December, 2017

The Opinion of Machines

Curtis E.A. Karnow

THE COLUMBIA
SCIENCE & TECHNOLOGY
LAW REVIEW

VOL. XIX STLR.ORG FALL 2017

ARTICLE

THE OPINION OF MACHINES[†]

Curtis E.A. Karnow[*]

[*] Judge of the California Superior Court, County of San Francisco.

I. INTRODUCTION

> People understand the linear algebra behind deep learning [neural networks]. But the models it produces are less human-readable. They're machine-readable. They can retrieve very accurate results, but we can't always explain, on an individual basis, what led them to those accurate results.[1]

> When I watch these games, I can't tell you how tense it is. I really don't know what is going to happen.[2]

A specific software architecture, neural networks, not only takes advantage of the virtually perfect recollection and much faster processing speeds of any software, but also teaches itself and attains skills no human could directly program. We rely on these neural networks for medical diagnoses, financial decisions, weather forecasting, and many other crucial real-world tasks. In 2016, a program named AlphaGo beat the top-rated human player of the game of Go.[3] Only a few years ago, this had been considered impossible.[4] High-level Go requires remarkable skills, not just of

---

1. Cade Metz, *AI Is Transforming Google Search. The Rest of the Web Is Next*, WIRED (Feb. 4, 2016) (quoting Chris Nicholson), https://www.wired.com/2016/02/ai-is-changing-the-technology-behind-google-searches.

2. Cade Metz, *What the AI Behind AlphaGo Can Teach Us About Being Human*, WIRED (May 19, 2016) (quoting David Silver, one of AlphaGo's creators), https://www.wired.com/2016/05/google-alpha-go-ai/. *See also* Nick Sibicky, *Nick Sibicky Go Lecture #256 - Alpha vs. Go*, YOUTUBE (June 29, 2017), https://www.youtube.com/watch?v=yfUzW0gH8ts (discussing the games AlphaGo plays against itself and includes a quote from Nick Sibicky, a strong Go player and professional Go instructor: "There are a lot of things I don't understand.").

3. David Silver et al., *Mastering the Game of Go with Deep Neural Networks and Tree Search*, 529 NATURE 484, 488 (2016), http://web.iitd.ac.in/~sumeet/Silver16.pdf.

4. *See, e.g.*, Alan Levinovitz, *The Mystery of Go, the Ancient Game That Computers Still Can't Win*, WIRED (May 12, 2014), https://www.wired.com/2014/05/the-world-of-computer-go/ ("But the fact is that of all the world's deterministic perfect information games – tic-tac-toe, chess, checkers, Othello, xiangqi, shogi – Go is the only one in which computers don't stand a chance against humans."). *See also* George Johnson, *To Test a Powerful Computer, Play an Ancient Game*, N.Y. TIMES (July 29, 1997), http://www.nytimes.com/1997/07/29/science/to-test-a-powerful-computer-play-an-ancient-game.html ("'It may be a hundred years before a computer beats humans at Go–maybe

calculation, at which computers obviously excel, but, more critically, of judgment, intuition, pattern recognition, and the weighing of ineffable considerations such as positional balance.[5] These skills cannot be directly programmed. Instead, AlphaGo's neural network[6] trained itself with many thousands and, later, millions of games—far more than any individual human could ever play[7]—and now routinely beats all human challengers.[8] Because it learns and concomitantly modifies itself in response to experience, such a network is termed *adaptive*.[9]

As detailed below, neural networks are used throughout industry and science. They are proposed for missile launch and

---

even longer,' said Dr. Piet Hut, an astrophysicist at the Institute for Advanced Study in Princeton, N.J.").

5. This is so because the number of possible permutations is practically infinite. The number of possible Go games far, far exceeds the number of atoms in the universe, and mere calculation cannot beat even a modestly good human player. *Number of Possible Go Games*, SENSEI'S LIBR. (Mar. 24, 2016), http://senseis.xmp.net/?NumberOfPossibleGoGames. This is as opposed to chess, which has far fewer options than Go. For chess, a so-called brute force approach can beat top human players. *Frequently Asked Questions: Deep Blue*, IBM, https://www.research.ibm.com/deepblue/meet/html/d.3.3a.shtml (last visited Oct. 28, 2017). *See* Cade Metz, *In a Huge Breakthrough, Google's AI Beats a Top Player at the Game of Go*, WIRED (Jan. 27, 2016), https://www.wired.com/2016/01/in-a-huge-breakthrough-googles-ai-beats-a-top-player-at-the-game-of-go/ ("When Deep Blue topped world chess champion Gary Kasparov in 1997, it did so with what's called brute force. In essence, IBM's supercomputer analyzed the outcome of every possible move, looking further ahead than any human possibly could. That's simply not possible with Go."). *See also* Johnson, *supra* note 4 (providing a good explanation of the differing complexities as between Go and chess).

6. See Christopher Burger, *Google DeepMind's AlphaGo: How It Works*, TASTEHIT (Mar. 16, 2016), https://www.tastehit.com/blog/google-deepmind-alphago-how-it-works/, for a general discussion of AlphaGo's neural network. Neural networks are so called because they operate in layers, each with different function. *See also* IAN GOODFELLOW ET AL., DEEP LEARNING 6 (MIT Press 2016) (Draft Version), http://www.deeplearningbook.org/version-2016-03-11/index.html.

7. *See generally* Metz, *supra* note 2.

8. *See AlphaGo Confirmed as Master/Magister*, AM. GO ASS'N. (Jan. 4, 2017), http://www.usgo.org/news/2017/01/alphago-confirmed-as-mastermagister (reporting that on January 4, 2017, AlphaGo was confirmed as the secret player defeating fifty of the top Go players in the world).

9. Mohamad Hassoun, *What Is a Neural Network and How Does Its Operation Differ from That of a Digital Computer? (In Other Words, Is the Brain like a Computer?)*, SCI. AM. (May 14, 2017) https://www.scientificamerican.com/article/experts-neural-networks-like-brain.

interception.[10] This Article argues that as these systems are deemed reliable, juries should be entitled to rely on their expert opinions as well.

