10 (Recitals 39 and 50).

[217] *See id.* at art. 5(1)(b).

[218] *See id.* at art. 5(1)(c); *see also id.* at 7 (Recital 39).

[219] *See id.* at art. 5(1)(c).

The GDPR also provides for the "Right to Erasure" (a/k/a the "Right to be Forgotten") and Article 7 requires that consent be revocable.[220] Article 17(1) provides the data subject with the right to demand the erasure of personal data about themselves without undue delay, and that the data controller must comply when the data subject withdraws their consent.[221] This may be virtually impossible to accomplish once that data has been ingested into a machine-learning algorithm.

More important for present purposes is Article 22, which prohibits automated decision-making in certain circumstances.[222] Article 22(1) provides that a data subject may not be subject to a decision made solely on the basis of automated processing, including profiling, if that decision produces legal or similar effects.[223] Automated decision-making is the process of making a decision solely by automated means, without any human involvement.[224] These decisions can be based on factual data (*i.e.*, data provided by the data subject or observed about them) as well as digitally created profiles (*i.e.*, derived or inferred data). Examples of automated decisions include an online decision to award credit or a loan, eligibility for social service benefits or the amount of same, recruiting decisions about whether to interview a candidate for a position based on an automated analysis of their résumé, or decisions about providing a medical treatment to patients based on predictions about the likelihood of success given the presence or absence of certain group characteristics. The GDPR restricts only certain, solely automated decisions: ones that either affect a person's legal status or rights, or those that have a significant effect on an individual's circumstances, reputation, behavior, or choices.[225] The latter is not terribly well defined. The GDPR includes other collateral rights such as the right to request a review if an individual is unhappy with a solely automated decision.[226] The decision maker must be able to show how and why it reached the decision it did, and the system should be able to provide an audit trail

---

[220] *Id.* at art. 7(3), art. 17(1); *see also id.* at 12–13 (Recitals 65 and 66).

[221] *See id.* at art. 17(1); *see also id.* at 12–13 (Recital 65).

[222] *See id.* at art. 22; *see also id.* at 14 (Recital 71).

[223] *See id.* at art. 22(1); *see also id.* at 14 (Recital 71).

[224] Note that there is some variability across jurisdictions concerning the definition of "automated" when it comes to decision-making systems. For example, the Canadian government's definition allows partial human involvement in what is defined as an "automated decision system." *See Directive on Automated Decision-Making: Appendix A - Definitions*, CANADA.CA, https://www.tbs-sct.gc.ca/pol/doc-eng.aspx?id=32592#appA [https://perma.cc/RR7E-5DLP] (An automated decision system "[i]ncludes any technology that either assists or replaces the judgment of the human decision-makers."). Others, however, would refer to that as a "semi-automated system." The preferable term for an AI system that has no human involvement may therefore be an "autonomous decision-making system."

[225] GDPR at art. 22(1); *see also id.* at 14 (Recital 71).

[226] *Id.* at art. 22(3); *see also id.* at 14 (Recital 71).

showing the key decision points that formed the basis for the decision.[227] There must be a process in place for individuals to challenge or appeal the decision, taking into account the factors upon which the original decision was based, as well as any additional evidence the individual can assemble to support their claim.[228] Right now, it is primarily up to lawyers and judges in the United States to provide these kinds of protections to individuals that have been subjected to automated decision-making.[229]

In addition to violations of privacy in connection with personal data, AI itself can be alarmingly intrusive. Recently, one of the authors (Grossman) received the following message: "Hi Maura. I'm Neville, the co-founder of XXXXXXXX. I would like to discuss our remote proctoring features for online assessments & see if this can be useful to you. Our tech comes with face recognition, 2 face detection, mobile & book detection geo tagging and much more!" (company name redacted).[230]

Because most educational instruction has moved online during the COVID-19 pandemic, the use of AI-based surveillance techniques for the purposes of proctoring exams has seen an increase at educational institutions. Proctorio is another fully automated "comprehensive learning integrity tool" used to monitor for cheating during exams.[231] It requires the student to sit in a quiet place without anyone else present in the room, which can disproportionally affect students coming from disadvantaged economic

---

[227] *See id.* at 14 (Recital 71).

[228] *See also id.* at art. 22(3).

[229] Two notable exceptions to this are the Fair Credit Reporting Act ("FCRA"), enacted in 1970, and the Equal Credit Opportunity Act ("ECOA"), enacted in 1974, both of which address automated decision-making in the context of machine-based credit underwriting models. The Federal Trade Commission ("FTC") Act [of 1914] authority to prohibit unfair and deceptive practices has also been used to address consumer injury arising from the use of AI and automated decision-making. *See* Andrew Smith, *Using Artificial Intelligence and Algorithms*, FED. TRADE COMM'N (Apr. 8, 2020, 9:58 AM), https://www.ftc.gov/news-events/blogs/business-blog/2020/04/using-artificial-intelligence-algorithms [https://perma.cc/K9Y8-5Z4V]. In a recent case involving a photo application that the FTC claimed deceived consumers about the use of facial recognition technology and the retention of photos and videos of users who had deactivated their accounts, as part of the proposed settlement with the company, Everalbum, Inc., the company was not only required to "obtain consumers' express consent before using facial recognition technology on their photos and videos," but also to "delete models and algorithms it developed by [impermissibly] using the photos and videos uploaded by its users." Press Release, *Fed. Trade Comm'n, California Company Settles FTC Allegations It Deceived Consumers about use of Facial Recognition in Photo Storage App*, FTC.GOV (Jan. 11, 2021), https://www.ftc.gov/news-events/press-releases/2021/01/california-company-settles-ftc-allegations-it-deceived-consumers [https://perma.cc/UQ7Z-KTFV].

[230] Invitation from Neville Katila, Co-Founder & Director at Eduswitch Solutions Private Limited, to connect on LinkedIn (Sept. 5, 2020) (on file with author Grossman).

[231] *See* PROCTORIO, https://proctorio.com [https://perma.cc/P8E6-KN9T]. For an unvarnished student's take on Proctorio, *see* Cassie Finley (@Angry_Cassie), TWITTER (Sept. 2, 2020, 10:26 PM), https://twitter.com/Angry_Cassie/status/1301360994044850182 [https://perma.cc/D2QT-VV3A].

backgrounds that may only have access to Wi-Fi in public or shared spaces. When registering for the exam, the test-taker must provide a photo ID using the computer's webcam, which will be compared to the test-taker's face on the day of the exam using facial recognition software.[232] The intent behind this process, to make sure that someone else is not taking the exam in place of the test- taker, is not unreasonable. But not everyone looks the same as their photo ID, perhaps because of weight gain or loss, illness, or gender transition. Before the start of the exam, the webcam must slowly be moved around the room to record the test-taker's surroundings,[233] ostensibly to confirm that nearby areas are free of materials that could be used to cheat. But what if the camera records a roommate's illicit paraphernalia or illegal reading materials? Proctorio also records all sounds in the room, flags "suspicious behavior," like taking one's eyes off the screen, and scans for plagiarism.[234] We do not know what other analyses Proctorio performs, and students do not have a choice to refuse such surveillance tools. We can expect to see an expansion in their use into employment and other settings.[235]

E. Lack of Transparency and Explainabilty

One of the most widely proposed solutions to the "black-box"[236] problem of AI is to require transparency and explainability, both in terms of how the AI system works, as well as how it reached its decision (*i.e.*, approved or denied for a loan), classification (*i.e.*, eligible for a prime loan versus a sub-prime loan), or prediction/conclusion (*i.e.*, the user will like the following movies, or the user should make the following grammatical corrections).[237] Cynthia Rudin argues that models for high-stakes decisions

---

[232] *See ID Verification*, PROCTORIO, https://proctorio.com/platform/id-verification [https://perma.cc/24DD-BLJW].

[233] *See, e.g.*, *Online Proctoring*, PROCTORIO, https://proctorio.com/products/online-proctoring [https://perma.cc/5JSC-42VC]; *Desk Scan Setting and Exam Environment under Frequently Asked Questions*, PROCTORIO, https://proctorio.com/frequently-asked-questions [https://perma.cc/W8LT-GJ2W].

[234] *See Behavior under Frequently Asked Questions*, PROCTORIO, https://proctorio.com/frequently-asked-questions [https://perma.cc/YB3F-DUBR]; *Plagiarism*, PROCTORIO, https://proctorio.com/platform/plagiarism [https://perma.cc/H7VM-2PKZ].

[235] Recently, a colleague of author Grossman suggested that these kinds of monitoring tools might be useful to judges and adverse parties to assist them in assessing the credibility of witnesses during online depositions, hearings, and trials necessitated by the COVID-19 pandemic. The colleague had not considered the privacy implications.

[236] A "black box" is "anything that has mysterious or unknown internal functions or mechanisms." *Black Box*, MERRIAM WEBSTER.COM DICTIONARY, https://www.merriam-webster.com/dictionary/black%20box [https://perma.cc/R5CB-E9XD].

[237] *See, e.g.*, Ron Schmelzer, *Towards a More Transparent AI*, FORBES: COGNITIVE WORLD (May 23, 2020, 1:28 PM), https://www.forbes.com/sites/cognitiveworld/2020/05/23/towards-a-more-transparent-ai [https://perma.cc/K8TQ-2G2B]; Greg Satell & Josh Sutton, *We Need AI That Is*

must provide explanations that reveal their inner workings and that algorithms that are inherently black-box should be avoided for such decisions.[238] This remains an area of controversy.

The technical challenge of explaining AI decisions is known as the "interpretability problem,"[239] and an entire domain of research exclusively devoted to this problem has emerged, known as "Explainable AI" ("XAI").[240] Those who advocate for XAI believe that AI can only be trustworthy if it can be explained to humans, although they acknowledge that the level or type of explanation may vary for different applications or users. NIST has outlined four principles of XAI which include (i) explanation—that AI systems deliver accompanying evidence or the reason(s) for all outputs; (ii) meaningful—that AI systems provide explanations that are understandable to individual users; (iii) explanation accuracy—that the explanations correctly reflect the AI system's process for generating the outputs; and (iv) knowledge limits—that the AI system only operates under the conditions for which it was designed or when the system reaches sufficient confidence in its output.[241]

One transparency project, the Defense Advanced Research Project Agency ("DARPA") XAI program, aims to produce "glass-box" models that are explainable to a "human-in-the-loop" without sacrificing AI performance. [242] The term "glass box" has also been used to describe and monitor the inputs and outputs of an AI system with the purpose of verifying

Explainable, Auditable, and Transparent, HARV. BUS. REV. (Oct. 28, 2019), https://hbr.org/2019/10/we-need-ai-that-is-explainable-auditable-and-transparent [https://perma.cc/K4PG-NWQS]; Finale Doshi-Velez & Mason Kortz, *Accountability of AI Under the Law: The Role of Explanation* 11–12 (Berkman Klein Ctr. Working Grp., 2017), https://dash.harvard.edu/bitstream/handle/1/34372584/2017-11_aiexplainability-1.pdf [https://perma.cc/3LN8-M4X9]. For a slightly different take on the issue, see Kartik Hosanagar & Vivian Jair, *We Need Transparency in Algorithms, But Too Much Can Backfire*, HARV. BUS. REV. (July 25, 2018), https://hbr.org/2018/07/we-need-transparency-in-algorithms-but-too-much-can-backfire [https://perma.cc/422H-8ZXB].

[238] *See* Cynthia Rudin, *Stop Explaining Black Box Machine Learning Models for High Stakes Decisions and Use Interpretable Models Instead*, 1 NATURE MACH. INTEL. 206 (May 13, 2019).

[239] *See Explainable Artificial Intelligence*, WIKIPEDIA, https://en.wikipedia.org/w/index.php?title=Explainable_artificial_intelligence&oldid=1002122023 [https://perma.cc/KX52-M2D6].

[240] *See id.*

[241] *See* P. Jonathon Phillips et al., *Four Principles of Explainable Artificial Intelligence* 1–2 (NIST Working Paper No. 8312-draft 2020), https://www.nist.gov/system/files/documents/2020/08/17/NIST%20Explainable%20AI%20Draft%20NISTIR8312%20%281%29.pdf [https://perma.cc/K7M6-9DUM].

[242] WIKIPEDIA, *supra* note 239. *See also* Matt Turek, *Explainable Artificial Intelligence (XAI)*, DARPA, https://www.darpa.mil/program/explainable-artificial-intelligence [https://perma.cc/4QN9-EE8L]; David Gunning & David. W. Aha, *DARPA's Explainable Artificial Intelligence (XAI) Program*, 40 AI MAG. 44 (2019).

the system's adherence to certain social, ethical, and legal values, therefore producing value-based explanations.[243]

The problem with the requirement of transparency is that many modern AI techniques are not explainable because they are naturally opaque. While Decision Trees alone are explainable, when combined into Random Forests (*i.e.*, ensembles of decision trees) they lose a certain degree of interpretability. Unlike code, which can be examined for bugs, it is often not apparent how a machine-learning model has been developed or works, especially when it employs deep learning or neural networks. It may be inexplicable why an algorithm mistakes a 3D-printed turtle for a rifle, or a baseball for an espresso,[244] nor is there typically a way to "bug-fix" a way to the correct model. Developers can only improve the training data, choose different features or parameters to emphasize, and re-assess the output, but otherwise, it may not be obvious why the model is performing poorly. Another challenge is that AI models are not static; they constantly adapt and update over time. While there has been some development of models that are more interpretable, there is typically a tradeoff between accuracy and explainability, so explainable algorithms have not yet achieved widespread adoption, especially when they require a decrease in predictive performance.

It is worth bearing in mind that there is great variability in the situations in which the law requires explanations from humans, such as strict liability, no-fault divorce, national security-related decisions, and jury determinations—where little to no explanation is required—versus direct discrimination, where intent must be proven, or administrative decision-making where, at minimum, the decision must be shown to be non-arbitrary.[245] Even judicial decisions can vary in their need for transparency; decisions on discovery motions are granted considerable deference, while a decision by a judge delivering a criminal sentence must provide a thorough explanation.

Generating explanations is not without cost or effect, and the utility of explanations must be balanced against the time and cost of generating them, including the benefits that are lost by imposing that requirement.[246] By way of example, a doctor who was required to explain every diagnosis and

---

[243] WIKIPEDIA, *supra* note 239. For a more technical discussion *see* Arun Rai, *Explainable AI: from Black Box to Glass Box*, 48 J. ACAD. MARKETING SCI. 137 (2020).

[244] *See* James Vincent, *Google's AI Thinks This Turtle Looks Like a Gun, Which Is a Problem*, VERGE, https://www.theverge.com/2017/11/2/16597276/google-ai-image-attacks-adversarial-turtle-rifle-3d-printed [https://perma.cc/YXS9-MBYA]. *See also* Matthew Hutson, *A Turtle or a Rifle? Hackers Fool AIs into Seeing the Wrong Thing*, SCI., https://www.sciencemag.org/news/2018/07/turtle-or-rifle-hackers-easily-fool-ais-seeing-wrong-thing [https://perma.cc/3UU7-NA65].

[245] *See* Doshi-Velez & Kortz, *supra* note 237, at 5–6.

[246] *Id.* at 3.

treatment plan would likely make fewer mistakes, but would also see far fewer patients because they were busy making patient notes.[247] Moreover, it is well known that humans are notoriously inaccurate when providing post-hoc rationales for their decisions.[248] The need to provide explanations can also impact the decision-maker's choices in the same way that "observed particles behave differently."[249] Finally, it has also been shown that access to an explanation can actually decrease users' trust in some decisions.[250]

Some commentators have argued that there is no real difference between inexplicable AI systems and medications where the neurobiological mechanism through which the drugs operate is not well understood. They argue that this is why we have the Food and Drug Administration ("FDA") to ensure that appropriate testing is undertaken to ensure that drugs are safe before they are released to the public, and that the same should apply for high-impact AI systems.[251] The bottom line for lawyers and judges is that when an AI system is not transparent or explainable, then ensuring its validity and reliability increase in importance.

The *Loomis* case discussed above in section V highlights a related and critical issue that arises with respect to current AI systems and is increasingly likely to arise in court. Even when the data sources and training data are known, and the features, their weights, and parameter choices can be described, when it comes to litigation, AI providers generally assert that information concerning the data and the algorithms are proprietary trade secrets and refuse to disclose them, thereby impeding the ability to challenge their scientific validity and reliability, and to address the many other questions they raise. This is precisely what happened in the *Loomis* case, where Mr. Loomis challenged the Circuit Court's use of COMPAS at sentencing because it violated his due process rights when it interfered with his right "to be sentenced based upon accurate information, in part because the proprietary nature of COMPAS prevent[ed] him from assessing its

---

[247] *Id.*

[248] *Id.* (citing Richard E. Nisbett & Timothy D. Wilson, *Telling More Than We Can Know: Verbal Reports on Mental Processes*, 84 PSYCH. REV. 231 (1977)).