The admission of what we might call "machine opinion evidence" entails both a review of the requirements of providing an expert opinion as well as a survey of trial judges who understand the technology and so are able to rule on admissibility and ensure the opinion is correctly framed for the jury. Judges must have enough knowledge to handle the technical issues. Furthermore, appreciating the risks involved, judges must also have the legal authority to decide whether the software is scientifically reliable. Many judges do not have this knowledge, and current law[11] may not tolerate that sort of admissibility analysis. This Article may assist on those two problems, by providing both a detailed outline of the mechanism of neural networks as well as a brief, if moderately technical, background useful to an evaluation of the reliability of machine opinion.

Judges and lawyers alike are familiar with the ability of experts to sway juries with their professed independence and apparent authority. Thus, there is a high risk that juries will view computer systems with even greater authority, as such systems are ostensibly free of bias, independent of the parties, and error-free.[12] Especially in this context, trial judges must carefully undertake their gate-keeping functions[13] and ensure that only reliable evidence gets to the jury.

---

10. *E.g.*, Jinke Xiao et al., *Improved Clonal Selection Algorithm Optimizing Neural Network for Solving Terminal Anti-Missile Collaborative Intercepting Assistant Decision-Making Model*, 644 COMM. COMPUT. & INFO. SCI. 216, 216-31 (2016); Michael B. McFarland & Anthony J. Calise, *Adaptive Nonlinear Control of Agile Antiair Missiles Using Neural Networks*, 8 IEEE TRANSACTIONS ON CONTROL SYSTEMS TECH. 749, 749-56 (2000), http://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=865848; Eric Wahl & Kamran Turkoglu, *Non-Linear Receding Horizon Control Based Real-Time Guidance and Control Methodologies for Launch Vehicles*, 2016 IEEE AEROSPACE CONF. (2016), http://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=7500857.

11. This Article focuses on California law, and uses it as a reasonable example of the state of the law applicable more generally in United States jurisdictions.

12. Erin E. Kenneally, *Gatekeeping Out of the Box: Open Source Software as a Mechanism to Assess Reliability for Digital Evidence*, 6 VA. J. L. & TECH. 13, 39 (2001) ("[D]igital evidence may carry an aura of infallibility in the public's eyes . . . .").

13. *E.g.*, Daubert v. Merrell Dow Pharm., Inc., 509 U.S. 579, 600-01 (1993) ("Rule 702 confides to the judges some gatekeeping responsibility . . . ."); Sargon Enters., Inc. v. Univ. of S. Cal., 288 P.3d 1237, 1250 (Cal. 2012) ("[T]rial courts have a substantial 'gatekeeping' responsibility."). To be clear, this Article is

While there are some court opinions involving neural networks, such as in patent cases,[14] there appears to be no state or federal case discussing the admissibility of what one might term "machine opinion," that is, an evidentiary statement generated by software, which no human can fully explain. A number of commentators, however, have suggested that such evidence should be admissible. They have explored, for example, facial recognition software, which reports the probability that a fuzzy picture is that of a defendant in circumstances in which no human could make a comparable estimate.[15] Commentators have advocated for the use of software to prove fraud in the healthcare industry, which would require pattern detections in large amounts of data.[16] Relatedly, a California Supreme Court Justice has explored the implications of relying on software to generate decisions for administrative agencies and questioned the sort of review courts might give to those decisions.[17] Further examples are provided below.

With the two goals of providing (a) an introduction to the technology of neural networks and (b) an argument for the admissibility of machine opinion, this Article introduces the technology by first looking at the relatively familiar operation of technology-assisted review (TAR) of documents in a typical case. The Article then outlines the extensive application of neural

---

addressed specifically to the threshold issue of the *admissibility* of opinions. While reliability of an opinion is or should be the most important factor in tests for both admissibility (*e.g.*, Wendell v. GlaxoSmithKline LLC, 858 F.3d 1227 (9th Cir. 2017)) and subsequent acceptance by the trier of fact (the judge or the jury), admissibility is distinct from whether the opinion is ultimately treated as persuasive by the trier of fact.

14. *E.g.*, Neuromedical Sys., Inc. v. Neopath, Inc., No. 96 Civ. 5245 (JFK), 1998 WL 264845, at *4 (S.D.N.Y. May 26, 1998).

15. *E.g.*, John Nawara, *Machine Learning: Face Recognition Technology Evidence in Criminal Trials*, 49 U. LOUISVILLE L. REV. 601 (2011). There are interesting Confrontation Clause issues. For example, see Joseph Clarke Celentino, Note, *Face-to-Face with Facial Recognition Evidence: Admissibility Under the Post-Crawford Confrontation Clause*, 114 MICH. L. REV. 1317 (2016).

16. Neil Issar, *More Data Mining for Medical Misrepresentation? Admissibility of Statistical Proof Derived from Predictive Methods of Detecting Medical Reimbursement Fraud*, 42 N. KY. L. REV. 341 (2015). For other suggestions, see Andrea Roth, *Machine Testimony*, 126 YALE L.J. 1972, 2021 (2017).

17. Mariano-Florentino Cuéllar, *Cyberdelegation and the Administrative State*, STANF. PUB. LAW (2016) (Working Paper), https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2754385. *See also* Mariano-Florentino Cuéllar, *Artificial Intelligence and the Administrative State*, PPR NEWS (Dec. 19, 2016), https://www.theregreview.org/2016/12/19/artificial-intelligence-and-the-administrative-state/.

networks in the real world, which is used later to argue that systems trusted in the field should be trusted in court. The Article then turns to the law of evidence, focusing on the rules governing the admissibility of computer-stored and computer-generated data, including animations and simulations. The theme of those sections is, again, that reliability drives admissibility. This sets the stage for the Article's central contention, made through four arguments, that the output of neural networks be admissible in court. The Article ends by invoking the need for meaningful cross-examination, setting out the risks—and so the likely targets of that cross-examination—which attend the admission of opinions generated by neural networks.

I. NEURAL NETWORKS

A. An Introduction for Lawyers: Predictive Coding

Many lawyers are already familiar with the basic technology since they use neural networks in their technology-assisted review (TAR) of voluminous electronic documents.[18] With productions of millions of emails and other documents, it is not only futile to have humans review these pages but also usually cheaper and almost always more accurate when TAR searches for relevant items. The software uses predictive coding. The program is trained on a preliminary or starter set ("seed set") of documents, selected by humans as representative of the universe of documents at issue. Then, the system is provided a sample of the general production documents. The system then offers an opinion regarding what is relevant and what is not. Humans train the system by noting errors, and the system then iteratively refines its ability to discriminate. It does this by weighing various aspects of the documents, such as keywords and series of words, to generate a probability that the item is relevant. When the system is sufficiently accurate with respect to its training (or "control set") documents, it is then applied to the entire corpus of the production—the many millions of documents at issue—and marks those which it determines are

18. *See*, *e.g.*, Shannon Brown, *Peeking Inside the Black Box: A Preliminary Survey of Technology Assisted Review (TAR) and Predictive Coding Algorithms for Ediscovery*, 21 SUFFOLK J. TRIAL & APP. ADVOC. 221 (2016); Aaron T. Goodman, *Predictive Coding and Electronically Stored Information: Computer Analytics Combat Data Overload*, ARIZ. ATT'Y 26 (July/Aug. 2016).

relevant. "Predictive discovery is faster, cheaper and more accurate than traditional discovery approaches."[19]

Several observations can be made about TAR's predictive discovery system. No one knows why the system selects a document: once the system is trained, no script can be provided to a human sorter to imitate the system's selection of documents. That is, there is no way to accurately summarize the criteria used. Nevertheless, parties rely on predictive coding in very high-stakes litigation. It is treated as reliable.

B. Under the Hood: Hidden Layers

Having noted the legal profession's general familiarity with and reliance on a sort of neural network, this section provides a short introduction to a typical mechanism of these programs.

At the risk of conflating uses of the term "expert," neural networks can be contrasted with a classic "expert system." The classic expert system is simply a collection of rules, expressly preprogrammed by a human. For example, imagine a car repair expert system that asks a series of scripted questions and then spits out an answer. A human expert scripted each of the questions and created the matrix such that a certain set of responses generate a scripted output.[20] The operations are "hard-coded" into the

---

19. Joseph H. Looby, *E-Discovery – Taking Predictive Coding Out of the Black Box*, FTI J. (Nov. 2012), http://ftijournal.com/article/taking-predictive-coding-out-of-the-black-box-deleted, (relying on Maura R. Grossman & Gordon V. Cormack, *Technology-Assisted Review in E-Discovery Can Be More Effective and More Efficient than Exhaustive Manual Review*, 17 RICH. J. L. & TECH. 11 (2011)). *See also* VERITAS TECHS. CORP., PREDICTIVE CODING DEFENSIBILITY 3 (2015), https://www.veritas.com/content/dam/Veritas/docs/white-papers/21290290_GA_ENT_WP-Predictive-Coding-Defensibility-Measuring-Accuracy-with-Random-Sampling-EN.pdf ("Despite the widespread misconception that linear review is the electronic discovery process 'gold standard,' exhaustive manual review is surprisingly inaccurate, considering its high cost. Academic research on legal review as part of the TREC Legal Track has shown linear review is often only 40-60 percent accurate. Predictive coding technology involves an iterative process that senior attorneys follow to train software on review criteria, creating a mathematical model that predictive coding software uses to generate 'predictions' of how the remaining documents would otherwise be tagged if reviewed by an experienced attorney. Studies show that predictive coding can achieve much higher levels of accuracy at a fraction of the time and cost.").

20. FREDERICK P. BROOKS, JR., THE MYTHICAL MAN-MONTH 191 (2d ed. 1995). Brooks in his classic text (originally published in 1975) calls these now relatively simple expert systems "inference engines." While the same term could be used for neural networks, the means of inference and their flexibility differ profoundly.

software.[21] Some legal work can probably be done with these systems.[22] The significant point is that humans understand these classic expert systems and can explain each step they perform.

Before delving into systems able to learn (such as the TAR system discussed above), it is important to note that these systems learn, of course, from new data. But that normally requires structured data, which means humans must, in effect, interpret the data from the world, rendering it into formats acceptable to the program.[23]

Representational learning systems, and deep learning systems in particular, do not require this human intervention. These programs can be exposed to data from the real world and be taught—and later, teach themselves—the relationship between (i) raw data and (ii) higher-level representations and abstract concepts.[24] Neural networks are a type of representational learning system; some of them are deep, and some are shallow, as described below. They solve problems that cannot be solved by fixed programs written by humans.[25]

Neural networks are arranged such that humans do not perceive the actual operation: the weighing of probabilities. Humans do not fix the way in which elements are weighed, and they usually do not even identify *which* elements are weighed. The networks organize themselves. Recent results are even more surprising: networks have trained themselves on unlabeled data to recognize, for example, faces and cats—that is, the systems make these discriminations without first being fed examples of the items to be discriminated.[26] These systems use statistics and algorithms derived from probability theory[27] to navigate uncertain and ambiguous data to generate results, and then teach themselves to revise their own algorithms in order to increase accuracy.

---

21. GOODFELLOW ET AL., *supra* note 6, at 2.

22. *See generally*, L. Thorne McCarty, *Reflections on Taxman: An Experiment in Artificial Intelligence and Legal Reasoning*, 90 HARV. L. REV. 837 (1977).

23. GOODFELLOW ET AL., *supra* note 6, at 2-3.

24. *Id.* at 4-5.

25. *Id.* at 96.

26. Quoc V. Le et al., *Building High-Level Features Using Large Scale Unsupervised Learning*, 2012 PROC. 29TH INT'L CONF. ON MACHINE LEARNING 127 (2012), https://static.googleusercontent.com/media/research.google.com/en//archive/unsupervised_icml2012.pdf.

27. GOODFELLOW ET AL., *supra* note 6, at 52-79.

A good example of a neural network is one used for image analysis, such as recognizing faces or other features in pictures.[28] The system first accepts input. In the previous example, further imagine this is a series of pixels that, for simplicity's sake, will be either black or white, on or off, on a grid, perhaps 200 by 200 (*i.e.*, 4,000) pixels or dots. These are processed by a series of computing routines, each one in effect a processor or "node." The system's first task is to recognize whether the inputs are on or off–let's call that the work of the first layer of nodes. The second task is to determine whether there are edges. Three black dots in a row might be an edge; perhaps seven are very likely to be an edge, and ten in a row are extremely likely to be so. Edge detection might, for example, be then the second layer of processing. Depending on how nodes are adjusted, some of the nodes might "vote" that there is an edge or not an edge. At this stage, the system does not know if it is looking at a face or a baseball.