[249] *Id.* at 3 (citing William F. Messier Jr. et al., *The Effect of Accountability on Judgment: Development of Hypotheses for Auditing; Discussions; Reply*, 11 AUDITING 123 (1992)). In physics, this phenomenon is known as the "observer effect," which is "the disturbance of an observed system by the act of observation." *Observer Effects (Physics)*, WIKIPEDIA, https://en.wikipedia.org/w/index.php?title=Observer_effect_(physics)&oldid=1000916691 [https://perma.cc/BEW2-N7G8].

[250] *See* Hosanagar & Jair, *supra* note 237.

[251] *See* Andrew Tutt, *An FDA for Algorithms*, 69 ADMIN. L. REV. 83 (2017). The FDA even recently published its own paper, see FDA, ARTIFICIAL INTELLIGENCE AND MACHINE LEARNING (AI/ML)-BASED SOFTWARE AS A MEDICAL DEVICE (SAMD) ACTION PLAN (2021).

accuracy."[252] Northpointe, Inc., the developer of COMPAS, considers it to be "a proprietary instrument and a trade secret," and it accordingly declined to disclose how the risk scores were determined or how the factors were weighted.[253] Mr. Loomis argued that because COMPAS's developer would not disclose this information, he was denied information that the Circuit Court considered at his sentencing and therefore the ability to refute it.[254] Mr. Loomis further contended that unless he could review how the factors were weighed and how the risk score was determined, the accuracy of the COMPAS assessment could not be verified.[255] From a technical perspective, these were all valid arguments. Yet, the Wisconsin Supreme Court gave them short shrift, responding that while it agreed that Mr. Loomis could not review and challenge how the COMPAS algorithm calculates risk, he could review and challenge the resulting risk scores themselves.[256] The Court concurred with Mr. Loomis that "the risk scores do not explain how the COMPAS program uses information to calculate the risk scores. However, Northpointe's 2015 Practitioner's Guide . . . explains that the risk scores are based largely on static information (criminal history), with limited use of some dynamic variables (i.e. criminal associates, substance abuse)."[257] Thus, the Court asserted, "to the extent that Loomis's risk assessment is based upon his answers to questions and publicly available data about his criminal history, Loomis had the opportunity to verify that the questions and answers listed on the COMPAS report were accurate."[258] Loomis also "had an opportunity to challenge his risk scores by arguing that other factors or information demonstrate their inaccuracy."[259] Despite citing to studies that have raised questions about the accuracy of COMPAS and its tendency to disproportionally classify minority offenders as higher risk because of factors that may be out of their control, [260] the Court held that the tool could nonetheless be used with appropriate warnings, including that:

(1) the proprietary nature of COMPAS had been invoked to prevent disclosure of information relating to how factors are weighed or how risk scores are to be determined; (2) [COMPAS] compares defendants to a national sample, but no cross-validation study for a Wisconsin population

[252] *See* Wisconsin v. Loomis, 371 Wis. 2d 235 (2016), *cert. denied*, 137 S. Ct. 2290 (2017) ¶¶ 34, 46.

[253] *Id.* ¶ 51.

[254] *See id.*

[255] *See id.* ¶ 52.

[256] *See id.* ¶ 53.

[257] *Id.* ¶ 54.

[258] *Id.* ¶ 55.

[259] *Id.* ¶ 56.

[260] *See id.* ¶¶ 59–64.

has yet been completed; (3)some studies of COMPAS risk assessment scores have raised questions about whether they disproportionately classify minority offenders as having a higher risk of recidivism; and (4) risk assessment tools must be constantly monitored and re-normed for accuracy due to changing populations and subpopulations.[261]

The many risks of AI raise questions as to the advisability of protecting the rights of RNA providers over the rights of criminal defendants.[262] Warnings, alone, do not make up for denying a party's ability to challenge the accuracy of an AI tool. The Court never addresses why a protective order would be insufficient protection for Northpointe, and this question can be expected to be an area that will be highly litigated as we move forward.

F. Lack of Accountability

Another place where we can expect to see significant challenges for lawyers and judges is in the area of accountability of AI, and the legal and regulatory frameworks that surround it. At present, there are relatively few laws or regulations governing AI and automated decision making.[263] The

---

[261] *Id.* ¶ 66.

[262] Indeed, the AI Now Institute (an interdisciplinary research institute affiliated with New York University that is dedicated to understanding the social implications of AI technologies), see AINOW, www.ainowinstitute.org [https://perma.cc/G8LZ-MSPJ], has recommended that "AI companies should waive trade secrecy and other legal claims that stand in the way of accountability in the public sector. Vendors and developers who create AI and automated decision systems for use in government should agree to waive any trade secrecy or other legal claim that inhibits full auditing and understanding of their software. Corporate secrecy laws are a barrier to due process: they contribute to the 'black-box effect' rendering systems opaque and unaccountable, making it hard to assess bias, contest decisions, or remedy errors. Anyone procuring these technologies for use in the public sector should demand that vendors waive these claims before entering into any agreements." MEREDITH WHITTAKER ET AL., AI NOW INST., AI NOW REPORT 2018, 5 (2018), https://ainowinstitute.org/AI_Now_2018_Report.pdf [https://perma.cc/L5U6-8KM].

[263] *See* Mark MacCarthy, *AI Needs More Regulation, Not Less*, BROOKINGS INST. (Mar. 9, 2020), https://www.brookings.edu/research/ai-needs-more-regulation-not-less [https://perma.cc/S747-WHA4]; Devin Coldewey, *AI Desperately Needs Regulation and Public Accountability, Experts Say*, TECHCRUNCH (Dec. 7, 2018) (discussing the AI NOW REPORT 2018, *supra* note 262), https://techcrunch.com/2018/12/07/ai-desperately-needs-regulation-and-public-accountability-experts-say [https://perma.cc/A747-UPPM]. For other views on the regulation of AI, *see generally* Richard Diffenthal et al., *Artificial Intelligence – Time to Get Regulating?*, GLOBAL MEDIA TECH. & COMM. Q. (2018); Oren Etzioni, *How to Regulate Artificial Intelligence*, N.Y. TIMES (Sept. 1, 2017), https://www.nytimes.com/2017/09/01/opinion/artificial-intelligence-regulations-rules.html [https://perma.cc/NWC4-TGD5]; Matthew U. Scherer, *Regulating Artificial Intelligence Systems: Risks, Challenges, Competencies, and Strategies*, 29 HARV. J. L. & TECH. 353 (2016). For a discussion of some of the arguments against the regulation of AI, see Andres Fogg, *Artificial Intelligence Regulation: Let's Not Regulate Mathematics!*, IMPORT.IO (Oct. 13, 2016), https://www.import.io/post/artificial-intelligence-regulation-lets-not-regulate-mathematics [https://perma.cc/ARZ4-ZNAX]. And finally, for useful resources surveying global AI governance and regulation issues, curated by the Multidisciplinary Institute on Artificial Intelligence ("MIAI") at Grenoble Alpes, see *AI Governance and Regulation*, AI-REGULATION, https://ai-regulation.com/ai-governance [https://perma.cc/AL6V-LZEE]. On April 21,

existing frameworks seem ill-prepared to address the unique capabilities and characteristics of AI. For example, in August 2019, Stephen L. Thaler of the Artificial Inventor Project[264] tried to patent two inventions—a food container and a flashing warning light—with the U.K.'s Intellectual Property Office ("UKIPO") and the European Patent Office ("EPO").[265] The inventor on the patent was listed as "DABUS."[266] Both regulators held that while the inventions themselves were patent-worthy, the applications were rejected because the "inventor" was not a natural person (*i.e.*, a human).[267] The UKIPO's decision was upheld by the U.K. High Court in October 2020, and was further appealed to the U.K. Court of Appeal, which, in September 2021, also denied the patent.[268] The result was no different in the United States, where in April 2020, the U.S. Patent and Trademark Office ("USPTO") likewise ruled that AI systems cannot be credited as an inventor in a patent, stating that: "[U]nder current law, only natural persons may be named as an inventor in a patent application."[269] The USPTO's decision was appealed to

2021, the European Commission released a long-awaited, comprehensive draft regulation on AI. *See Europe Fit for the Digital Age: Commission Proposes New Rules and Actions for Excellence and Trust in Artificial Intelligence*, European Commission Press Release (Apr. 21, 2021), https://ec.europa.eu/commission/presscorner/detail/en/IP_21_1682.

[264] The Artificial Intelligence Project is a "site dedicated to a project seeking intellectual property rights for the autonomous output of artificial intelligence." *The Artificial Inventor Project*, ARTIFICIAL INVENTOR, https://artificialinventor.com [https://perma.cc/WJE3-DENF].

[265] A copy of the two patent applications, EP3564144 (food container) and EP3563896 (neural flame), can be found here: *Patents and Patent Applications*, ARTIFICIAL INVENTOR, https://artificialinventor.com/patent-applications [https://perma.cc/YK97-3UCJ]. *See also* Amy Sandys, *UK High Court Rejects Idea of Inventor by AI system Dabus*, JUVE PATENT (Oct. 9, 2020), https://www.juve-patent.com/news-and-stories/cases/uk-high-court-rejects-idea-of-invention-by-ai-system-dabus [https://perma.cc/BQ2J-44NU].

[266] Sandys, *supra* note 265. A technical description of DABUS, which stands for "device for the autonomous bootstrapping of unified sentience," *id.*, can be found here: *DABUS Described*, IMAGINATION ENGINES, https://imagination-engines.com/dabus.html [https://perma.cc/6D94-R5K3].

[267] Sandys, *supra* note 265. An appeal of the EPO denial is scheduled to be heard by the EPO Board of Appeal on Dec. 21, 2021. *See* Seiko Hidaka, *Court of Appeal – AI Generated Inventions Denied UK Patent in DABUS Case*, GOWLING WLG (Sept. 23, 2021), https://gowlingwlg.com/en/insights-resources/articles/2021/ai-invention-denied-patent-in-dabus-case [https://perma.cc/G25F-5ZDY].

[268] *See* Cynthia O'Donoghue & Angelika Bialowas, *UK Court of Appeal Rules AI is Not an Inventor*, REEDSMITH TECH. LAW DISPATCH (Sept. 26, 2021), https://www.technologylawdispatch.com/2021/09/in-the-courts/uk-court-of-appeal-rules-ai-is-not-an-inventor [https://perma.cc/UC8E-FMHX].

[269] *Petition Decision: Inventorship Limited to Natural Persons*, USPTO BULLETIN (Apr. 27, 2020), https://content.govdelivery.com/accounts/USPTO/bulletins/287fdc9 [https://perma.cc/37QD-RQG2]; *see also* Jon Porter, *US Patent Office Rules That Artificial Intelligence Cannot Be a Legal Inventor*, VERGE, https://www.theverge.com/2020/4/29/21241251/artificial-intelligence-inventor-united-states-patent-trademark-office-intellectual-property [https://perma.cc/X5UV-Q35K]. For a discussion of the reasons why the U.S. should grant AI inventor status, see Ernest Fok, *Challenging the International Trend: The Case for Artificial Intelligence Inventorship in the United States*, 19 SANTA CLARA J. INT'L LAW 51 (2021).

the U.S District Court for the Eastern District of Virginia. On September 9, 2021, the Court affirmed the decision of the USPTO.[270] However, two months before that, in July of 2021, South Africa was the first country to award DABUS a patent for its AI-generated invention.[271] Australia followed shortly thereafter.[272]

A December 2018 report published by the AI Now Institute makes the point that AI-based tools have been deployed with little regard to their potential negative effects or even sufficient documentation of their positive ones.[273] They lament that untested algorithms are employed in places where they can deeply affect thousands, if not millions of people, with no systems in place to monitor or stop them, other than limited ethical precepts often propounded by the very same companies that created the systems.[274] One particularly egregious example that AI Now cites surfaced in June 2018, immediately after the U.S. Department of Homeland Security implemented a family separation policy that forcibly removed immigrant children from their families, when it was revealed that U.S. Immigration and Customs Enforcement ("ICE") had altered its own risk-assessment algorithm so that it produced only one result: it recommended "detain" for 100% of the immigrants in custody.[275] Another concerning example described a voice recognition system in the U.K. designed to detect immigration fraud, which cancelled thousands of visas resulting in the deportation of people in error.[276] Documents leaked in July 2018, revealed that IBM Watson was rendering

---

[270] Gourdin Sirles & Baldassare Vinti, *Update on Artificial Intelligence: Court Rules That AI Cannot Qualify As "Inventor,"* THE NAT'L L. REV. (Sept. 9, 2021), https://www.natlawreview.com/article/update-artificial-intelligence-court-rules-ai-cannot-qualify-inventor [https://perma.cc/D4NZ-F2VJ].

[271] *See* Sam Udovich, *Recent Developments in Artificial Intelligence and IP Law: South Africa Grants First Patent for AI-Created Invention*, THE NAT'L L. REV. (Aug. 3, 2021), https://www.natlawreview.com/article/recent-developments-artificial-intelligence-and-ip-law-south-africa-grants-world-s [https://perma.cc/ZE2J-4PS2].

[272] *See* John Collins, Natalie Shoolman & Rose Jenkins, *Robots Are Taking Over the Patent World – AI Systems or Devices Can Be "Inventors" Under the Australian Patents Act*, KLUWER PATENT BLOG (Sept. 8, 2021), http://patentblog.kluweriplaw.com/2021/09/08/robots-are-taking-over-the-patent-world-ai-systems-or-devices-can-be-inventors-under-the-australian-patents-act [https://perma.cc/ZK3Y-RYTM].

[273] *See generally* WHITTAKER ET AL., *supra* note 262.

[274] *See generally id.*

[275] *See id.* at 10 (citing Nikhil Sonnad, *US Border Agents Hacked Their 'Risk Assessment' System to Recommend Detention 100% of the Time*, QUARTZ (June 26, 2018), https://qz.com/1314749/us-border-agents-hacked-their-risk-assessment-system-to-recommend-immigrant-detention-every-time [https://perma.cc/28U4-VU8N]).

[276] *See id.* (citing Nikhil Sonnad, *A Flawed Algorithm Led the UK to Deport Thousands of Students*, QUARTZ (May 3, 2018), https://qz.com/1268231/a-toeic-test-led-the-uk-to-deport-thousands-of-students [https://perma.cc/4U5S-8DXA]).

"unsafe and incorrect" cancer treatment recommendations.[277] An investigation conducted in September 2018, unearthed the fact that IBM was also working in concert with the New York City Police Department ("NYPD") to build an "ethnicity-detection" algorithm to search faces based on race, using police camera footage of thousands of people on the streets of New York taken without their knowledge or consent.[278] This is likely just the tip of the iceberg.

On this basis, the AI Now Institute argues, compellingly, that the "frameworks presently governing AI are not capable of ensuring accountability," and that "[a]s the pervasiveness, complexity, and scale of these systems grow, the lack of meaningful accountability and oversight—including basic safeguards of responsibility, liability, and due process—is an increasingly urgent concern."[279] The responsibility for oversight will undoubtedly fall to the legal justice system until there is direct intervention by regulatory agencies.

Finally, AI Now highlights the large gap between those who develop and profit from AI, and those most likely to suffer the consequences of its ill effects.[280] They emphasize several reasons for this discrepancy, including insufficient government oversight, insufficient governance structures within tech companies, a highly concentrated AI sector subject to constant pressure to innovate and commercialize, power asymmetries between the tech companies and the people they serve, and a vast cultural divide between those responsible for technical research and development and the diverse populations on which AI systems are deployed.[281] In an earlier report from September 2018, the AI Now Institute, in collaboration with the Center on Race, Inequality, and the Law and the Electronic Frontier Foundation ("EFF"), discussed the growing number of legal challenges to the use of autonomous systems by government agencies in decisions that affect individual rights, such as Medicaid and disability rights, public teacher

[277] *See id.* (citing Casey Ross & Ike Swetlitz, *IBM's Watson Supercomputer Recommended 'Unsafe and Incorrect' Cancer Treatments, Internal Documents Show*, STAT (July 25, 2018), https://www.statnews.com/wp-content/uploads/2018/09/IBMs-Watson-recommended-unsafe-and-incorrect-cancer-treatments-STAT.pdf [https://perma.cc/H2PN-LG8L]).

[278] See *id.* (citing George Joseph & Kenneth Lipp, *IBM Used NYPD Surveillance Footage to Develop Technology That Lets Police Search By Skin Color*, INTERCEPT (Sept. 6, 2018, 6:00 AM), https://theintercept.com/2018/09/06/nypd-surveillance-camera-skin-tone-search [https://perma.cc/2ZR5-DGWW]).