The second layer's output–"There is an edge here" or "There is no edge here"–is the input for the next layer, which might be called a shape detector, or eye detector, for example. At this third layer, the edges are determined to either fit together in a certain shape, or not. The output here might be something like, "There is an eye" or a nose, or some other elemental shape. That output is the input of the next (fourth) layer, which could be a face recognition layer. Given the input of eyes and noses or other shapes, it generates a final output: "We have a face" or "We do not have a face here" or, if the penultimate layer were trained to look for things like wheels, side panels, cabs, and so on, it might report "It's a truck." At each layer, the input is likely to vary greatly because edges come in all sorts of shapes and sizes, and can sometimes manifest either in a few pixels or many more. These edges, at subsequent layers, to a greater or lesser extent conform to an eye, or nose, or wheel, or head, and so on, and those elements in turn conform to a truck or baseball to greater or lesser extent. The output of one layer to the next layer is a probabilistic value. Depending on the system's training, it might take only a weak probability to send an affirmative vote up the chain so to speak, or it might take a high degree of certainty to send that "Yes, it's an edge" or "Yes, this is wheel." A layer may have some but not all of the input it needs to be certain of a conclusion, and so, in effect, its nodes vote on the *degree* of certainty about its conclusion. The nodes in the network that, in the end, either do or do not send on

28. *See generally id.* at 6, fig. 1.2.

a "yes" to the next layer are adjustable—and here is where the training comes in.

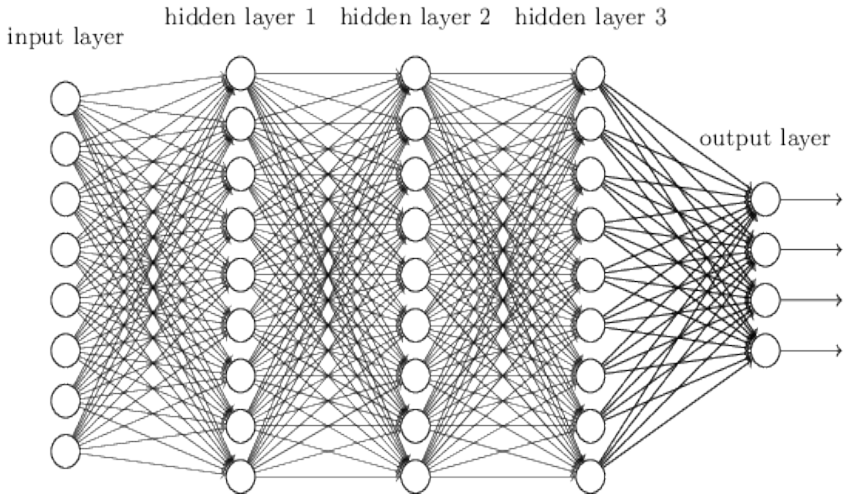
During training, the system makes adjustments to the nodes, assigning more or less weight to inputs from earlier layers. In the classic training session, the system is fed a large number of labeled pictures (or, in the TAR context, documents), and is provided human feedback. It is told if it reached the correct decision. If not, the system experiments internally, adjusting the weighting of its nodes until it maximizes the number of correct estimates or final outputs. The classic example is a "back propagational neural network" in which the final output error is used to go "back" and tweak the nodes' weights, run another effort, and note the extent to which the output improves. Whether technically correct or not, the comparisons to human learning are obvious:[29] children are taught that various things are dogs or cats by repeatedly correcting the child's output statements ("Doggie!" or "Kitty!") until, by and large, the output is correct. As with neural nets, humans can measure, and ultimately have some faith in, the accuracy of the output, but will have no idea what the internal state of the network (or of the child's brain) looks like or exactly why it is so. In a neural network, the internal state is just a very large number of weights, *i.e.*, numbers. The layers in between the initial input and the final output are thus often referred to as "hidden layers."[30] As the system traverses the layers from the raw data input to the final

29. Our intuition that artificial neural networks mimic our biological ones may be right. David Hubel and Torstein Wiesel were awarded the 1981 Nobel Prize in physiology or medicine for work on the information processing systems in the visual cortex, which use the equivalent of hidden layers of neural network. *Press Release: The Nobel Prize in Physiology or Medicine 1981*, THE NOBEL ASSEMBLY OF KAROLINSKA INSTITUTE (Oct. 9, 1981), https://www.nobel prize.org/nobel_prizes/medicine/laureates/1981/press.html. Others caution that the brain is like a neural network only by way of analogy and metaphor. *E.g.*, Chris Chatham, *10 Important Differences Between Brains and Computers*, SCIENCEBLOGS (Mar. 27, 2007), http://scienceblogs.com/developing intelligence/2007/03/27/why-the-brain-is-not-like-a-co/. The issue is irrelevant here. For those interested, artificial neural networks probably will not have the same number of neuron equivalents as humans until around 2050, GOODFELLOW ET AL., *supra* note 6, at 21; but around then artificial networks may advance very, very rapidly, unconstrained by the relatively slow processing speeds and limited storage abilities of humans.

30. Currently, networks with about ten layers are termed "deep" or "very deep." *See* Jürgen Schmidhuber, *Deep Learning in Neural Networks: An Overview*, 61 NEURAL NETWORKS 85, 88 (2015), http://www.sciencedirect.com/ science/article/pii/S0893608014002135.

output, it reaches conclusions about increasingly complex and abstract concepts.[31]

Here is a simple diagram of a five-layer network:[32]



Supervised networks train using labeled data and then estimate answers from new input. As noted, neural networks can train themselves, taking great advantage of the amount of digitized data which has vastly increased in recent years.[33] "Big data" allows programs much more room to train and self-correct their mechanisms. While the line between supervised and unsupervised learning is not fixed,[34] unsupervised learning examines unlabeled data, compares it to random data, and extracts a series of features common to the non-random data. These features are, of course, abstractions from the input layer. The parameters of that layer may then be fixed and its output examined by the next layer as input, extracting common features for the next level of abstraction. A simple example is a clustering program, which reviews a large amount of input, makes conclusions concerning common features, and then sorts the inputs into different groups. This can all be done with unlabeled data, and human corrective input is not required. For example, engineers at Google created an early iteration of AlphaGo, which taught itself to recognize cats. Without telling it

31. GOODFELLOW ET AL., *supra* note 6, at 8.

32. MICHAEL NIELSEN, *Chapter 5: Why Are Deep Neural Networks Hard to Train?*, *in* NEURAL NETWORKS AND DEEP LEARNING (2017), http://neuralnetworksanddeeplearning.com/chap5.html.

33. GOODFELLOW ET AL., *supra* note 6, at 19-20. Discussed below in Part IV.D.

34. *Id.* at 100.

anything about cats, the engineers simply let it examine 13,026 pictures of cats and 23,974 pictures without cats. Even though the engineers provided no indication as to which was which, the system eventually detected the common cat features on its own, and reported its discovery of that common entity.[35]

By sorting data into groups or clusters defined by common aspects, unsupervised learning systems thus create a series of higher-level abstractions. Then, these systems learn to improve themselves. Assume a system which distinguishes—*i.e.*, separately clusters, as described above—digits from not digits; or cats from not cats. Now the higher layers (those generating conclusions such as "This is a digit" or "This is a cat") perform a top-down analysis, instructing the lower layers on what, more specifically, to look for as they make their lower-level determinations. For example, a top-down pass might in effect say, "Look for about two to three strokes for digits" or "Look for whiskers and a certain shape of ear for cats," which then iteratively improves the performance of the system as a whole.[36] The system teaches itself.

C. Uses of Neural Networks

For reasons expressed below, it is important to note the extensive reliance on neural networks. Traditional means of analysis are ill-equipped to handle massive amounts of data—familiarly known as "big data." But neural networks can be used to extract patterns and find needles in these big data haystacks. For example, these networks are used for automated bank loan application approval, credit card fraud detection, as well as a wide spectrum of other uses in the financial markets. They are used for medical diagnoses and x-ray interpretation. They are also used for process controls in factories, in scientific research, and of course, in data mining in many contexts.[37] One text notes these uses:

---

35. *See* Le et al., *supra* note 26, at 1 ("Contrary to what appears to be a widely-held intuition, our experimental results reveal that it is possible to train a face detector without having to label images as containing a face or not . . . . We also find that the same network is sensitive to other high-level concepts such as cat faces and human bodies.").

36. For a more technical but still somewhat approachable discussion, see Geoffrey E. Hinton, *Learning Multiple Layers of Representation*, 11 TRENDS COGNITIVE SCI. 428 (2007), http://www.cs.toronto.edu/~hinton/absps/tics.pdf.

37. A quick Google Scholar review demonstrates the breadth of splendid scientific research covering this topic. Further applications include: Financial: Stock Market Prediction; Credit Worthiness; Credit Rating; Bankruptcy Prediction; Property Appraisal; Fraud Detection; Price Forecasts; Economic Indicator Forecasts; Medical: Medical Diagnosis; Detection and Evaluation of

Detection of medical phenomena. A variety of health-related indices (*e.g.*, a combination of heart rate, levels of various substances in the blood, respiration rate) can be monitored. The onset of a particular medical condition could be associated with a very complex (*e.g.*, nonlinear and interactive) combination of changes on a subset of the variables being monitored. Neural networks have been used to recognize this predictive pattern so that the appropriate treatment can be prescribed.

Stock market prediction. Fluctuations of stock prices and stock indices are another example of a complex, multidimensional, but in some circumstances at least partially-deterministic phenomenon. Neural networks are being used by many technical analysts to make predictions about stock prices based upon a large number of factors such as past performance of other stocks and various economic indicators.

Credit assignment. A variety of pieces of information are usually known about an applicant for a loan. For instance, the applicant's age, education, occupation, and many other facts may be available. After training a neural

---

Medical Phenomena; Patient's Length of Stay Forecasts; Treatment Cost Estimation; Industrial: Process Control; Quality Control; Temperature and Force Prediction; Science: Pattern Recognition; Recipes and Chemical Formulation Optimization; Chemical Compound Identification; Physical System Modeling; Ecosystem Evaluation; Polymer Identification; Recognizing Genes; Botanical Classification; Signal Processing: Neural Filtering; Biological Systems Analysis; Ground Level Ozone Prognosis Odor Analysis and Identification; Educational: Teaching Neural Networks; Neural Network Research; College Application Screening; Predict Student Performance; Data Mining: Prediction Classification; Change and Deviation Detection; Knowledge Discovery; Response Modeling; Time Series Analysis; Sales and Marketing: Sales Forecasting; Targeted Marketing; Service Usage Forecasting; Retail Margins Forecasting; Operational Analysis: Retail Inventories Optimization; Scheduling Optimization; Managerial Decision Making; Cash Flow Forecasting; HR Management: Employee Selection and Hiring; Employee Retention; Staff Scheduling; Personnel Profiling; Energy: Electrical Load Forecasting; Energy Demand Forecasting; Short and Long-Term Load Estimation; Predicting Gas/Coal Index Prices; Power Control Systems; Hydro Dam Monitoring; Other: Sports Betting; Making Horse and Dog Racing Picks; Quantitative Weather Forecasting; Games Development; Optimization Problems, Routing; Agricultural Production Estimates. *Neural Network Software Applications*, ALYUDA, http://www.alyuda.com/products/neurointelligence/neural-network-applications.htm (last visited Nov. 1, 2017).

network on historical data, neural network analysis can identify the most relevant characteristics and use those to classify applicants as good or bad credit risks.

Monitoring the condition of machinery. Neural networks can be instrumental in cutting costs by bringing additional expertise to scheduling the preventive maintenance of machines. A neural network can be trained to distinguish between the sounds a machine makes when it is running normally ("false alarms") versus when it is on the verge of a problem. After this training period, the expertise of the network can be used to warn a technician of an upcoming breakdown, before it occurs and causes costly unforeseen "downtime."