[279] *Id.* at 7.

[280] *See id.*

[281] *See id.*

employment evaluations, juvenile criminal risk assessment, and criminal DNA analysis.[282]

As the development, commercialization, and use of AI proliferates, so too will questions about how the risks of AI will be apportioned. These questions will be complicated by the vast sea of machine-learning applications in which humans are more or less in- or on-the-loop, and where the systems themselves continuously learn and can act in increasingly unpredictable ways. The present state of the law governing liability for AI systems does not specify who should be held accountable for errors and accidents caused by AI, and under what circumstances. There are many possibilities: the data collector/analyst, the inventor, the designer/developer, the manufacturer, the retailer, the user, the AI itself, some combination of the above, or none at all. The choice of who to hold accountable, and when, is not without consequences for those who can afford to enter the field and for the future of innovation itself.

Some commentators have argued that AI should not be humanized and, from an ethical (and therefore legal) vantage point, should not be treated differently from any other technology, equipment, or tool "we use to extend our own abilities and to accelerate progress on our own goals."[283] It also has been noted by others that concepts from tort and products liability law (*e.g.*, design or manufacturing defect, failure to warn, negligent operation, and strict liability) have been applied and will continue to develop creatively in

[282] *See* AI NOW INST., LITIGATING ALGORITHMS: CHALLENGING GOVERNMENT USE OF ALGORITHMIC DECISION SYSTEMS (2018).

[283] *See* JOANNA J. BRYSON, CLOSE ENGAGEMENTS WITH ARTIFICIAL COMPANIONS: KEY SOCIAL, PSYCHOLOGICAL, ETHICAL AND DESIGN ISSUES 63 (Yorick Wilks ed., 2010). *But see* Ryan Abbott, *The Reasonable Computer: Disrupting the Paradigm of Tort Liability*, 86 GEO. WASH. L. REV. 1, 4–5 (2018) ("This Article employs a functional approach to distinguish an autonomous computer, robot, or machine from an ordinary product. Society's relationship with technology has changed. Computers are no longer just inert tools directed by individuals."). *See also* Iria Giuffrida, *Liability for AI Decision-Making: Some Legal and Ethical Considerations*, 88 FORDHAM L. REV. 439, 440 (2019) (addressing "whether AI merits a new approach to deal with the liability challenges it raises when humans remain 'in' or 'on' the loop."); Karni A. Chagal-Feferkorn, *Am I An Algorithm or a Product? When Products Liability Should Apply to Algorithmic Decision-Makers*, 30 STAN. L. & POL'Y REV. 61, 82–86 (2019) ("Thinking algorithms, despite their nature as information-based and although they may frequently cause damage regardless of a defect, may thus nevertheless be governed by products liability."); Frank H. Easterbrook, *Cyberspace and the Law of the Horse*, U. CHI. LEGAL F. 207 (1996) (arguing that the legal system is dynamic and capable of coping with new challenges by so-called new technologies).

this context.[284] But others believe that may not be the case,[285] because machine-learning applications are dynamic; they use data that is constantly updated and combined in new ways. AI systems are by their nature not intended to be static; they are systems that learn and adapt often in unpredictable ways. Therefore, they pose issues that are more complex than tools that are stable. Moreover, when the technology is "black box," it may be inherently impossible to determine how and why the system reached the conclusion it did or to reverse engineer the decision-making process. Classic tort law assigns liability based on fault. For example, a product is defectively designed when a reasonable alternative was possible and could have avoided foreseeable harm. These questions about alternative design or foreseeability simply may not be answerable when it comes to AI.

Consider, for example, the following hypothetical:

> Company is responsible for operating a dam and generating hydroelectric power. Company decides to modernize in order to be more efficient. It replaces its human-operated control system with a fully autonomous AI system. To enable the AI to function, Company installs a large number of sensors throughout the dam and the area in which the dam is. They collect temperature, moisture, stress, and other readings and send them via the internet to the AI. The "AI" actually consists of a number of components. The primary component is located in Company's primary corporate office some five hundred miles away. It constantly monitors the sensor data and varies water flow on a continuous basis. It implements its decisions via instructions to its implementation module in the dam control room on site. Meanwhile the AI modifies its programming based upon its ongoing experience of the interaction

[284] *See, e.g.*, *Artificial Intelligence Litigation: Can the Law Keep Pace with the Rise of the Machines?*, QUINN EMANUEL URQUHART & SULLIVAN, LLP (Dec. 2016), https://www.quinnemanuel.com/the-firm/publications/article-december-2016-artificial-intelligence-litigation-can-the-law-keep-pace-with-the-rise-of-the-machines [https://perma.cc/3LYN-CC3K]. The treatment of AI systems under criminal law poses unique issues because of the *mens rea* requirement for imposing criminal liability. For a discussion of these issues, see Francesca Lagioia & Giovanni Sartor, *AI Systems Under Criminal Law: A Legal Analysis and a Regulatory Perspective*, 33 PHIL. & TECH. 433 (2020). For an interesting take on how the law might address artificially intelligent robots that misbehave, see Mark A. Lemley & Bryan Casey, *Remedies for Robots*, 86 U. CHI. L. REV. 1311 (2019).

[285] *See, e.g.*, Matthew U. Scherer, *supra* note 263, at 388–92. There is extensive debate in both the literature and the popular press over whether AI should be regulated and what form that regulation (if any) should take. *See* sources cited *supra* note 263. *See also, e.g.*, *Regulation of Artificial Intelligence*, WIKIPEDIA, https://en.wikipedia.org/w/index.php?title=Regulation_of_artificial_intelligence&oldid=1000320251 [https://perma.cc/ZM5S-TCVN]; R. David Edelman, *Here's How to Regulate Artificial Intelligence Properly*, WASH. POST (Jan. 13, 2020), https://www.washingtonpost.com/outlook/2020/01/13/heres-how-regulate-artificial-intelligence-properly [https://perma.cc/5857-NNR9]; Paul Chadwick, *To Regulate AI We Need Laws, Not Just a Code of Ethics*, GUARDIAN (Oct. 28, 2018), https://www.theguardian.com/commentisfree/2018/oct/28/regulate-ai-new-laws-code-of-ethics-technology-power [https://perma.cc/J8G3-FCW2].

> of all the monitored sensor factors in order to produce the most electricity at the cheapest operating cost while maintaining community safety. The AI is also connected, via the internet, to other dam systems so that it can learn from how those systems are operating.
>
> One night, the AI fully opens the emergency floodgates and floods one thousand homes downstream. Company investigates and cannot determine causation. Possibilities include: defective AI design; defective AI training; defective sensor design and/or manufacture; unforeseen consequences from multiple data inputs in real world circumstances; erroneous AI operation based upon sensor or remote data; and external interference, that could have been accidental or intentional, by either one or more private actors or on behalf of a foreign organization or nation. Notably, the sensors are from multiple companies and may have never been used together prior, certainly not in the instant configuration.[286]

It is entirely possible that causation and fault may be indeterminate under these circumstances.[287] Given multiple potential tortfeasors, courts are usually able to apportion damages in a reasonable manner, but that assumes that both the tortfeasors and causation can be identified.[288] While an analysis of proper legal and regulatory regimes for AI and/or whether we need an entirely new regime such as legal personhood for AI[289] is beyond the scope of this paper, it is clear that questions about accountability for AI failures will remain in the hands of the courts for the foreseeable future.

---

[286] Giuffrida, *supra* note 283, at 446–47 (attributed to Professor Frederic I. Lederer, Chancellor Professor of Law, William & Mary Law School, *see Fredric I. Lederer*, WM. & MARY L. SCH., https://law2.wm.edu/faculty/bios/fulltime/filede.php [https://perma.cc/C8UT-DGYN]).

[287] *See id.* While, in some respects, the hypothetical presented might sound like a classic engineering malpractice claim, or something that would be prohibited by a regulator, it raises the issue of how to harness advancements in technology without, at the same time, hamstringing innovation through the litigation process. There is always a risk-reward tradeoff with advances in technology; early adopters assume greater risk than late adopters. A further discussion of this issue, however, is beyond the scope of this paper.

[288] *Id.*

[289] *See* MIREILLE HILDEBRANT, *Legal Personhood for AI?*, *in* LAW FOR COMPUTER SCIENTISTS 237 (2019). *See also* John-Stewart Gordon, *Artificial Moral and Legal Personhood*, 36 AI & SOC. (2020); Tyler L. Jaynes, *Legal Personhood for Artificial Intelligence: Citizenship as the Exception to the Rule*, 35 AI & SOC. 343 (2020). For a discussion of the perspective of the European Parliament on this issue, see Markus Häuser, *Do Robots Have Rights? The European Parliament Addresses Artificial Intelligence and Robotics*, CMS LAW-NOW (June 4, 2017), https://www.cms-lawnow.com/ealerts/2017/04/do-robots-have-rights-the-european-parliament-addresses-artificial-intelligence-and-robotics [https://perma.cc/HH6E-X4E3].

G. Lack of Resilience

Resilience refers to the degree to which AI systems can detect and resist both intentional and unintentional efforts to cause machine-learning models to fail, or to otherwise adapt to risk.[290] While researchers have developed measures to protect AI systems from such failures,[291] sophisticated hackers quickly learn ways to circumvent these defensive measures, and so it goes in a vicious cycle.

One of the biggest challenges that has emerged along with the introduction of digital evidence is the ease with which it can be altered through means such as spoofing.[292] Recently, a family law attorney in California reported on fake evidence used in several of his divorce cases.[293] In one particular matter, where a husband had been granted temporary custody of the children, the wife submitted text messages as evidence of domestic abuse perpetrated by the husband.[294] The wife was granted a Domestic Violence Restraining Order ("DVRO") and the three children were removed from the custody of husband, with no visitation permitted, pending resolution of the charges related to the threats contained in the text messages he had allegedly sent her.[295] The only problem was that the husband had not sent the text messages.[296] All the wife did was change the name associated with someone else's phone number in her cell phone to her husband's name

---

[290] *See* Nathan Michael, *Shield AI Fundamentals: On Resilient Intelligence*, SHIELD AI (June 25, 2019), https://www.shield.ai/content/2019/6/25/shield-ai-fundamentals-on-resilient-intelligence [https://perma.cc/Y5CD-KWEM].

[291] *See, e.g.*, Shilin Qui et al., *Review of Artificial Intelligence Adversarial Attack and Defense Technologies*, 9 APPLIED SCI. 909 (2019); Ali Chehab et al., *Machine Learning for Network Resilience: The Start of a Journey*, PROC. 2018 5TH INT'L CONF. ON SOFTWARE DEFINED SYS. ("SDS"). 59 (2018); Yevgeniy Vorobeychik, *Adversarial AI*, PROC. 25TH INT'L JOINT CONF. ON ARTIFICIAL INTELL. ("IJCAI-16") 4094 (2016).

[292] "Spoofing is the act of disguising a communication from an unknown source as being from a known, trusted source. Spoofing can apply to emails, phone calls, and websites, or can be more technical, such as a computer spoofing an IP address. . . . Spoofing can be used to gain access to a target's personal information, spread malware through infected links or attachments, bypass network access controls, or redistribute traffic to conduct a denial-of-service attack. Spoofing is often the way a bad actor gains access in order to execute a large cyber attack. . . ." *What Is Spoofing? Spoofing Defined, Explained, and Explored*, FORCEPOINT, https://www.forcepoint.com/cyber-edu/spoofing [https://perma.cc/7KJP-EC6X].

[293] M. Jude Egan, *Deep Fakes in Divorce Court: Manipulated Electronic Evidence and What to Do About It*, LEGALTECH NEWS (Aug. 20, 2020), https://www.law.com/therecorder/2020/08/20/deep-fakes-in-divorce-court-manipulated-electronic-evidence-and-what-to-do-about-it [https://perma.cc/TWC6-8V9D]. Fake evidence is becoming a major challenge for judges and lawyers. *See* Matt Reynolds, *Courts and Lawyers Struggle with Growing Prevalence of Deepfakes*, ABA J., https://www.abajournal.com/web/article/courts-and-lawyers-struggle-with-growing-prevalence-of-deepfakes [https://perma.cc/N9M6-GXE3].

[294] *See* Egan, *supra* note 293.

[295] *Id.*

[296] *Id.*

and then sent herself the threatening texts.[297] When she printed out the texts to attach to the application, the husband's name appeared at the top of the messages and made it appear as if he had sent the messages.[298] Since judges in California often read DVRO requests on written pleadings, without notice to the other party, the restrained party may not have an opportunity to challenge the Temporary Restraining Order ("TRO") until a hearing is held.[299] In this case, the hearing was continued for almost four months due to the intervening Christmas Holiday and other events.[300] At the hearing, the husband was able to offer his monthly phone statement showing that he had never sent his wife a single text message on the date at issue, or any other day that month.[301] The judge dismissed the DVRO, but did not award the husband full custody of the children.[302] Fake evidence can be so sophisticated and convincing that it can take a forensic examiner to determine whether the evidence is real or not, but such expert assistance can be quite expensive in the average case. Photographs, audiotapes, and video images are also easily manipulated. While humans have a strong tendency to believe their own eyes and ears, and digital evidence has traditionally been given considerable credence, things are not always what they seem to be. This problem can cause judges to be reticent to grant domestic violence TROs when they are needed, and to be suspicious of other evidence that is actually authentic. This problem will only be exacerbated by AI.

"Adversarial AI" refers to the use of the very power of AI to pose malicious threats.[303] Such techniques attempt to fool machine-learning models by supplying deceptive input(s), most often to cause some kind of malfunction. Adversarial AI can be used to attack just about any kind of system built on AI technology, from causing an automated email message to disclose sensitive data such as credit card numbers, to tricking a computer

---

297 *See id.*

298 *Id.*

299 *Id.*

300 *Id.*

301 *Id.*

302 *Id.*

303 *See Adversarial Machine Learning*, WIKIPEDIA, https://en.wikipedia.org/w/index.php?title=Adversarial_machine_learning&oldid=1001999631 [https://perma.cc/6HKC-5SET]; Ben Dickson, *What Is Adversarial Machine Learning?*, TECHTALKS (July 15, 2020), https://bdtechtalks.com/2020/07/15/machine-learning-adversarial-examples [https://perma.cc/D9HF-88M8]. For a more technical discussion of adversarial machine learning, see, for example, Kevin Eykholt et al., *Robust Physical-World Attacks on Deep Learning Visual Classification*, PROC. 2018 IEEE/CVR CONF. ON COMPUTER VISION & PATTERN RECOGNITION 1625 (2018). For a taxonomy of different types of attacks on machine-learning technologies, and a variety of defenses against those attacks, see Marco Barrero et al., *Can Machine Learning Be Secure?*, Proc. 2006 ACM SYMP. ON INFO., COMPUTER, AND COMM. SECURITY ("ASIACCS '06") 16 (2006).

vision system in an automated vehicle by placing stickers on road signs, causing the vehicle to mistake a stop sign for a merge or speed limit sign.[304] By way of example, McAfee attacked Tesla's former Mobileye system, fooling it into driving 50 mph over the speed limit, by adding a two-inch strip of black tape to a speed limit sign.[305] Adversarial patterns on glasses or clothing can deceive facial recognition systems.[306] The possibilities are endless.

The mechanisms through which such attacks operate can differ, as can their specificity; a targeted attack attempts to allow a specific intrusion or disruption (*e.g.*, an attempt to gain access to personal information), whereas an indiscriminate attack creates general mayhem.

Evasion attacks are the most prevalent form of adversarial attack.[307] Spammers and hackers attempt to evade detection by obfuscating the content of spam or malware.[308] Samples of data are modified to evade detection so as to be classified as legitimate.[309] Another example of an evasion might be a spoofing attack against a biometric verification system, in which fake biometric traits may be exploited to impersonate a legitimate user.[310] Poisoning, on the other hand, is the adversarial contamination of training

---

[304] *See* Madeleine Clare Elish, *When Humans Attack: Re-thinking Safety, Security, and AI*, POINTS: DATA & SOC. (May 14, 2019), https://points.datasociety.net/when-humans-attack-re-thinking-safety-security-and-ai-b7a15506a115 [https://perma.cc/V7SQ-X834]; MILES BRUNDAGE ET AL., THE MALICIOUS USE OF ARTIFICIAL INTELLIGENCE: FORECASTING, PREVENTION, AND MITIGATION (2018).

[305] Brian Barrett, *Security News This Week: A Tiny Piece of Tape Tricked Teslas Into Speeding Up 50 MPH*, WIRED, https://www.wired.com/story/tesla-speed-up-adversarial-example-mgm-breach-ransomware [https://perma.cc/5EUV-6RZL].