Engine management. Neural networks have been used to analyze the input of sensors from an engine. The neural network controls the various parameters within which the engine functions, in order to achieve a particular goal, such as minimizing fuel consumption.[38]

Closer to the legal realm, neural nets are developed or proposed for the electronic discovery uses noted above, as well as, for example, detecting gunshot residue,[39] demographic analysis of crime patterns,[40] automated detection of smuggling,[41] and other uses[42] including for legal services.[43]

---

38. STATSOFT, INC., *Neural Networks, in* ELECTRONIC STATISTICS TEXTBOOK (2013), http://www.statsoft.com/Textbook/Neural-Networks (last visited Nov. 9, 2017).

39. Regina Verena Taudte et al., *Development of a UHPLC Method for the Detection of Organic Gunshot Residues Using Artificial Neural Networks*, 7 ANAL. METHODS 7447 (2015), http://pubs.rsc.org/en/content/articlepdf/2015/ay/c5ay00306g.

40. Xingan Li & Martti Juhola, *Country Crime Analysis Using the Self-Organizing Map, with Special Regard to Demographic Factors,* 29 AI & SOC'Y 53 (2014), http://link.springer.com/article/10.1007/s00146-013-0441-7.

41. N. Jaccard et al., *Automated Detection of Smuggled High-Risk Security Threats Using Deep Learning*, ARXIV (Sept. 9, 2016), https://arxiv.org/pdf/1609.02805.pdf.

42. Michael Aikenhead, *The Uses and Abuses of Neural Networks in Law*, 12 SANTA CLARA COMPUT. & HIGH TECH. L.J. 31 (1996) (legal reasoning); Mohammadreza Ebrahimi et al., *Detecting Predatory Conversations in Social Media by Deep Convolutional Neural Networks*, 18 DIGITAL INVESTIGATION 33 (2016), http://www.sciencedirect.com/science/article/pii/S1742287616300731; Neil Issar, *More Data Mining for Medical Misrepresentation? Admissibility of Statistical Proof Derived from Predictive Methods of Detecting Medical*

II. ADMITTING OUTPUT OF SOFTWARE

This section examines basic rules for the admissibility of computer-generated evidence generally, in preparation for a later discussion of the admissibility of machine opinion.

Evidence introduced at trial, including software, may involve issues of hearsay and, more generally, of reliability. Authenticity is an aspect of reliability: thus, a document must be authenticated because otherwise it is not reliable. Hearsay objections are pertinent to some computer outputs but not to others. Software is used to generate simulations and animations—two very different types of evidentiary creatures with different admissibility requirements. As discussed below, the rules governing the admissibility of simulations, in particular, are useful but insufficient in deciding whether the output of neural networks should be admitted.

But first, a brief taxonomy of computer-generated evidence will be useful.[44] As Justice Simons has noted, "We must distinguish from this computer-generated data [such as data associated with credit card swipes and cell phone use], a written or electronic document prepared by a person, and then electronically stored in a computer. Electronic storage does not make the document computer-generated."[45]

A. The Filing Cabinet

Reimbursement Fraud, 42 N. KY. L. REV. 341 (2015) (statistical detection evidence); Georgia Koukiou & Vassilis Anastassopoulos, *Neural Networks for Identifying Drunk Persons Using Thermal Infrared Imagery*, 252 FORENSIC SCI. INT'L 69 (2015), http://www.sciencedirect.com/science/article/pii/S0379073815001681; John Nawara, *Machine Learning: Face Recognition Technology Evidence in Criminal Trials*, 49 U. LOUISVILLE L. REV. 601 (2011) (reliability of face recognition systems); Dominik Olszewski, *Fraud Detection Using Self-Organizing Map Visualizing the User Profiles*, 70 KNOWLEDGE-BASED SYS. 324 (2014), http://www.sciencedirect.com/science/article/pii/S0950705114002652.

43. John O. McGinnis & Russell G. Pearce, *The Great Disruption: How Machine Intelligence Will Transform the Role of Lawyers in the Delivery of Legal Services*, 82 FORDHAM L. REV. 3041 (2014).

44. *See generally*, Gregory P. Joseph, *A Simplified Approach to Computer-Generated Evidence and Animations*, 43 N.Y.L. SCH. L. REV. 875 (1999).

45. MARK SIMONS, CAL. EVID. MANUAL § 2:2 (2017).

Much of what is thought of as computer-generated evidence (CGE)[46] is not. What is sometimes termed CGE is actually generated by humans who input the data into computers; the computers act simply as storage systems much like filing cabinets. Letters, briefs, emails, PowerPoints, much of our spreadsheets and other accounting data, and most photographs fit in this category. Aside from photographs, these types of data are collections of statements by humans, and so a hearsay objection may be made.[47] The objection can be met with a showing under the business records exception, for example.[48] Websites and chat room postings are also similarly housed in digital storage cabinets: people put the words there and other people can testify as to authenticity and related issues, as they would if the data came out of a physical filing cabinet. Databases contain human-entered data too, which fit under this rubric of human-generated information. Pictures posted to the Internet are generally authenticated and admitted the same way as any other photographs; that is, either someone testifies that she took the picture or someone familiar with the scene depicted testifies that the photo is accurate. Circumstantial evidence may suffice for admissibility.[49]

It is true that the act of processing electronic data into photographs or legible text involves computer processing, a translation of bits into a human-readable product. But the proponent of the evidence need not explain or defend this type of

---

46. Justice Simons calls it computer-generated information (CGI). "CGI" is also used to designate computer-enhanced film effects (computer-generated imagery). *Id.* In the interests of avoiding ambiguity, "CGE" is used here for computer-generated evidence.

47. SIMONS, *supra* note 45, at § 2:63; People v. Romeo, 193 Cal. Rptr. 3d 96, 108 (Cal. Ct. App.2015) ("[I]nformation residing in a computer database is still hearsay, often multilevel hearsay."); Joseph, *supra* note 44, at 878.