[306] *See, e.g.*, Aaron Holmes, *These Clothes Use Outlandish Designs to Trick Facial Recognition Software into Thinking You're Not Human*, BUS. INSIDER, https://www.businessinsider.com/clothes-accessories-that-outsmart-facial-recognition-tech-2019-10 [https://perma.cc/2RU7-JA5H]; John Seabrook, *Dressing for the Surveillance Age*, NEW YORKER (Mar. 9, 2020), https://www.newyorker.com/magazine/2020/03/16/dressing-for-the-surveillance-age [https://perma.cc/HVM8-7DW5]; Simen Thys et al., *Fooling Automated Surveillance Cameras: Adversarial Patches to Attack Personal Detection*, arXiv:1904.08653v1 [cs.CV] (Apr. 18, 2019), https://arxiv.org/pdf/1904.08653.pdf.

[307] WIKIPEDIA, *supra note* 303.

[308] *See id.*

[309] *See* Ilja Moisejevs, *Evasion Attacks on Machine Learning (or "Adversarial Examples")*, TOWARDS DATA SCI. (July 14, 2019), https://towardsdatascience.com/evasion-attacks-on-machine-learning-or-adversarial-examples-12f2283e06a1 [https://perma.cc/XR6Z-2CFQ]. For a more technical discussion of evasion, see, for example, Blaine Nelson et al., *Query Strategies for Evading Convex-Inducing Classifiers*, 13 J. MACH. LEARN. 1293 (2012).

[310] *See* Danny Thakkar, *Spoofing Fingerprint Scanner and Spoof Detection: How Do They Work?*, BAYOMETRIC, https://www.bayometric.com/spoofing-fingerprint-scanner-and-spoof-detection [https://perma.cc/CPH3-CEWF]; For a more technical discussion of spoofing, see, for example, Ricardo N. Rodrigues et al., *Robustness of Multimodal Biometric Fusion Methods against Spoof Attacks*, 20 J. Visual Lang. and Computing 169 (2009).

data.[311] An attacker may poison such data by injecting malicious samples that disrupt subsequent retraining.

Other than spam, perhaps the most well-known adversarial attacks are deepfakes, synthetic media in which a person in an existing image or video is replaced with someone else's likeness.[312] While faking content is not new, deepfakes leverage powerful machine-learning techniques to manipulate or generate visual and audio content with a high potential to deceive. AI-powered deepfakes are already being used in everyday attacks such as fraud. In one widely publicized U.K. case, a victim received a phone call from what he thought was his boss instructing him to wire money to the bank account of a supplier in Hungary.[313] The call and email that followed accurately replicated the mannerisms, accent, and diction of his employer.[314] The "Synthesizing Obama" program in 2017 modified video footage of former President Barack Obama to depict him mouthing the words contained in a separate audio track.[315] While this was an academic exercise, other such

---

[311] *See* Ilja Moisejevs, *Poisoning Attacks in Machine Learning*, TOWARDS DATA SCI., (July 14, 2019), https://towardsdatascience.com/poisoning-attacks-on-machine-learning-1ff247c254db [https://perma.cc/839Z-MESS]; For a more technical discussion of poisoning, see, for example, Gan Sun et al., *Data Poisoning Attacks on Federated Machine Learning*, Vol. 14, No. 8, J. of Latex Class Files, 1 (2015).

[312] *Deepfake*, WIKIPEDIA, https://en.wikipedia.org/w/index.php?title=Deepfake&oldid=1002384861 [https://perma.cc/B6LS-MTCS]; *see also* Ian Sample, *What Are Deepfakes and How Can You Spot Them?*, GUARDIAN, https://www.theguardian.com/technology/2020/jan/13/what-are-deepfakes-and-how-can-you-spot-them [https://perma.cc/L6LU-DLHV]; Aseem Kishore, *What Is a Deepfake and How Are They Made?*, ONLINE TECH TIPS, (May 23, 2019), https://www.online-tech-tips.com/computer-tips/what-is-a-deepfake-and-how-are-they-made [https://perma.cc/E2P2-642D].

[313] *See* Catherine Stupp, *Fraudsters Used AI to Mimic CEO's Voice in Unusual Cybercrime Case*, WSJ, (Aug. 30, 2019, 12:52 PM), https://www.wsj.com/articles/fraudsters-use-ai-to-mimic-ceos-voice-in-unusual-cybercrime-case-11567157402 [https://perma.cc/9RF5-G72R].

[314] Rahul Kashyap, *Are You Ready for the Age of Adversarial AI? Attackers Can Leverage Artificial Intelligence Too*, FORBES, https://www.forbes.com/sites/forbestechcouncil/2020/01/09/are-you-ready-for-the-age-of-adversarial-ai-attackers-can-leverage-artificial-intelligence-too/?sh=22337f3a4703 [https://perma.cc/ZM8P-YLR7].

[315] Daniel Akst, *The Researchers Who Synthesized Video of Barack Obama*, WSJ, https://www.wsj.com/articles/the-researchers-who-synthesized-video-of-barack-obama-1500655962 [https://perma.cc/L4MD-AZ6G]; For a copy of the research discussed in the WSJ article, see Supasorn Suwajanakorn et al., *Synthesizing Obama: Learning Lip Synch from Video*, 36 ACM TRANSAC. ON GRAPHICS ("SIGGRAPH 2017") (2017); For a video describing how the synthesized video was prepared, see Supasorn Suwajanakorn et al., *Synthesizing Obama: Learning Lip Sync from Audio;* SIGGRAPH (2017), GRAIL, https://grail.cs.washington.edu/projects/AudioToObama [https://perma.cc/HXZ3-T2J9]; For a more recent example of a deepfake video of Queen Elizabeth giving her annual Christmas speech, see Bruce Haring, *Queen Elizabeth 'Deepfake' Message Jabs Prince Harry and Meghan, Prince Andrew*, DEADLINE, https://deadline.com/2020/12/queen-elizabeth-deepfake-message-jabs-harry-meghan-prince-andrew-1234661642 [https://perma.cc/MAG6-EG26]; Of course, the two Canadian authors of this paper take issue with the Queen's jab at Canadians ("There are few things more hurtful than someone telling you they prefer the company of Canadians."). *Id.*

adversarial efforts are not. In January 2018, a proprietary desktop application called FakeApp was launched by an anonymous Reddit user. [316] It allowed users to easily create and share videos with their faces swapped with their friends.[317] Since then, FakeApp has been superseded by open-source alternatives such as faceswap.[318] Other deepfake efforts that are less amusing involve the alteration or manipulation of video related to well-known public officials. These are now being used to sow distrust in public and government institutions. Current events underscore the danger of these types of AI.

It is becoming increasingly difficult to distinguish material generated by AI from that generated by humans. In August 2020, a college student was able to use Generative Pre-trained Transformer 3 ("GPT-3"), one of the most powerful language-generating AI models to date, to create, in a matter of hours, a fake blog on productivity and self-help.[319] Many people hit

---

[316] *See* Dan Marino, *FakeApp: Groundbreaking or Dangerous?*, ARTEFACT, (Feb. 13, 2018), https://www.artefactmagazine.com/2018/02/13/fakeapp-groundbreaking-or-dangerous [https://perma.cc/K7R4-96HX]. To download FakeApp version 2.2.0, see *FakeApp*, MALAVIDA, https://www.malavida.com/en/soft/fakeapp/#gref [https://perma.cc/TSG7-XKDU]. For a discussion of another application (Zao) that allows users to add themselves into their favorite movies, see Ryan Gilbey, *A 'Deepfake' App Will Make Us Film Stars – But Will We Regret Our Narcissism?*, GUARDIAN, https://www.theguardian.com/technology/2019/sep/04/a-deep-fake-app-will-make-us-film-stars-but-will-we-regret-our-narcissism [https://perma.cc/95QU-4Z8H].

[317] Marino, *supra* note 316.

[318] Faceswap bills itself as "the leading free and Open Source multi-platform Deepfake Software." *Welcome*, FACESWAP, https://faceswap.dev [https://perma.cc/YK2J-AQWW]. Faceswap can be downloaded here: *Download*, FACESWAP, https://faceswap.dev/download [https://perma.cc/YTF4-F2JT].

[319] *See* Karen Hao, *A College Kid's Fake AI-generated Blog Fooled Tens of Thousands. This Is How He Made It.*, MIT TECH. REV. (Aug. 14, 2020), https://www.technologyreview.com/2020/08/14/1006780/ai-gpt-3-fake-blog-reached-top-of-hacker-news [https://perma.cc/8R5E-D7A7]. In July of 2020, one month earlier, another college student using GPT-3 had launched an AI startup that could write emails "automatically from a few fragmentary notes . . . by learning the way 'a user thinks and responds'. . . ." Ken Schachter, *Long Island Artificial Intelligence Startup Can Write Your Emails*, NEWSDAY, (Sept. 21, 2020, 6:08 PM), https://www.newsday.com/business/technology/othersideai-artificial-intelligence-software-melville-matt-shumer-1.49644292 [https://perma.cc/45RH-S3CD]. In November 2020, the New York Times asked GPT-3 to write one of its *Modern Love* columns by giving it a short prompt: "The following is a New York Times's Modern Love Column about a woman named Frances, her husband Dean, and the unlikely story of how they first met." In response, GPT-3 wrote the following:

> After my fiancé died, my mother told me to "get out there again." She wanted me to go to a singles bar. I told her I'd rather go to the dentist.
>
> "Just once," she said. "Just to see what it's like."
>
> One day, early last year, I found myself driving to a singles bar in winter snow. I sat in my car for 15 minutes, then drove away. The next day, I went back and sat in my car for another 15 minutes. I did this for a couple of weeks, until I finally mustered up the nerve to walk in.

"subscribe." We can expect the appearance of more and more (truly) fake news.

Even more horrifying, perhaps, is the recent retraction by one of the world's most highly cited medical journals—*The Lancet*—of an article on Hydroxychloroquine ("HCQ") that caused the World Health Organization and several Latin American nations to halt their trials of the medication for the treatment of COVID.[320] The authors of the study withdrew the article

---

> The place was mobbed. It was like one of those bad dreams where you show up for a test without having studied. I tried to size up the situation. I was a short, thin middle-aged woman in a sea of young, good-looking men. I didn't see anyone I wanted to meet. I decided to leave.
>
> As I walked toward the door, a man walking toward me accidentally bumped into me. "I'm sorry," he said. We shook hands, and I introduced myself. He asked where I was from, and I told him. . . . He was from a small town nearby. We talked for a while, and I asked him to have a drink with me.
>
> "I don't drink," he said.
>
> I was so shocked, I dropped my purse and my keys and had to pick them up. I don't drink, either, but I would have had a drink with this man, this stranger. "That's OK," I said.
>
> "I don't know," he said, and he seemed a little uncomfortable.
>
> "It's not a test," I said. "I don't care if you drink."
>
> "Well, in that case, I'll have a glass of water."

Cade Metz, *When A.I. Falls in Love*, N.Y. TIMES, (Nov. 24, 2020), https://www.nytimes.com/2020/11/24/science/artificial-intelligence-gpt3-writing-love.html [https://perma.cc/SH3W-SRRA]. For a discussion of what GPT-3 is, what it can do, and how it works, see Bernard Marr, *What Is GPT-3 And Why Is It Revolutionizing Artificial Intelligence?*, FORBES, (Oct. 5, 2020, 12:21 AM), https://www.forbes.com/sites/bernardmarr/2020/10/05/what-is-gpt-3-and-why-is-it-revolutionizing-artificial-intelligence/?sh=6a12d762481a [https://perma.cc/JJ4W-ZZ2W]. For a more technical discussion of GPT-3, see Tom B. Brown et al., *Language Models Are Few-Shot Learners*, arXiv2005.14165v4 [cs.CL] (July 22, 2020), https://arxiv.org/pdf/2005.14165.pdf [https://perma.cc/U592-GU8W]. For a less optimistic view of GPT-3, see Rob Toews, *GPT-3 Is Amazing—And Overhyped*, FORBES, (July 19, 2020, 6:56 PM), https://www.forbes.com/sites/robtoews/2020/07/19/gpt-3-is-amazingand-overhyped/?sh=4a59d1fb1b1c [https://perma.cc/UEE2-JZKL]; Tom Taulli, *Turing Test At 70: Still Relevant For AI (Artificial Intelligence)?*, FORBES, (Nov. 27, 2020, 12:59 PM), https://www.forbes.com/sites/tomtaulli/2020/11/27/turing-test-at-70-still-relevant-for-ai-artificial-intelligence/?sh=660c340e250f [https://perma.cc/C6UJ-KQBD] (noting that if you ask a GPT-3 system how many eyes the sun has, it responds that there is one, and if you ask it who was the president of the U.S. in 1600, it responds "Queen Elizabeth I").

[320] *See* Sarah Boseley & Melissa Davey, *Covid-19: Lancet Retracts Paper that Halted Hydroxychloroquine Trials*, GUARDIAN (June 4, 2020, 3:43 PM), https://www.theguardian.com/world/2020/jun/04/covid-19-lancet-retracts-paper-that-halted-hydroxychloroquine-trials [https://perma.cc/TXH5-LWD8].

because they determined that they could no longer vouch for the data obtained from a healthcare analytics company named Surgisphere.[321] After the paper was published, concerns were raised about the veracity of the data and the analysis of same conducted by the corporation. Surgisphere claimed to have collected data from 15,000 coronavirus patients who received HCQ alone, or in combination with antibiotics, from 1,200 hospitals around the world.[322] Subsequent investigations by The Guardian Australia, among others, revealed that the data was fake, when reporters contacted five Australian hospitals reported to have provided data and they denied it.[323] Moreover, the number of deaths reported in Australia due to coronavirus also did not match the numbers from the purported Australian database.[324]

There have been a number of federal and state efforts to enact regulatory responses to the problems posed by deepfakes,[325] but most have not yet been successful. On October 3, 2019, however, California Governor Newsom signed into law Assembly Bill Nos. 602 and 730, which respectively, provide individuals targeted by sexually explicit deepfake content made without their consent a cause of action against the content's creator, and prohibit the distribution of malicious deepfake audio or visual media targeting a candidate running for public office within 60 days of their election[326] Until better technology and more legislation emerge, the challenge of detecting deepfakes and addressing the mischief they may cause will fall in the hands of the U.S. courts. The remainder of this article will address the ways that lawyers and judges can test the veracity of the data used to fuel AI

---

[321] Mandeep R. Mehra et al., *Retraction—Hydroxychloroquine or Chloroquine with or Without a Macrolide for Treatment of COVID-19: A Multinational Registry Analysis*, LANCET, (June 5, 2020), https://www.thelancet.com/journals/lancet/article/PIIS0140-6736(20)31324-6/fulltext [https://perma.cc/9P8A-25XC].

[322] *See* Melissa Davey, *Questions Raised over Hydroxychloroquine Study Which Caused WHO to Halt Trials for Covid-19*, THE GUARDIAN, (July 1, 2020, 12:21 PM), https://www.theguardian.com/science/2020/may/28/questions-raised-over-hydroxychloroquine-study-which-caused-who-to-halt-trials-for-covid-19 [https://perma.cc/7QGU-889A]; *Medical Journal The Lancet Retracts its HCQ Article Based on Fake Data from a Dubious Company, Authors Say they Cannot Vouch for Data's Authenticity*, OPINDIA, (June 5, 2020), https://www.opindia.com/2020/06/lancet-retracts-article-study-hydroxychloroquine-trials-fake-data-surgisphere-who-clinical-trials-chicago-company [https://perma.cc/338G-KRRC].

[323] *See* OPINDIA, *supra* note 322.

[324] *See id.*

[325] *See* Matthew F. Ferraro, *Deepfake Legislation: A Nationwide Survey—State and Federal Lawmakers Consider Legislation to Regulate Manipulated Media*, WILMERHALE CLIENT ALERT (Sept. 25, 2019), https://www.wilmerhale.com/en/insights/client-alerts/20190925-deepfake-legislation-a-nationwide-survey [https://perma.cc/5DCY-M6P9].

[326] *See* K.C. Halm et al., *Two New California Laws Tackle Deepfake Videos in Politics and Porn*, DAVIS WRIGHT TREMAINE LLP: ARTIFICIAL INTELLIGENCE LAW ADVISOR (Oct. 11, 2019), https://www.dwt.com/blogs/artificial-intelligence-law-advisor/2019/10/california-deepfakes-law [https://perma.cc/U95R-G66Q].

tools, the bona fides of the tools themselves, and the output of such tools when they are presented in court as evidence.

VI. ESTABLISHING VALIDITY AND RELIABILITY

A. Testimony, Expert Testimony, or Technology?

Because AI employs technology to emulate or exceed human cognitive ability, the question arises as to whether evidence gleaned from AI should be judged by the standard of direct witness testimony, expert witness testimony, or measurement using established technology.