48. Cal. Evid. Code § 1271 (West 2017).

49. *See* EDWARD A. RUCKER & MARK E. OVERLAND, 4 CAL. CRIM. PRACTICE: MOTIONS, JURY INSTRUCTIONS AND SENTENCING § 48:18 (4th ed. 2017) ("The normal rules of admissibility apply to evidence obtained from social networking and other online sites. Authentication of a photograph on a Web site may be provided by expert testimony if there is no one qualified to authenticate it from personal observation. In addition, authentication may be provided from its contents or subject matter (People v. Valdez, 135 Cal. Rptr. 3d 628 (Cal. Ct. App. 2011) (photograph from a social-networking web page alleged to have been authored by defendant was sufficiently authenticated by its content to be admissible))."). *See also*, Steven Goode, *The Admissibility of Electronic Evidence*, 29 REV. LITIG. 1, 24–25 (2009) (discussing the use of circumstantial evidence to authenticate in federal court); Paul W. Grimm et. al., *Authenticating Digital Evidence*, 69 BAYLOR L. REV. 1, 15 (2017) (showing that circumstantial evidence is widely used to authenticate).

processing as it is assumed that "a computer's print function has worked properly."[50] In short, "[p]rintouts are admissible and presumed to be an accurate representation of the data in the computer."[51] That being said, even when the print function presumably works correctly, the printed data may still be hearsay, because it was input by humans.[52]

B. Data Created by Internal Operations

Computers may be fed data and, based on their programming, generate new data, the accuracy of which depends on the validity of the programing. The simplest examples are spreadsheet cells containing formulas that simply execute something like a mini program written by the user, such as "Multiply cell B3 with B4 and put the result here." If the human selected the wrong cells, or directed a multiplication when it ought to have been division, the result (*i.e.*, "Profits this year were $100") will be wrong. Drawing programs can automatically generate circles and squares, but whether these are accurate depends on whether the algorithm is correct. In short, software may or may not have bugs,[53] which may affect the validity of its outputs.

---

50. People v. Goldsmith, 326 P.3d 239, 246 (Cal. 2014), *quoting* People v. Hawkins, 121 Cal. Rptr. 2d 627, 643 (Cal. Ct. App. 2002). Of course, as with any other presumption, the other side is free to attack it. But the presumption is almost always enough to get the evidence before the trier of fact (*e.g.*, the jury). *See*, *e.g.*, People v. Martinez, 990 P.2d 563, 580-82 (Cal. 2000) (problems with printouts may be subject to cross-examination but typically will not bar admissibility). This is because the opponent does not have evidence that the printout is inaccurate. But if she does have evidence of inaccuracy, *the presumption is no longer in effect*, and the burden returns to the proponent of the printout to establish that it is, in fact, accurate. *See* People v. Rekte, 181 Cal. Rptr. 3d 912, 918-19 (Cal. Ct. App. 2015).

51. BERNARD WITKIN ET AL., 1 CAL. EVID. § 231(b)(3) (5th ed. 2012). New federal rules of evidence, effective December 1, 2017, will make it even easier in federal court to meet basic authentication requirements for computer stored data. *See,* FED. R. EVID. 902(13)-(14). *See also* Grimm et al., *supra* note 49, at 39.

52. Printouts offered for their truth usually have to qualify under some exception to the hearsay rule, such as under the business records exception. *See* Aguimatang v. Cal. State Lottery, 286 Cal. Rptr. 57, 72-73 (Cal. Ct. App. 1991); People v. Lugashi, 252 Cal. Rptr. 434, 439 (Cal. Ct. App. 1988).

53. Actually, all software in general use is sufficiently complex in that it has bugs. "All software contains bugs or errors in the code. Some of these bugs have security implications, granting an attacker unauthorized access to or control of a computer. These vulnerabilities are rampant in the software we all use. A piece of software as large and complex as Microsoft Windows will contain hundreds of them, maybe more." Bruce Schneier, *Why the NSA Makes Us More Vulnerable to Cyberattacks: The Lessons of WannaCry*, FOREIGN AFFAIRS

Typically, it is suggested that because the result of internal computer processing is not a human statement, hearsay is not implicated. For example, the venerable legal author Bernard Witkin declared:[54]

> *Distinction: Computer's Internal Operations.* A printout of the results of a computer's internal operations is not hearsay evidence at all, and thus the business records exception is inapplicable. Such a printout does not represent the output of statements placed in the computer by an out-of-court declarant. With a machine, there is no possibility of conscious misrepresentation. "[T]he true test for admissibility of a printout reflecting a computer's internal operations is not whether the printout was made in the regular course of business, but whether the computer was operating properly at the time of the printout." (People v. Hawkins (2002) 98 C.A.4th 1428, 1449, 1450, 121 C.R.2d 627 [in prosecution arising from defendant's having taken source code from computer system of former employer, trial judge did not err in admitting computer printouts showing when computer files were last accessed, where evidence was introduced showing that computer was functioning properly and its clock was accurate] . . . .).[55]

But the results of internal processing may use human-entered data as inputs, and that data can be challenged on a variety of grounds, including hearsay. Sometimes, data appears to come directly to the computer without human intervention, such as by way of sensors and digital imaging, in which case it may be thought of as part of the internal processing of the computer. There may be some difference between the standards used for the admission of the results of internal processing and those used in connection with sensor input. In the former situation, as discussed below, the proponent must present some foundation (but not much) on the accuracy of the system, but the inputs of real-time

---

(May 30, 2017), https://www.foreignaffairs.com/articles/2017-05-30/why-nsa-makes-us-more-vulnerable-cyberattacks. The complexity of software is essential, not an accident. *See* BROOKS, *supra* note 20, at 183. And that complexity may lead to unexpected results—which is, in fact, all that is meant by "bug."

54. Bernard Witkin, a former Reporter of Decisions for the California Supreme Court, was the author of, among other things, these standard compendia of California law: SUMMARY OF CAL. LAW (11th ed. 2017), CAL. PROC. (5th ed. 2008), and CAL. EVID. (5th ed. 2012).

55. WITKIN ET AL., *supra* note 51, at § 231(b)(3).

sensor information, such as automated photographs, "are presumed to be accurate,"[56] a test reminiscent of that applied to the "print function" of a computer.

Turning then directly to classic internal processing, the basic rule for admissibility is succinctly captured in this unpublished opinion:

> The test for admissibility of machine created information is whether the computer was operating properly at the time of the printout . . . . The admissibility of computer records also does not require establishing the accuracy, maintenance, reliability or the acceptability of the computer's hardware or software. Our Supreme Court has noted that mistakes can occur with computer generated information. However, such mistakes should not affect admissibility but be developed on cross-examination.[57]

This "machine created information" or computer-generated evidence (CGE) includes, for example, metadata, such as timestamps and author-identification information,[58] which is automatically created by the machine. But even some human input is implicated for metadata, such as setting the time either on the

---

56. This is "especially [so for] government-maintained computers, [which] are presumed to be accurate. Thus, a witness with the general knowledge of an automated system may testify to his or her use of the system and that he or she has downloaded the computer information to produce the recording. No elaborate showing of the accuracy of the recorded data is required. Courts in California have not required 'testimony regarding the "'acceptability, accuracy, maintenance, and reliability of . . . computer hardware and software'" in similar situations.'" People v. Dawkins, 179 Cal. Rptr. 3d 101, 110 (Cal. Ct. App. 2014) (internal citations omitted).