Consider, for example, a smart digital assistant that "listens" to everything that goes on in a home, an automobile, or within "earshot" of a mobile phone. Arguably, the digital assistant is a direct witness to what it hears. At the same time, the digital assistant may employ sophisticated technology like voice recognition to draw conclusions regarding the identity of the speaker, their tone of voice, and the words that are spoken. It may also act as a verbatim recording device, capturing sound, time, global position, speed, and motion, and perhaps video. Some or all of this information may be stored in the device or transmitted to the cloud where it may be retrieved even if the device is lost or destroyed.[327]

When author Cormack's credit card was declined in Australia, he was sent the following voicemail transcript:

(800) 466-7295 4 Jul 2014, 9:15 am

> Yeah. This is an urgent call for Gordon. Cormac, yum the T. V. Canada Trust Loss Prevention center. This is not a telemarketing call. We would like to verify some recent activity on your T E D U. S. Dollar visa card, ending in. 8 Yeah, 0 Your yeah 1. Whether protection and security of your T V credit card account is very important that we speak to you. Please call us toll free at 1(800) 466-7295. You may call us back 24 hours a day, seven days a week. Yeah, the number again is 1(800) 466-7295. Thank you for choosing P D, Canada Trust goodbye.

This message was incorrectly marked spam and never delivered to Cormack's email and was discovered only when Cormack telephoned a bank representative, who told him that a voice message had been left for him. The effort to find this message resulted in the serendipitous discovery of two other important messages that had also been blocked by the spam filter:

---

[327] *See, e.g.*, Anthony Cuthbertson, *Amazon Ordered to Give Alexa Evidence in Double Murder Case*, INDEPENDENT (Nov. 14, 2018), https://www.independent.co.uk/life-style/gadgets-and-tech/news/amazon-echo-alexa-evidence-murder-case-a8633551.html [https://perma.cc/U9TR-M4RA].

+1 XXX-XXX-XXXX 10 Jun 2014, 11:11 am

> Yeah, Hi. My name is calling. I'm calling with the Canada Revenue Agency, This message is for Gordon have a question regarding some self-employed earnings from U 2013 tax returns. Please call me back. Toll free number is 1(XXX) XXX-XXXX (XXX) XXX-XXXX. Thank you. Bye. (Phone numbers redacted).

+1 XXX-XXX-XXXX 18 Jun 2014, 10:00 am

> Hi, My name is Clint calling with the Canada Revenue Agency doing a follow up on the message I left on June 10th. Certain court and to the questions and some self employed or drinks from the 2013 tax. Please give a call back. Toll free number is 1(XXX) XXX-XXXX (XXX) XXX XXX. Thank you. (Phone numbers redacted.)

While these communications played no role in any legal controversy, it is easy to imagine a situation in which similar communications could have. Are the transcripts genuine? Are they accurate? Were they in fact blocked by a spam filter? Did the bank, the revenue agent, the spam filter, and the intended recipient exercise reasonable diligence to ensure that the communications were successful? Should the recipient, having read the transcript, be deemed to have been notified of its content? Should he have assumed that they were real rather than a scam or phishing attack?[328]

Establishing the provenance of the transcript involves several factors: (i) whether a call was really placed from the specified phone number to the recipient at the specified time; (ii) what voice recognition system was used to produce the transcript; (iii) what version and configuration was used, and how was it trained; and (iv) whether the proffered text is an accurate reproduction of the transcript?

Accuracy does not mean perfection. Clearly there are errors in each of the examples. The name of the bank is T.D. [Canada Trust] not T.V. or T.E.D. or P.D. The revenue agent's name was neither "calling" nor "Clint." "Self employed or drinks" presumably should be "self-employed earnings."

[328] A phishing attack is a "fraudulent attempt to obtain sensitive information or data, such as usernames, passwords and credit card details or other sensitive details, by impersonating oneself as a trustworthy entity in a digital communication. Typically carried out by email spoofing, instant messaging, and text messaging, phishing often directs users to enter personal information at a fake website which matches the look and feel of the legitimate site. Phishing is an example of social engineering techniques used to deceive users. Users are lured by communications purporting to be from trusted parties such as social networking websites, auction sites, banks, mails/messages from friends or colleagues/executives, online payment systems or IT administrators." *Phishing*, WIKIPEDIA, https://en.wikipedia.org/w/index.php?title=Phishing&oldid=1002208250 [https://perma.cc/6X4X-386E].

There are several spelling mistakes. Notwithstanding these errors, it might be argued that the transcripts convey accurately enough the substance of the voicemail messages, and also the spoken telephone numbers, which were correctly transcribed.

Determining whether a transcript is *accurate enough* is fraught with challenges: Precisely defining and quantifying what is meant by "accuracy," estimating the accuracy of a particular transcript, determining what threshold of accuracy is sufficient, and determining the reliability with which a transcription tool meets this threshold.

As a term of art, the accuracy of a transcript typically refers to the fraction or percentage of words that are correctly transcribed. To evaluate accuracy, according to this definition, it is necessary to define, in turn, what is meant by a word, and what is meant for that word to be correctly translated. Is "T.D." one word or two, and is its correct spelling "T.D." or "TD"? How is the spurious E in "T E D" to be counted? Is the telephone number 1(800) 466-7295 a word? It was probably spoken as ten words: "one eight hundred four six six seven two nine five." Are homonyms or sound-alike words correct or incorrect?

Any quantitative assessment of accuracy depends on such arbitrary but necessary choices. For a reasonable set of choices, we might determine that the first voicemail message contained 120 words, of which 100 were correctly transcribed, or 83% accuracy. Error—the complement of accuracy—is 17%, or one in six. It can be argued that this transcript could be considered accurate enough for many purposes.

But this is not to say that the transcription tool always achieves 83% accuracy, or that all transcripts achieving 83% accuracy are sufficiently accurate to assume the recipient has knowledge. In the first transcript, TD was consistently misspelled, but arguably, the words "Canada Trust" provided essential context. Imagine if the caller had referred to the bank as simply TD—would the recipient be able to determine that the call was not just another phishing attempt? Would the accuracy be considered acceptable?

The error rate in this transcript was 17%, or one-in-six words. Imagine a different transcription in which one in six of the digits of the telephone number were transcribed incorrectly. Would such accuracy be considered acceptable?

Admittedly, these are contrived examples, and generally, we find that measured accuracy and acceptable accuracy are well correlated. Researchers and developers take advantage of this correlation to evaluate and improve their AI systems, under the assumption that improving measured accuracy tends to improve the reliability with which an AI system achieves its

intended purpose: here, a transcript sufficient to convey the substance of the message.

B. Benchmarks and Goodhart's Law

In 1975, Charles Goodhart, acting as a member of the Bank of England's Policy Committee, observed that "any observed statistical regularity will tend to collapse once pressure is placed upon it for control purposes."[329] In other words, when a statistical measure of effectiveness, accuracy, or reliability is used as a target or acceptance criterion, it ceases to be a valid measure. The reason for this effect is that, although the measure may be apt if the measurement is conducted independent of what is being measured, it is no longer independent and therefore, no longer apt, if the thing being measured is influenced by the measurement.[330] In more common terms, the purpose of a college examination is defeated if the examinees are aware of the questions beforehand.

Benchmarks and statistical measures are very useful tools for monitoring and improving the effectiveness of AI technologies. But if these benchmarks are public or used repeatedly, technologies will evolve—whether intentionally or not—to optimize their performance with respect to the benchmark and the chosen measure of success, not the general problem for which the benchmark is intended to be a representative example, or the underlying property that the measure was designed to estimate.

---

[329] David Manheim & Scott Garrabrant, *Categorizing Variants of Goodhart's Law*, https://arxiv.org/pdf/1803.04585.pdf [https://perma.cc/2LCS-996D]. *See also Goodhart's law*, WIKIPEDIA, https://en.wikipedia.org/w/index.php?title=Goodhart%27s_law&oldid=999730673 [https://perma.cc/J3NS-Y89X].

[330] A famous example of Goodhart's law is the "Cobra effect." *See* Cedric Chin, *The Four Flavors of Goodhart's Law*, HOLISTICS BLOG, https://www.holistics.io/blog/four-types-goodharts-law [https://perma.cc/ZF5X-AK73]. So, the story goes, the British Colonial Government in India was becoming concerned about the increasing number of venomous cobras in Delhi, so it began offering a bounty for each dead cobra that was delivered. *Id.* Initially, this was a successful strategy; locals brought in large numbers of the slaughtered snakes. *Id.* But over time, enterprising individuals started to breed cobras in order to kill them for the supplemental income. *Id.* When the government abandoned the bounty, the cobra breeders released their cobras into the wild and Delhi experienced a surge in its snake population. *Id.* Similarly, in 1902, the French Colonial government in Hanoi created a bounty program to reduce the rat population. *Cobra effect*, WIKIPEDIA, https://en.wikipedia.org/w/index.php?title=Cobra_effect&oldid=1002053645 [https://perma.cc/NP3X-MPN7]. To collect the bounty, locals needed to provide the severed tail of a rat. *Id.* Shortly thereafter, Vietnamese officials began to notice an increasing number of rats running around the city without tails. *Id.* It turned out that the rat catchers would capture the rats, sever their tails, and release them back into the sewers so they would procreate, produce more rats, and therefore generate more revenue. *Id.* So, too, when a court indicates that claims and defenses must be based on "evidence," this can lead to pressures and incentives to massage and manipulate such "evidence," either by optimizing for a metric that defeats the metric's goal or that reduces its predictive effect. *See* Chin, *supra* note 330. This has also been referred to as "Adversarial Goodhart." *Id.*

This issue was brought to the fore recently with respect to vehicle emissions testing. Given a standard evaluation protocol and a measure of success, the systems learn (or are taught) to behave differently when they are being tested, and to optimize not actual emissions, but whatever the test instruments register.[331]

In an ideal world, the accuracy and reliability of AI tools should be established by *independent* testing. Even so, it is necessary to consider carefully whether the results from such testing actually transfer to the problem at hand. In practice, progress in AI has occurred so quickly that often such independent testing has not yet occurred. Some AI tools have been rigorously tested by their developers; others, not so much. Some vendors disclose the nature of the testing they have conducted, but rarely do they disclose detailed protocols and results. Should they be required to do so, as are the purveyors of drugs, medical devices, and safety-critical equipment? Until such time as requirements like these are implemented, unvetted AI technologies will continue to be deployed, and it will be necessary to estimate their effectiveness and reliability on an ad-hoc basis. It would be unwise to consider such ad-hoc determinations as judicial notice, absent rigorous independent testing.

As an example, consider the voice transcription results shown above. There is reason to believe that the major corporation providing the transcription service has tested its software and has a reputational (if not economic) incentive for it to work well. And, perhaps, it works well enough for its intended purpose in this particular example. That transcript might even be offered in evidence to demonstrate that Cormack had notice, provided its provenance could be established. But the authors would not suggest that all transcription software, or indeed all transcriptions provided by this particular company, are necessarily accurate or should automatically be admitted as evidence.

As particular AI tools mature, the standards for their acceptance as evidence should tighten, as should the criteria to be used in assessing the weight of the evidence provided by them.

---

[331] *See, e.g.*, Benjamin Hulac, *Volkswagen Uses Software to Fool EPA Pollution Tests*, SCI. AM. (Sept. 21, 2015), https://www.scientificamerican.com/article/volkswagen-uses-software-to-fool-epa-pollution-tests [https://perma.cc/HN5J-463L]; *Volkswagen emissions scandal*, WIKIPEDIA https://en.wikipedia.org/w/index.php?title=Volkswagen_emissions_scandal&oldid=1000735588 [https://perma.cc/BD5K-62TY].

VII. EVIDENTIARY PRINCIPLES THAT SHOULD BE CONSIDERED IN EVALUATING THE ADMISSIBILITY OF AI EVIDENCE IN CIVIL AND CRIMINAL TRIALS

A. Adequacy of the Federal Rules of Evidence in Addressing the Admissibility of AI Evidence

As the above discussion illustrates, understanding what AI is, and how it functions in the many different applications in which it is used, is a complex and challenging undertaking. This complexity is no less present when lawyers and judges are faced with the task of determining how to evaluate the admissibility of AI evidence when it is offered to support and defend claims in civil and criminal cases. To date, there have been few, if any, court decisions squarely addressing this topic, and the cases that have referenced AI evidence often have done so in a cursory or tangential manner.[332] Th challenge is compounded by the fact that the Federal Rules of Evidence[333] are amended infrequently, and the process of amendment is slow, because it is governed by the procedural requirements of the Rules Enabling Act.[334] In contrast, technology, and especially AI technology, changes at near-breakneck speed, and often is incorporated into routine use by individuals, organizations, corporations, and governments long before it is the subject of evidentiary scrutiny in a particular case. For this reason, it is

[332] *See, e.g.*, Wisconsin v. Loomis, *supra* note 145. The *Loomis* Court discussed AI technology in the context of due process challenges to its use during a sentencing, where the rules of evidence are inapplicable. *See, e.g.*, FED. R. EVID. 1101(d)(3). It therefore provides no real help in evaluating the standards to be used when AI evidence is being offered during trials where evidence rules do apply.

[333] Every state in the United States has adopted its own rules of evidence, some of which are identical or nearly identical to the Federal Rules of Evidence, and some of which differ in significant respects. Nevertheless, the evidentiary concepts that govern admissibility of AI evidence are fundamental and are found in all compilations of the rules of evidence. Further, the Federal Rules of Evidence are frequently cited as persuasive authority even in states that have evidence codes that differ from the Federal Rules. For that reason, the authors will refer to the Federal Rules of Evidence in this paper because of their national scope and their influence on state codifications of the rules of evidence.

[334] *See* 28 U.S.C. §§ 2072–2077. Section 2073 of the Rules Enabling Act (the "Enabling Act") authorizes the Judicial Conference of the United States Courts to appoint a standing committee on rules of practice, procedure, and evidence, and individual committees for the rules of civil, criminal, appellate, and bankruptcy procedure, and the rules of evidence. The meetings of the standing committee, as well as those of the individual committees, are open to the public, minutes are kept of their proceedings, and there must be sufficient advance public notice of committee meetings. When one of the individual committees recommends a new rule (or amendment) it must prepare a proposed rule (or amendment) and explanatory note. The standing committee reviews and approves proposed new rule (or amendment), and it then is transmitted to the U.S. Supreme Court for review and approval. Section 2074 of the Enabling Act requires the Supreme Court to transmit the proposed new rule (or amendment) to Congress not later than May 1 of the year in which a proposed new rule (or amendment) is to become effective. The proposed new rule (or amendment) then takes effect on December 1 of that year, unless revised or rejected by Congress. *See id.*

not an unfair question to ask whether the Federal Rules of Evidence provide an adequate analytical framework to evaluate whether AI evidence ought to be admitted in court cases.[335]

But the Federal Rules of Evidence are nothing if not resilient, and they are designed to be used in a manner that is not static or inflexible. Rule 102 provides: "These rules should be construed so as to administer every proceeding fairly, eliminating unjustifiable expense and delay, and *promote the development of evidence law*, to the end of ascertaining the truth and securing a just determination." (emphasis added).[336] As this paper argues, the existing Federal Rules of Evidence are adequate for the task of evaluating AI evidence, provided they are applied flexibly.

We will start with the rules that define what relevant evidence is, then discuss the rules that govern how to authenticate evidence, and, finally, focus on the rules that govern how to admit scientific, technical, and specialized evidence. In the process, we will focus primarily on the evidentiary issues associated with relevance and authenticity, the two areas that create most of the evidentiary challenges for admitting AI evidence. Other evidence doctrines, such as the hearsay rule,[337] and the original writing rule,[338] can be encountered, but these rules present less of a concern than authenticity. Why? Because the focus of the hearsay rule is intentionally assertive statements made by human declarants,[339] and AI applications, by their very nature, involve machine-generated output.[340] While the evidence may, and

---

[335] *See* Lorraine v. Markel Am. Ins. Co., 241 F.R.D. 534, 542–43 (D. Md. 2007) (courts have rejected arguments calling for abandoning the existing rules of evidence and adopting more demanding rules to govern admissibility of electronic evidence). *See also* Michael M. Martin, Stephen A. Salzburg, and Daniel J. Capra, 5 *Federal Rules of Evidence Manual* § 901.02[9], at 901–19 (12th ed. 2019) (noting that the "basic authentication principles . . . [of the Fed. R. Evid.] have been found to be sufficiently adaptable to all forms of electronic evidence.").

[336] FED. R. EVID. 102.

[337] *See* FED. R. EVID. 801–07.

[338] *See* FED. R. EVID. 1001–08.

[339] *See* FED. R. EVID. 801(a)–(c).

[340] "Because human design, input, and operation are integral to a machine's credibility, some courts and scholars have reasoned that a human is the true 'declarant' of any machine conveyance. But while a designer or operator might be partially epistemically or morally responsible for a machine's statements, the human is not the sole source of the claim. . . . The machine is influenced by others but is still a source whose credibility is at issue." Andrea Roth, *Machine Testimony*, 127 Yale L.J. 1972, 1978–79 (2017). While it may be a useful analogy to compare the factually assertive output of an AI algorithm as a "statement," akin to one made by a human declarant, for purposes of stressing the importance of not accepting algorithmic output without critical analysis, this analogy has its limits. First, algorithms, unlike human beings, cannot intentionally "lie," they have no "demeanor" that a jury can evaluate for clues of deception or candor, and they cannot be subjected to an "oath" to impress upon them the duty to be truthful. Therefore, anthropomorphically characterizing the results of AI programs as having potential "credibility" problems adds little to what lawyers and judges must consider in deciding whether AI evidence may be considered by a jury. At its root, the hearsay rule is intended to promote the reliability

often does, take the form of an express or implied factual assertion (*e.g.*, "this is the photo of the person depicted in the surveillance video"; "this is the sector of the city that is likely to have the greatest potential for criminal activity on a particular date and time"; "this job applicant is most qualified for the vacancy being filled"), and may be offered for its substantive truth, the source is not a *human* declarant, therefore it is not properly regarded as hearsay.[341] Rather, the key issue is *authenticity*—how accurately does the AI system that generated the evidence produce the result that its proponent claims it does. Similarly, the original writing rule imposes a requirement that proof of the content of writings, recordings, and photographs must be made by introducing an original or duplicate original,[342] but those terms are defined interchangeably, and broadly, so they are seldom difficult to comply with, unless a witness is called who merely describes what he or she observed as the output of the AI system, instead of introducing a copy. This seldom occurs for the simple reason that having a human describe the contents of the output of an AI system that produces a written, recorded, or photographic result robs it of most of the weight that the evidence would have if the jury were shown the output itself (once properly authenticated).

B. Relevance

Federal Rule of Evidence 401 defines relevance. It states: "Evidence is relevant if: (a) it has any tendency to make a fact more or less probable than

---

of testimonial evidence, and the many hearsay exceptions all share a common denominator of being sufficiently reliable and accurate to allow the jury to consider them without the need to have the human declarant appear before them to assess credibility. If validity and reliability are the common goals, then, at least for AI, it is much more usefully analyzed under the lens of the authenticity rules, and the rules governing admissibility of evidence regarding experts, than by strained analogies to the hearsay rule.

[341] *See, e.g.*, U.S. v. Wallace, 753 F.3d 671, 675 (7th Cir. 2014) (rejecting confrontation-clause challenge to the admissibility of a video recording showing an exchange of drugs between two people because there was no human declarant to be cross examined and there was no showing that the conduct involved was intended by the participants to be an assertion, therefore there was no hearsay "statement," as contemplated by Fed. R. Evid. 801(a), and no "declarant," as contemplated by Fed. R. Evid. 801(b)); U.S. v. Lizarraga-Tirado, 789 F. 3d 1107, 1109-10 (9th Cir. 2015) (rejecting hearsay challenge to a satellite image and accompanying GPS coordinates. The Court found that the satellite image, exclusive of any labels and markers, was not hearsay because it contained no "assertion," as Fed. R. Evid. 801(a) requires. Similarly, because the geolocation coordinates of a particular point on the image was identified by a "tack," it was not hearsay since it was automatically generated by the Google Earth program. The Court held that "[a] tack placed by the Google Earth program and automatically labeled with the GPS coordinates isn't hearsay," because it contains no "statements" made by a "human" declarant.). These same analyses apply with equal force to the content and output of AI systems. *See also* 31 Charles A. Wright and Victor J. Gold, *Federal Practice and Procedure: Evidence* §7103, at 4 (Supp. 2018) ("While machine produced evidence like a readout from a global positioning system raises an issue under Rule 901, it does not also raise a hearsay issue because such evidence does not contain the statement of a person.").

[342] *See* FED R. EVID. 1001(e) (defining duplicates and duplicate originals), 1002 (setting forth the substantive rule), and 1004–1007 (setting forth exceptions to the rule).

it would be without the evidence; and (b) the fact is of consequence in determining the action." This is a relatively low bar to admitting evidence, because even evidence that has slight tendency to prove or disprove facts that are important to resolve a civil or criminal case meet this standard.[343] Examined in isolation, it could be argued that AI evidence that has not adequately been examined to determine its validity and reliability still has some tendency to prove a disputed issue. Rule 401 does not require perfection, only a tendency to prove or disprove.

But Rule 401 must not be read in isolation; it must be considered in conjunction with its evidentiary neighbors, Rules 402 and 403. Rule 402 states: "Relevant evidence is admissible unless any of the following provides otherwise: the United States Constitution; a federal statute; these rules [of evidence]; or other rules prescribed by the Supreme Court. Irrelevant evidence is not admissible."[344] In essence, Rule 402 creates a presumption that relevant evidence is admissible, even if it is only minimally probative, unless other rules of evidence or sources of law require its exclusion. But, while the first part of Rule 402 is flexible, the second part is immutable: Irrelevant evidence is never admissible.

Rounding out Rules 401 and 402 is Rule 403, which is designed to level the evidentiary playing field. It provides: "The court may exclude relevant evidence if its probative value is substantially outweighed by a danger of one or more of the following: unfair prejudice, confusing the issues, misleading the jury, undue delay, wasting time or needlessly presenting cumulative evidence."[345] As it relates to the admissibility of AI evidence, Rule 403 has three important features. First, it establishes a "balancing test" for determining whether relevant evidence may be considered by the judge or jury. This scale "tilts" towards admissibility of relevant evidence.[346] It is inadmissible only if its probative value (*i.e.*, its ability to prove or disprove important facts presented in a case) is substantially outweighed by the adverse consequences listed in the rule. It is not enough that relevant evidence will be prejudicial to the party against which it is introduced—after all, all evidence offered by a plaintiff against a defendant is intended to be

---

[343] *See, e.g.*, MICHAEL M. MARTIN ET AL., 1 FEDERAL RULES OF EVIDENCE MANUAL § 402.02[1] 401, 406–7 (12th ed. 2019) ("To be relevant it is enough that the evidence has a *tendency* to make a consequential fact even the least bit more probable or less probable than it would be without the evidence. The question of whether relevance is thus different from whether evidence is *sufficient* to prove a point. . . . It should be emphasized that 'any tendency' is enough. The fact that the evidence is of weak probative value does not make it irrelevant.") (emphasis in original)).

[344] FED. R. EVID. 402.

[345] FED. R. EVID. 403.

[346] *See, e.g.*, United States v. Terzado-Madruga, 897 F. 2d 1099, 1117 (11th Cir. 1990) (The balancing test of Fed. R. Evid. 403 "should be struck in favor of admissibility.").

prejudicial in the sense that it is offered to show that the defendant is liable. It is excludable only if its prejudice is *unfair* to that party.[347] Similarly, Rule 403 will tolerate a degree of confusion on the part of the judge or jury that must evaluate the evidence, even if it tends to mislead them, provided that these adverse consequences do not substantially outweigh the tendency of the evidence to prove important facts in the case. But even though the balancing in Rule 403 favors admissibility, the fact that the rule clearly establishes that judges must consider unfairness, be aware that confusion may result, and be careful to discern whether the jury may be misled, is extremely important, especially when applied to the admissibility of AI evidence. After all, the court cannot evaluate technical evidence for prejudice, confusion, or assess whether it misleads without understanding how it works. And judges cannot assess whether a jury will be misled or confused by AI evidence unless they have an appreciation for whether the AI application meets acceptable standards of validity and reliability, which may differ depending on what the evidence is being offered to prove, and the adverse consequences flowing from allowing a jury composed of lay persons to consider that evidence in reaching its verdict.

Second, Rule 403 makes it clear that it is the trial judge who is charged with the responsibility of reviewing the evidence in the first instance to determine whether the jury may hear it. This obligation flows from another rule of evidence, Rule 104(a), which states: "The court must decide any preliminary question about whether a witness is qualified, a privilege exists, or evidence is admissible. In so deciding, the court is not bound by evidence rules, except those on privilege."[348] This is all well and good, but implicit in this delegation of responsibility is the notion that the judge must have the tools to make this preliminary determination. The hallmark feature of the American justice system is that it is an adversary process. This means that it is the responsibility of the parties, not the judge, to develop and present the factual evidence that will be offered to the jury for its consideration. When it comes to technical evidence like AI, the judge often is in a battle of wits unarmed, as the court is not involved in the investigation of the facts

[347] *See* United States v. Guzman-Montanez, 756 F.3d 1, 7 (1st Cir. 2014) ("[T]he law shields a defendant against unfair prejudice not against all prejudice. '[A]ll evidence is meant to be prejudicial; it is only unfair prejudice which must be avoided.'"); Martin, *supra* note 343, § 403.02[3], at 403,410–11 ("Evidence is not 'prejudicial' merely because it is harmful to the adversary. After all, if it didn't harm the adversary, it wouldn't be relevant in the first place. Rather, the rule refers to the negative consequences of 'unfair' prejudice. Unfair prejudice is that which could lead the jury to make an emotional or irrational decision, or to use the evidence in a manner not permitted by the rules of evidence.").

[348] FED. R. EVID. 104(a). The party introducing the evidence bears the burden of proving that the offered evidence meets the requirements of Rule 104(a) by a preponderance of the evidence. *See* Martin, *supra* note 343 § 104.02[9], at 104–12.

underlying a case, or the marshalling of evidence to prove or disprove it. What this means is that it is the obligation of lawyers who intend to offer (or challenge) AI evidence to do the hard work necessary to show the judge how the AI system works (*i.e.*, produced its output), why the evidence will enlighten not confuse, and promote a just outcome, not one that is unfair. To do this, they must understand the AI system and its output themselves, and that can be a challenge for lawyers who more often than not are generalists, not specialists in the many scientific and technical disciplines that underlie AI systems and their related evidence.

For their part, the trial judge must raise with the parties well in advance of the trial the question of whether they intend to offer AI or similarly technical evidence at trial, and as part of the pretrial scheduling process, impose deadlines for disclosing an intention to introduce such evidence, and for challenging its admissibility sufficiently far in advance of trial to allow the judge to have a hearing (which may require the testimony of witnesses). Determinations about whether AI evidence meets adequate thresholds of validity and reliability sufficient for it to be considered by the jury do not lend themselves to last minute, on-the-fly assessments, and should not be attempted or allowed in the middle of a trial itself.

Finally, it should be obvious that a judge cannot make the determinations required by Rules 401 through 403 unless the party offering the AI evidence is prepared to disclose underlying information concerning, for example, the training data and the development and operation of the AI system sufficient to allow the opposing party (and the judge) to evaluate it, and the party against whom the AI evidence will be offered to decide whether and how to challenge it. If a party intends to rely on facts that are the product of AI applications in a civil or criminal trial, they should not be permitted to withhold from the party against whom that evidence will be offered the information necessary to determine the validity (*i.e.*, the degree of accuracy with which the AI tool measures what it purports to measure), and the reliability (*i.e.*, the consistency with which the AI algorithm correctly measures what it purports to measure), of the AI evidence. If they are prohibited from doing so by the claims of proprietary information or trade secrets raised by the company that developed the AI application, the trial judge should give the proponent of the AI evidence a choice: disclose the underlying evidence (under the provisions of an appropriate protective order), or otherwise demonstrate its validity and reliability. If the proponent is unwilling or unable to do so, they should be precluded from introducing the evidence at trial.[349]

[349] In addition to evidentiary concerns associated with admitting AI evidence against a party that has been denied sufficient information with which to assess its validity and reliability, this can also raise

The long and the short of it is not hard to grasp. Invalid or unreliable AI systems produce results that have insufficient tendency to prove or disprove disputed facts in a trial. Neither the trial judge nor the party against whom AI evidence is offered should be required to accept at face value the unproven claims of the proponent of the evidence that it is valid and reliable. This takes us to the next important area where the Federal Rules of Evidence provide guidance: the process of authentication.

C. Authentication of AI Evidence

Federal Rule of Evidence 901(a) sets forth in plain terms what is meant by the requirement that AI evidence must be authenticated in order to be considered by the jury. It states: "To satisfy the requirement of authenticating . . . an item of evidence, the proponent must produce evidence sufficient to support a finding that the item is what the proponent claims it is."[350] Rule 901(b) then lists ten non-exclusive ways in which a party can

procedural due process issues if the proponent of the evidence is a government entity. In *Houston Fed. of Teachers, Local 2415 v. Houston Ind. Schl. Dist.*, *supra* note 98, the Court denied the school district's motion for summary judgment on the plaintiffs' procedural due process claims largely because the plaintiff school teachers had been "denied access to the computer algorithms and data necessary to verify the accuracy of their [teacher evaluation] scores." *Id.* at 1177. The school district used an AI-based evaluation system developed by a third-party vendor to evaluate teacher performance in order to determine whether to renew the employment of public school teachers. *Id*. The vendor claimed that the algorithms and related software were trade secrets and refused to allow the plaintiffs the ability to test their validity. *Id.* The Court concluded that the inability of the teachers to ensure the correct calculation of their evaluation scores exposed them to the risk of "mistaken deprivation" of their jobs and refused to grant summary judgment to the school district on the teachers' procedural due process claims. *Id.* at 1180. Similarly, in a more recent opinion, the Superior Court of New Jersey, Appellate Division, rejected claims of trade-secret protection as a bar to producing source code to permit the defendant in a criminal case to evaluate the validity and reliability of the State's DNA analysis software used to prove that the defendant's DNA was present, reversing the decision of the trial judge that blocked the disclosure of the source code. The Court held that "[w]ithout . . . [access to the source code] defendant is relegated to blindly accepting the company's assertions as to its reliability. *And, importantly, the judge would be unable to reach an informed reliability determination . . . as part of his gatekeeping function.* Hiding the source code is not the answer. The solution is producing it under a protective order." State v. Pickett, 466 N.J. Super. 270, 246 A.3d 279 (App. Div. 2021) (emphasis added)). Compare these two cases with the decision in *Wisconsin* v. *Loomis*, *supra* note 145, where the Wisconsin Supreme Court rejected due process challenges to the use of the AI-powered COMPAS system for evaluating defendant recidivism risk for purposes of sentencing defendants. *Id.* at 271 ¶86. In *Loomis*, the Court was unpersuaded that the defendant had been denied access to information necessary to evaluate the validity of the COMPAS software, on similar claims of proprietary trade secrets. *Id*. at 257–64 ¶¶46–65. In light of the discussion in this article, it is our view that the *Loomis* Court unwisely dismissed the defendant's legitimate challenges to the validity and reliability of the COMPAS system, while the *Houston Fed. of Teachers* and *Pickett* Courts correctly recognized the inherent unfairness associated with allowing claims of trade secrets to preclude litigants from testing the validity and reliability of critical AI evidence that is being offered against them. In *Pickett*, the Court cogently explained why the trial judge, as well as the party against whom the electronic evidence will be offered, needs this information to rule on its accuracy.

[350] FED. R. EVID. 901(a).

accomplish this task.[351] The examples that most readily lend themselves to authenticating AI evidence are: Rule 901(b)(1) (testimony of a witness with knowledge that an item is what it is claimed to be); and Rule 901(b)(9) (evidence describing a process or system and showing that it produces an accurate result).

When authenticating AI evidence using Rule 901(b)(1), the testimony of the witness called to accomplish this task must comply with other rules of evidence. For example, Rule 602 requires that the authenticating witness have personal knowledge of how the AI technology functions.[352] It states: "A witness may testify to a matter only if evidence is introduced sufficient to support a finding that the witness has personal knowledge of the matter. Evidence to prove personal knowledge may consist of the witness's own testimony. This rule does not apply to a witness's expert testimony under Rule 703."[353]

There are some important features of Rule 602 that tend to be overlooked by some lawyers and judges. There is an understandable tendency to call the fewest number of witnesses as possible to authenticate evidence. When a single person possesses all the knowledge needed to do so, then that is all that is required. But if this paper has shown anything, it is that AI applications seldom are the product of a single person possessing personal knowledge of all the facts that are needed to demonstrate that the technology and its output are what its proponent claims them to be. Data scientists may be required to describe the data used to train the AI system. Developers may be required to explain the features and weights that were chosen for the machine-learning algorithm. Technicians knowledgeable about how to operate the AI system may be needed to explain what they did when they used the tool, and the results that they obtained. These technicians, however, may be entirely at sea when asked to explain how the data was

---

[351] *See* FED. R. EVID. 901(b)(1)–(10).

[352] *See* 31 Charles A. Wright & Victor J. Gold, *Federal Practice and Procedure: Evidence* §7103 24–25 (1st ed. 2000), which states that "[f]or purposes of analyzing the scope of Rule 901, the most important additional relationship is the one between that provision and Rule 602. . . . Both Rules 602 and 901 identify elemental qualities that make evidence worthy of consideration. Since the provisions perform similar functions, it is important to know when evidence is subject to the personal knowledge requirement of Rule 602 and when it is subject to the authentication or identification requirement of Rule 901. Rule 602 applies only to testimonial evidence. . . . Rule 901 does not apply to testimonial evidence; it applies to all other evidence. The distinction can be misleading, however, because it might be taken to suggest that Rules 602 and 901 never apply to the same evidence. In fact, these provisions are simultaneously applied where testimony is the means by which some respect of non-testimonial evidence is relayed to the jury."; *See, also id*. at 25, n.33 ("Further, perhaps the most common way to establish authenticity or identity is with testimony that satisfies the personal knowledge requirement of Rule 602. *See Rule 901(b)(1).*" (emphasis added)).

[353] FED. R. EVID. 602.

collected or cleansed, how the algorithm that underlies the AI system was programmed, or how the system was tested to show that it produces valid and reliable results. An example illustrates this nicely.

As mentioned above, a Canadian company named BlueDot developed an algorithm that allowed it to examine data from a large number of publicly available sources—as varied and diverse as medical bulletins, livestock reports, and airline flight information—enabling it to accurately predict, as early as December, 2019, where the COVID-19 virus would spread.[354] Development of the algorithm required a team that included, among other disciplines, engineers, ecologists, geographers, and veterinarians.[355] Once developed, the algorithm had to be trained for over a year to learn how to detect 150 pathogens.[356] If evidence derived from use of the BlueDot algorithm was being offered into evidence at trial, the party seeking to introduce it would be required to show how it could accurately and reliably accomplish what its developers claimed it could. Given the number of specialties involved in the tool's development, and the length and complexity of the process by which it was "trained" to analyze data from so many disparate sources, it is difficult to imagine how a single person would be able to testify from personal knowledge in order to do so.

Of course, Rule 602 would not require authentication by a single person possessing personal knowledge of all of the information needed to authenticate the BlueDot's AI technology, if the person chosen for this task qualified as an expert witness under Rules 702 and 703.[357] Rule 702 provides that: "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."[358]

---

354 *See* Bill Whitaker, *supra* note 94.

355 *See id.*

356 *See id.*

357 *See* Charles A. Wright & Victor J. Gold, *supra* note 352 §7103, at 26, stating that "[t]he connections between Rule 901 and the rules governing opinion evidence are also of consequence. Rules 701 and 702 impose general limits on the admissibility of lay and expert opinion testimony. . . . Rule 901(b) seems to assume that opinion evidence may be admitted under Rules 701 and 702 in certain limited contexts. . . ." One such context is Rule 901(b)(3), which provides that authentication may be accomplished by "[a] comparison with an authenticated specimen by an expert witness or the trier of fact." Similarly, Rule 901(b)(5) states that authentication of the identity of a person's voice may be accomplished by "[a]n opinion identifying a person's voice."

358 FED. R. EVID. 702.

Importantly, Rule 703 states that: "An expert may base an opinion on facts or data in the case that the expert has been made aware of or personally observed. If experts in the particular field would reasonably rely on those kinds of facts or data in forming an opinion on the subject, they need not be admissible for the opinion to be admitted."[359] If the requirements of Rules 702 and 703 were met, then, a party that wanted to authenticate an AI system that was developed by a team of individuals with scientific, technical, or specialized knowledge beyond the personal knowledge of any one person could do so with a single qualified expert. But that is a big "if," because, as will be seen, the requirements of Rules 702 and 703 are quite demanding when applied as intended by the Federal Rules of Evidence.

The key takeaway point is that lawyers must keep in mind, and judges must be vigilant to require, that the person or persons called to authenticate AI evidence either have personal knowledge of the authenticating facts or qualify as an expert that is permitted to incorporate into their testimony information from sources beyond their own personal knowledge, provided it is sufficiently reliable.[360]

The second authenticating rule most suited to AI evidence is Rule 901(b)(9). It permits authentication by "[e]vidence describing a process or system and showing that it produces an accurate result."[361] Of course, to do so, the party that wishes to introduce the AI evidence would face the exact challenges just described in the discussion of Rule 901(b)(1)—calling a single person or persons themselves possessing personal knowledge of all the authenticating facts or qualifying as an expert under Rules 702 and 703.[362]

---

[359] FED. R. EVID. 703.

[360] *See, e.g.*, Fed. R. Evid. 703. *See also* United States v. Frazier, 387 F.3d 1244, 1260 (11th Cir. 2004), for a discussion of the importance of a trial judge to diligently fulfill their "gatekeeping" function under Fed. R. Evid. 104(a) to ensure the "reliability and relevancy of expert testimony" because an expert's opinion "can be both powerful and quite misleading because of the difficulty in evaluating it." The Court in *Frazier* noted that "[i]ndeed, no other kind of witness is free to opine about a complicated matter without any firsthand knowledge of the facts in the case and based upon otherwise inadmissible hearsay if the facts or data are 'of a type reasonably relied upon by experts in the particular field in forming opinions or inferences upon the subject.'" (internal citations omitted)); *Cooper* v. *Smith & Nephew, Inc.*, 259 F.3d 194, 199 (4th Cir. 2001)("While Rule 702 was intended to liberalize the introduction of relevant expert evidence, courts 'must recognize that due to the difficulty of evaluating their testimony, expert witnesses have the potential to be both powerful and quite misleading.'") (internal citation omitted)).

[361] FED. R. EVID. 901(b)(9).

[362] There are two additional rules of evidence that may be used to authenticate AI evidence that are closely related to Rules 901(b)(1) and 901(b)(9). They are Fed. R. Evid. 902(13), which allows authentication of "[a] record generated by an electronic process or system that produces an accurate result, as shown by a certification of a qualified person"; and Fed. R. Evid. 902(14), which allows authentication of "[d]ata 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." Rules 902(13) and (14) would allow the proponent of AI evidence to authenticate it by substituting the certificate of a qualified witness

An important feature of authentication needs careful consideration in connection with admitting AI evidence. Normally, a party has fulfilled its obligation to authenticate non-testimonial evidence by producing facts that are sufficient for a reasonable factfinder to conclude that the evidence more likely than not is what the proponent claims it is.[363] In other words, by a mere preponderance. This is a relatively low threshold—51%, or slightly better than a coin toss.[364] However, as we have shown in this paper, not all AI evidence is created equal. Some AI systems have been tested and shown to be valid and reliable. Others have not, when, for example, efforts to determine their validity and reliability have been blocked by claims of proprietary information or trade secret. Furthermore, some of the tasks for which AI technology has been put to use can have serious adverse consequences if it does not perform as promised—such as arresting and criminally charging a person based on flawed facial recognition technology or sentencing a defendant to a long term of imprisonment based on an AI system that has been trained using biased or incomplete data that inaccurately or differentially predicts the likelihood that the defendant will reoffend.

The greater the risk of unacceptable adverse consequences, the greater the need to show that the AI technology is unlikely to produce those consequences. Judges, tasked with making the initial determination of admissibility of AI evidence under Rule 104(a), should be skeptical of admitting AI evidence that has been shown to be accurate by no more than an evidentiary coin toss. They should insist that the proponent of the evidence establish the validity and reliability of the AI to a degree that is commensurate with the risk of the adverse consequences likely to occur if the technology does not perform as claimed. And if the proponent of the evidence fails to do so, then the trial judge should evaluate under Rule 403

for their live testimony. But it must be stressed that the qualifications of the certifying witness and the details of the certification that the evidence produces an accurate and reliable result must be the same as would be required by the in-court testimony of a similarly qualified witness. Rules 902(13) and (14) are not invitations for boilerplate or conclusory assertions of validity and reliability and should not be allowed to circumvent the need to demonstrate, not simply proclaim, the accuracy and reliability of the system or process. *See, e.g.*, Wright & Gold, supra note 352 §7147, at 43, stating that "[n]ewly adopted Rule 902(13)] allows the authenticity foundation that satisfies Rule 901(b)(9) [process or system producing accurate results] to be established by a certification rather than the testimony of a live witness. If the certification provides information that would be insufficient to authenticate the record if the certifying person testified, then authenticity is not established under Rule 902(13)." The same applies for the certification in Rule 902(14), certified data copied from an electronic device, storage medium, or file.

[363] *See, e.g.*, Lorraine v. Markel Am. Ins. Co., *supra* note 335 at 542; United States v. Safavian, 435 F.Supp.2d. 36, 38 (D.D.C. 2006); United States v. Holmquist, 36 F.3d 154, 168 (1st Cir. 1994) ("the standard for authentication, and hence for admissibility, is one of reasonable likelihood.").

[364] *See, e.g.*, Martin, *supra* note 343 § 901.02[1], at 901–07 ("[The requirement to authenticate or identify evidence imposed by Rule 901(a)] is a mild standard—favorable to admitting the evidence.").

whether the probative value of AI authenticated by a mere preponderance is substantially outweighed by the danger of unfair prejudice to the adverse party or would confuse or mislead the jury to an unacceptable degree,[365] taking into consideration the nature of the adverse consequences that could occur if the AI technology is insufficiently accurate or reliable.

What is the best, fairest way to do so? We believe it is to employ Rule 102, which requires the rules of evidence to be "construed so as to administer every proceeding fairly . . . and promote the development of evidence law"[366] to "borrow" from Rule 702 and the cases that have interpreted it, when determining the standard for admitting scientific, technical, or other specialized information that is beyond the understanding of lay jurors and generalist judges. These factors are commonly referred to as the *Daubert* factors and are discussed next.

D. Usefulness of the Daubert Factors in Determining Whether to Admit AI Evidence

As previously noted, Federal Rule of Evidence 702 requires that introduction of evidence dealing with scientific, technical, or specialized knowledge that is beyond the understanding of lay jurors be based on a sufficient facts or data and reliable methodology that has been applied reliably to the facts of the particular case.[367] These factors were added to the evidence rules in 2000 to bolster the rule in light of the U.S. Supreme Court's decisions in *Daubert* v. *Merrell Dow Pharmaceuticals, Inc.*, 509 U.S. 579 (1993), and *Kumho Tire Co.* v. *Carmichael*, 119 S. Ct. 1167 (1999).[368] Therefore, while Rule 702 was not intended to codify the *Daubert* decision, the factors discussed in that decision relating to determining the reliability of scientific or technical evidence are quite informative when determining whether Rule 702's reliability[369] requirement has been met. As described in the Advisory Committee Note to the amendment of Rule 702 that went into effect in 2000, the "*Daubert* Factors" are: "(1) whether the expert's technique or theory can be or has been tested . . . ; (2) whether the technique

---

365 *See* FED. R. EVID. 403.

366 FED. R. EVID. 102.

367 *See* FED. R. EVID. 702 (b)-(d). *See also generally* In re Paoli R.R. Yard PCB Litig., 35 F.3d 717, 742 (3d Cir. 1994), which helpfully discusses the importance of the reliability factor in the *Daubert* analysis, and the obligation of the trial judge to "take into account" all of the factors listed in *Daubert* that are relevant to determining the reliability of the scientific or technical evidence that is being offered into evidence.

368 *See* Advisory Committee Note, FED. R. EVID. 702 (2000).

369 In legal parlance the "reliability" of scientific or technical evidence usually refers to its trustworthiness, as opposed to the narrower technical definition of "reliability" used in this paper. The legal concept of reliability encompasses both validity (*i.e.*, accuracy) and reliability (*i.e.*, consistency across similar circumstances).

or theory has been subject to peer review and publication; (3) the known or potential rate of error of the technique or theory when applied; (4) the existence and maintenance of standards and controls; and (5) whether the technique or theory has been generally accepted in the scientific [or technical] community."[370]

The usefulness of borrowing these factors in assessing whether AI evidence should be admitted is readily apparent. To authenticate AI technology, its proponent must show that it produces accurate, that is to say valid, results. And it must perform reliably, meaning that it consistently produces accurate results when applied in similar circumstances. When the accuracy and reliability of technical evidence has been verified through independent testing and evaluation of the AI system that produced it, the methodology used to develop the evidence has been published and subject to review by others in the same field of science or technology, when the error rate associated with the AI system use is not unacceptably high, when the standard testing methods and protocols have been followed, and when the methodology used is generally accepted within the field of similar scientists or technologists, then it has been authenticated. It does what its proponents say it does. And introducing it produces none of the adverse consequences that Rule 403 is designed to guard against.

In contrast, when the validity and reliability of the system or process that produces AI evidence has not properly been tested, when its underlying methodology has been treated as a trade secret by its developer preventing it from being verified by others, when applying the method produces unacceptably high error rates, when corners were cut and standard procedures were not followed when it was developed or employed, or when the methodology is not accepted as reliable by others in the same field, then it is hard to maintain with a straight face that it does what its proponent claims it does, which ought to render it inauthentic and inadmissible. The bottom line is that if a lawyer intends to rely on AI evidence to prove their case, they would be foolish not to consider these five factors and marshal the facts to show compliance with as many of them as they can. And if the reader is a judge that takes seriously their obligation to employ the rules of evidence during a trial "to the end of ascertaining the truth and securing a just determination,"[371] they will insist that the party offering evidence produced by an AI system to prove its case adequately has shown that it does what its proponent claims it does, to a degree of certainty commensurate with the risk of an unacceptably bad outcome if it turns out that the technology was unreliable. Failing that, the AI evidence should be excluded for insufficiency

[370] *See* Advisory Committee Note, *supra* note 368.

[371] FED. R. EVID. 102.

of authentication (Rule 901(a)), failure to show the use of reliable methodology that was replied applied to the facts of the case (Rule 702), and/or excessive danger of unfair prejudice, or of confusing or misleading the jury (Rule 403).

E. Practice Pointers for Lawyers and Judges

If both lawyers and judges accept that there are multiple types and uses of AI, and that there are many potential issues with it—for example, risk of bias, lack of robust testing and validation, function creep, potential lack of transparency and explainability, and possible lack of resilience—which can all affect the validity and reliability of AI evidence, and they recognize the need to authenticate it properly before it is admitted into evidence (and the need to follow the rules that govern how to do so), then the question arises: How should lawyers faced with introducing or challenging AI evidence, and judges who must rule on its admissibility, go about doing so? Below, we offer some practical suggestions with the hope that they will make this task less daunting in practice.

1. What problem was the AI created to solve?

As we have shown, the essence of AI technology comes down to the data and the algorithm or algorithms that were developed to govern it. Algorithms are a set of rules or procedures for solving a problem or accomplishing an end. So, the starting place for determining the admissibility of AI technology is to define the problem that the AI was designed to solve. Knowing this is essential to assessing the validity of the system (*i.e.*, its accuracy in performing these functions); its reliability (*i.e.*, the consistency with which it produces the same or substantially similar results when applied under substantially similar circumstances); and whether it is being used for purposes for which it was not designed (*i.e.*, there has been substantial function creep). The proponent of the evidence needs to know its design objective in order to advance the evidence necessary to secure its admissibility. Opposing parties need to know this information to be able to intelligently assess whether its admissibility may be challenged. And judges need to know this to be able to rule on the admissibility of the evidence derived from the AI system. Relevance is not an abstract concept. Evidence is relevant only to the extent that it has the ability to prove or disprove facts that are consequential to the resolution of a case. The problem that the AI was developed to resolve—and the output it produces—must "fit" with what is at issue in the litigation. Without knowing what the AI was designed and programmed to do, none of these fundamental questions can be answered.

2. How was the AI developed, and by whom?

One of the issues that affects the validity and reliability of AI evidence is whether its design was influenced by intended or unintended bias. Was the data used to train the AI representative, or skewed? Is it representative of the proper target population? If not trained with overtly discriminatory data, were discriminative proxies used in the training process? What assumptions, norms, rules, or values were used to develop the system? Were the people who did the programming themselves sufficiently qualified or experienced to ensure that there was not inadvertent bias that could impact the validity and reliability of the output of the system? Have the programmers given due consideration to the population that will be affected by the performance of the system? It does not require Napoleonic insight to realize that these questions cannot be answered without knowledge about the details of the data that was used as input for purposes of training, how the AI system was developed, including the design choices that were made, how the system was operated, and how the output was interpreted. When the party offering the output of an AI system into evidence thwarts efforts to obtain this information by asserting that it is proprietary or a trade secret, this should be a red flag for both the adverse party and the court. And judges should be particularly careful not to allow a party planning to introduce AI evidence to hide behind claims of proprietary information or trade secrets without careful consideration of the consequence to the party against whom the AI evidence will be offered. Will allowing trade-secret claims to shield disclosure of how the AI evidence was developed, trained, and functions prevent the party against whom it will be introduced from having a fair opportunity to learn how the AI works so they can prepare a defense? If so, how are they to frame evidentiary challenges to its use? Adverse parties who are refused access to the information they need to assess AI's validity and reliability on the basis of claims of trade secrets should challenge these designations and seek a ruling from the court that either grants them access to the information that they reasonably need (subject to proper protective measures,) or prohibits the introduction of the AI evidence at trial. And judges must ask themselves how they can fulfill their gatekeeping role in ruling on the admissibility of the AI evidence if presented with little more than a "black-box" AI program and a conclusory claim that it consistently functions as it was designed to.

3. Was the validity and reliability of the AI sufficiently tested?

We have repeatedly stressed the importance of the concepts of validity and reliability in assessing whether AI evidence should be admitted as evidence. The proponent of AI evidence should be required to demonstrate that the AI system that produced the evidence being offered has been tested (preferably independently) to confirm that it is both valid for the purpose for

which it is being offered, and reliable. If it was not tested, why not? And why should the court even consider allowing the introduction of the output of an untested AI system? Who designed and carried out the testing? Was it the same people who developed the system in the first place? If so, was the methodology used to test the system standard or otherwise reasonable, adhering to procedures accepted as appropriate by the relevant scientific or technological community familiar with the subject matter at the heart of the AI system? Under what conditions did the testing occur, and how do they compare to the circumstances under which the system is now being used? Was the system tested both for validity and reliability? Has the validity and reliability been confirmed by others who are independent of the developers? Are the results of the testing still available so that they may be reviewed by the adverse party and the court? The answers to these questions should inform the court's decision as to whether the evidence should be admitted at all. Allowing the introduction of AI evidence that has not been shown to be valid and reliable for the purpose for which it is being introduced substantially increases the risk that its probative value (if any) is substantially outweighed by the danger of unfairly confusing or misleading the factfinder. This is particularly so if the AI evidence is the primary evidence being offered to prove an essential element of the proponent's case.

4. Is the manner in which the AI operates "explainable" so that it can be understood by counsel, the court, and the jury?

As we discussed earlier, an important factor in evaluating the admissibility of AI evidence is whether the functioning of the system that produced it can be explained to lay persons unfamiliar with the technology and methodology involved, so they can understand how the system operates, how it achieves its results, and thus, evaluate the amount of weight they are willing to give to it. Recall our earlier discussion of "XAI" ("Explainable AI") and the principles advanced by the National Institute of Standards and Technology.[372] In NIST's draft publication titled *Four Principles of Explainable Artificial Intelligence*, the authors explained why it is important for the developers of AI programs to be able to explain to others—even if only in general terms—how they work. Notably, they stated:

With recent advances in artificial intelligence (AI), AI systems have become components of high-stakes decision processes. The nature of these decisions has spurred a drive to create algorithms, methods, and techniques to accompany outputs for AI systems with explanations. This drive is motivated in part by laws and regulations which state that decisions including those from automated systems, provide information

372 *See* discussion *supra* at page 61; Phillips et. al., *supra* note 241.

about the logic behind those decisions and the desire to create trustworthy AI.[373]

Based on these calls for explainable systems, it can be assumed that the inability or failure to articulate an answer can affect the level of trust users will afford that system. Suspicions that the system is biased or unfair can raise concerns about harm to oneself and to society. This may slow societal acceptance and adoption of AI technology, as members of the general public oftentimes place the burden of meeting societal goals on manufacturers and programmers themselves. Therefore, in terms of societal acceptance and trust, developers of AI systems may need to consider that multiple attributes of an AI system can influence public perception of the system. Explainable AI is one of several properties that can increase trust in AI systems. "Other properties include resiliency, reliability, elimination of bias, and accountability."[374]

The four principles of explainable AI are defined as follows:

> Explanation: Systems deliver accompanying evidence or reason(s) for all outputs;
>
> Meaningful: Systems provide explanations that are understandable to individual users;
>
> Explanation Accuracy: the explanation correctly reflects the system's process for generating the output; and
>
> Knowledge Limits: The system only operates under conditions for which it was designed or when the system reaches a sufficient confidence in its output.[375]

Although written from the perspective of scientists interested in the development of valid and reliable AI methods, the discussion emphasizes the same themes that underlie the purpose of the rules of evidence: that when technical information is offered during a trial, the proponent of that evidence must demonstrate that it is sufficiently trustworthy for the jury to credit it in making its decision. If the proponent of the evidence cannot even explain how the AI system operates in a way that can be understood by the trier of fact (including assuring them that it only is being used under the conditions

[373] Phillips et al., *supra* note 241 at 1.

[374] *Id.*

[375] *Id.* at 2. (emphasis in original).

for which it was designed and that there is sufficient confidence in its accuracy), then the evidence produced from it should not be admitted by the court.

5. What is the risk of harm if AI evidence of uncertain trustworthiness is admitted?

As we have explained, the Federal Rules of Evidence do not require that all risk of error be eliminated before scientific and technical evidence may be admitted. After all, evidence is relevant if it has any tendency, however slight, to prove or disprove facts that are important to deciding a case.[376] And authenticity is established if the proponent demonstrates that the evidence more likely than not is what it purports to be.[377] The argument could be made that even AI evidence shown to be valid and reliable for a particular purpose, but which is being offered to prove something for which its validity and reliability have not been established, has some tendency to prove what it is being offered to prove.

The expert witness rules[378]—which we argue should inform the decision of whether AI evidence is admissible—are probably the most helpful rules for evaluating the admissibility of AI evidence because they supply demanding standards: (i) whether there is a sufficient factual basis to support the evidence; (ii) whether the methods and principles used to generate the evidence were reliable; and (iii) whether they were reliably applied to the facts of the particular case.[379] And the *Daubert* Factors further focus the inquiry on the following: (i) whether the methodology was tested; (ii) whether there is a known error rate; (iii) whether the methods used are generally accepted as reliable within the relevant scientific or technical community that is familiar with the methodology; (iv) whether the methodology has been subject to peer review by others knowledgeable in the field; and (v) if standard procedures or protocols are applicable to the methodology, whether they were complied with.[380] But even this enhanced level of analysis does not require perfection. The ultimate question that must be decided in each case is whether the evidence is sufficiently valid and reliable for the purpose for which it is being offered. The answer to this question will depend on what is at stake if the fact finder credits AI evidence that is invalid and unreliable. Two factual scenarios will help to illustrate the import of this question.

---

376 *See* FED. R. EVID. 401.

377 *See* United States v. Holmquist, 36 F. 3d 154, 168 (1st Cir. 1994).

378 *See* FED. R. EVID. 702–03.

379 *See* FED. R. EVID. 702.

380 *See* Daubert v. Merrell Dow Pharmaceuticals, Inc., 509 U.S. 579, 593–94 (1993).

Imagine a civil case for breach of contract that seeks money damages. There have been terabytes of electronic documents generated by the parties that are potentially relevant to the resolution of the dispute. Reviewing them all manually, by humans, would be too time consuming and costly. A party requested to produce "all documents relevant to the dispute" by its adversary uses an AI system known in the eDiscovery community as "technology-assisted review" or "TAR"[381] to search the records and to identify those that are responsive to the request for production, and those that are not. The party produces the records deemed responsive by the TAR system, subject to a review for privilege. The requesting party is not satisfied with the production, the parties are unable to reach agreement, and they take the dispute to the court. The producing party touts the accuracy of its TAR system; the requesting party reels off reasons why it thinks the search methodology was unreliable and the production is incomplete. The judge must decide. Undoubtedly, the court will consider the "proportionality factors" set forth in Fed. R. Civ. P. 26(b)(1),[382] including what is at stake in the litigation, how important the information is to resolving the issues in dispute, how much the TAR process already cost the producing party, what it would cost to require further TAR (or other search and review efforts), how much more complete and accurate the production might be if more TAR (or other search and review methods) were performed, what the parties resources are, and whether what was produced—even if it does not include all of the available responsive documents—is sufficient to allow the requesting party a fair opportunity to prove its case.[383] Depending on how the judge weighs these factors they may rule that the production is "good enough," even if imperfect, or they may require further TAR (or other search and review methods), and decide who must pay for it. But unless the initial production is so clearly deficient as to hamstring the requesting party's ability to prove its case, the risk of ruling that no further production is required is not catastrophic to the requesting party. Expressed differently, the production, though imperfect, is sufficient, and the possibility that some undiscovered but responsive documents might not have been produced is not

381 *See* Grossman & Cormack, *supra* note 56.

382 *See* FED. R. CIV. P. 26(b)(1) ("Parties may obtain discovery regarding any nonprivileged matter that is relevant to any party's claim or defense and proportional to the needs of the party's case, considering the importance of the issues at stake in the action, the amount in controversy, the parties' relative access to relevant information the parties' resources, the importance of the discovery in resolving the issues, and whether the burden or expense of the proposed discovery outweighs it is likely benefit.").

383 FED. R. CIV. P. 26(g) requires that the search inquiry must be reasonable, not perfect; an attorney's signature on a discovery response certifies that it was based on a "reasonable inquiry." *See also* FED. R. CIV. P. 26(b)(1) defines the scope of discovery and provides that parties "may obtain discovery regarding any nonprivileged matter that is relevant to any party's claim or defense and proportional to the needs of the case. . . ."

the end of the world for the requesting party—the circumstances of the case will allow a degree of risk that the TAR system did not locate every possible responsive document.

Now, contrast this situation with one where a judge is tasked with sentencing a criminal defendant who has been convicted of possession with the intent to distribute a controlled substance. The defendant has a history of mental health problems, substance abuse, and two prior drug convictions: one for simple possession and the other for distribution. In fashioning a sentence, the judge will consider a number of factors to arrive at a sentence that is sufficient, but not excessive: the nature and circumstances of the offense; the safety of the public; the need to deter the defendant and others from committing similar crimes in the future; the history and characteristics of the particular defendant; whether the sentence should include drug testing and mental health treatment to lessen the risk that the defendant will recidivate; and perhaps other relevant factors.[384]

At sentencing, the prosecution argues that a prolonged jail sentence is needed to protect the public and to deter the defendant from committing future drug offenses. The prosecutor relies on an evaluation of the defendant performed by the court's probation department, which used an AI system similar to the COMPAS system we have discussed at length in this paper. That evaluation compared the defendant's characteristics to a national database of criminal convictions and determined that the defendant is 70% likely to recidivate within two years of his release from prison, unless the sentence includes both mental health treatment and substance abuse treatment. Focusing on the 70% recidivism prediction, the prosecutor argues that the judge should incarcerate the defendant for an extended period of time. The defense attorney argues that the AI system was not designed to be used to recommend the length of the sentence of incarceration, but rather to determine what services should be included in the sentence to mitigate the risk of recidivism once the defendant has been released from custody. The defense attorney also points out that the data used to make the recidivism prediction was gathered from a national database, not one that was representative of convictions in the state where the case has been brought. Nor has the AI been independently validated, and the defense attorney was not allowed access to the information needed to test the validity and reliability of the AI system, and so on.

---

[384] *See, e.g.*, 18 U.S.C. § 3553(a). The sentencing factors include: the nature and circumstances of the offense and the history and characteristics of the defendant; the need for the sentence to reflect the seriousness of the offense, promote respect for the law and provide just punishment for the offense; to afford adequate deterrence to criminal conduct, to protect the public from further crimes of the defendant; to provide the defendant with needed educational or vocational training, medical care, or other correctional treatment in the most effective manner. *See id.*

The judge must decide whether to rely on the AI recidivism prediction when deciding how long a sentence of incarceration they should impose. If the AI system has not been shown to be valid and reliable for the purpose for which it is being offered (*i.e.*, determining the length of a jail sentence), and the defense has not had a fair opportunity to challenge its validity and reliability because the developer of the software successfully asserted trade-secret protection, then the judge is faced with weighing the consequences of using what may be untrustworthy information to make a decision that will impact the defendant's personal freedom for a long period of time. The consequence of "getting it wrong" in this situation is substantial.

These two scenarios illustrate the point that must be emphasized. The greater the risk of adverse consequences (and the greater the magnitude of those consequences) in relying on AI evidence that is of uncertain validity and reliability, the greater the need for the trial judge to carefully consider whether to admit the AI evidence for the purpose for which it was offered. This is where Fed. R. Evid. 403 comes into play. The AI evidence may be relevant, and it may be valid and reliable for a purpose other than that which it is being offered to prove, but if the risk of unacceptable consequences to the defendant substantially outweighs its probative value, it should be excluded.

6. Timing Issues

It should be clear at this point that determining whether AI evidence should be admitted in a trial is complicated, requires a great deal of information, and is not the type of issue that is well suited to being resolved in the middle of a trial, or on the fly. Preparation is critical, by both the proponent and opponent of AI evidence. And the judge needs time to hear the competing evidence, to carefully review the supporting materials, and to decide. But since there is no rule of evidence that specifically addresses AI evidence, nor do the Federal Rules of Civil and Criminal Procedure directly require the disclosure of AI evidence, there is a risk that it may not be disclosed soon enough for disputes about its admissibility to be determined before trial. It is true that a party that intends to call a witness who would meet the definition of an expert witness under Fed. R. Evid. 702, in order to lay the foundation for AI evidence, would have to disclose the witnesses' opinions and the basis therefore, which should give its adversary and the court some advanced notice that AI evidence is going to be introduced.[385] But expert disclosures often are more general about the subjects of the expert's intended testimony than the rules actually require, so that the intent to introduce AI evidence may not be clearly flagged far enough ahead of trial.

[385] *See* FED. R. CIV. P. 26(b)(4); FED. R. CRIM. P. 16(a)(1)(G).

That means that the parties should communicate well ahead of trial to determine if AI evidence is going to be offered at trial, and reach agreement (or bring the matter to the attention of the court) about when such AI evidence will be disclosed, the extent to which the party against whom the AI evidence will be admitted will have access to the information needed to assess and challenge its validity and reliability, and whether the proponent of the AI evidence will assert proprietary information or trade-secret protection to deny the production of such information to the opposing party. And the trial judge should inquire during the pretrial stages of the case whether AI evidence will be introduced, set a deadline for its production, as well as for challenges to its admissibility, rule on any trade-secret claims, and schedule a hearing well before trial to insure that the court itself is adequately informed and has sufficient time to make a principled decision as far in advance of trial as possible. Finally, a trial judge faced with ruling on the admissibility of AI evidence need not rely solely on the arguments of the attorneys for the parties and their experts but can appoint a court expert as allowed by Fed. R. Evid. 706[386], if the circumstances so warrant.

CONCLUSION

Although the explosion in the use of AI within increasingly large sectors of our society is of relatively recent vintage, it is here to stay. AI is in a state of such rapid advancement that the law of evidence governing the circumstances under which AI technology and its output should be admitted into evidence in civil and criminal trials is not well developed. A growing number of commentators have written about the potential problems and concerns that impact whether AI evidence should be admitted, but there are few court decisions that have squarely addressed the admissibility of AI evidence in proceedings governed by the Federal Rules of Evidence or their state-law equivalents. But this will change, in due course, as it is inevitable that AI technology will be at the heart of disputes that will increasingly find their way into court. When this happens, lawyers and judges must be prepared to address the evidentiary issues that influence whether the AI evidence is to be admitted. Since AI systems are complex and highly technical, most lawyers and judges will be ill equipped for this task unless they have at least a rudimentary understanding of what AI is, how it operates, scientific and statistical evaluation, and the issues that need to be addressed in order to make decisions about its validity and reliability, and hence its admissibility. And, since there are, at present, no rules in the Federal Rules of Evidence that directly address AI evidence, lawyers and judges must rely

386 *See* FED. R. EVID. 706.

on the rules that do exist to provide an analytical framework to assist them with the challenges that await them when they must confront these issues. Our aim has been to lend a helping hand in this process. We have tried to describe in language that is not overly technical what AI is, the types of AI that presently exist, some of the challenges AI can pose, the principles that govern whether an AI system produces valid and reliable output, the issues that need to be considered when determining its evidentiary value in trials, and the available rules of evidence which—while not perfect for the task—are sufficient to insure fair outcomes, if followed. It is our hope that this article will be useful to lawyers and judge alike and will help to promote fair outcomes in trials in which AI evidence is sought to be admitted. At minimum, we hope that we have shown that when it comes to determining whether AI evidence should be admitted into evidence in civil and criminal trials, lawyers and judges cannot evaluate AI from a state of fundamental ignorance. In time, the court decisions will come, and there may even be new rules of evidence that give more specific guidance. But, in the meantime, we hope that this article will serve as a starting place.