57. People v. Johnson, No. F069414, 2016 WL 4482963, at *3 (Cal. Ct. App. Aug. 25, 2016) (unpublished) (internal citations omitted). *See also, e.g.*, People v. Hawkins, 121 Cal. Rptr. 2d 627, 642-43 (Cal. Ct. App. 2002) ("[T]he true test for admissibility of a printout reflecting a computer's internal operations is not whether the printout was made in the regular course of business, but whether the computer was operating properly at the time of the printout."). This foundational showing, that the computer was "operating properly," does not require much. *See People v. Martinez*, 990 P. 2d 563, 581 (Cal. 2000) ("[O]ur courts have refused to require, as a prerequisite to admission of computer records, testimony on the 'acceptability, accuracy, maintenance, and reliability of . . . computer hardware and software.'") (quoting *Lugashi,* 252 Cal. Rptr. at 441). Mistakes can be exposed by cross-examination. *Id. Accord Dawkins*, 179 Cal. Rptr. 3d at 110 (requiring no elaborate foundational showing, especially with government maintained computers); People v. Peyton, 177 Cal. Rptr. 3d 823 (Cal. Ct. App. 2014).

58. United States v. Hamilton, 413 F.3d 1138 (10th Cir. 2005).

machine itself or on some other machine to which it refers, creating the author's initials, and so on. Nevertheless, generally the data is automatically created, and so qualifies as CGE. Thus, no hearsay objection applies.

And while CGE does have to be validated with some foundational testimony, the bar is not high:

> First, the witness through whom the computer records are introduced is qualified if that witness generally understands the system's operation and possesses sufficient skill and knowledge to properly use the system and explain the resultant data, even if the witness is unable to perform every task from initial design and programming to final printout . . . . Second, testimony on the acceptability, accuracy, maintenance, and reliability of computer hardware and software need not be introduced, particularly where the data consists of retrieval of automatic inputs rather than computations based on manual entries . . . .[59]

The requirement generally to explain the system's operation seems to satisfy, or be the functional equivalent of offering "foundational evidence that the computer was operating properly."[60] What does it mean to show that the computer was operating "properly"? It seems to be no more than showing that it was operating as it usually does, explained by someone with some experience in using the system.[61] That is enough to meet the

---

59. SIMONS, *supra* note 45, at § 2:63 (internal citations omitted). *See also* text, *supra* note 53. But for a lengthy list of issues concerning this functionality that conceivably might be up for discussion, see Joseph, *supra* note 44, at 882. This list of issues may be significant as the focus of the (1) opposing side's attack on functionality, hoping to destroy the presumption of reliability (*cf. supra* note 50), or (2) the proponent's efforts thereafter to shoulder the real burden of demonstrating functionality. If the evidence is admitted, the issues could also be used to argue to the fact finder that the evidence is or is not persuasive.

60. *Hawkins*, 121 Cal. Rptr. 2d at 643. *Cf.*, *Goldsmith*, 326 P.3d at 248 (considering, among other factors, showing that "the evidence was properly received in the normal course and manner of Inglewood's operation of its ATES program").

61. *E.g.*, *Johnson*, 2016 WL 4482963, at *3 (admitting scanner evidence based on testimony of how staff used the scanner where the scanner's information "was from the night of the shooting because its data only lasted 'a day or two' and [the witness] correlated the scanner with the video surveillance system"); People v. Johnson, No. B224491, 2011 WL 4436451, at *2 (Cal. Ct. App. Sept. 26, 2011) (unpublished) ("Stoltz testified that the software company who owned the loan software designed a special program to extract 'miscellaneous' type transactions posted during a specific time period. The result

"minimal requirement for admissibility"[62] which, after all, still leaves the evidence subject to cross-examination and argument that the fact finder should disregard it.[63]

C. Simulations

1. Foundation

The low threshold for CGE may simply be a practical response to an otherwise impossible problem. Computers and their data are ubiquitous, but no one really knows how they work in detail.[64] No person, for example, can report on the detailed instructions used by the most ordinary operating system, not to mention the myriad interactions between operating systems and applications found in every business and most homes in this country. But the test is reasonable because it meets a fundamental predicate that notions of reliability in the legal world mirror those used in the "real" or ordinary world. The same line of reasoning dictates that business records are exempt from the hearsay rule: if the hearsay is good

---

of running the special program was a list of 'miscellaneous' type transactions for the period in question. This is sufficient to create an inference that the computer program was working properly."). As observed in *supra* note 51, new federal rules of evidence will ease the admissibility of computer evidence. Rule 902(13) of the new Federal Rules of Evidence *seems* to apply to CGE (while 902(14) applies to computer stored data), but the Committee Notes make it clear that only authenticity is established through the certification procedures of the rule, not reliability as such. For example, the Committee Notes state that "[s]imilarly, a certification authenticating a computer output, such as a spreadsheet, does not preclude an objection that the information produced is unreliable—the authentication establishes only that the output came from the computer." FED. R. EVID. 902(13) (advisory committee's note to 2016 amendment).

62. *Lugashi*, 252 Cal. Rptr. at 440.

63. *See also, e.g.*, People v. Nazary, 120 Cal. Rptr. 3d 143, 163-65 (Cal. Ct. App. 2010), *overruled on other grounds in* People v. Vidana, 377 P. 3d 805, 815-16 (Cal. 2016) (holding test of admissibility of machine-generated receipts from automated gas station island pumps is whether "machine was operating properly at the time of the reading").

64. MARK MEYSENBURG, INTRODUCTION TO PROGRAMMING USING PROCESSING 252 (3d ed. 2016) (regarding the typical fly-by-wire automated aircraft controls systems) ("No single person on the face of the earth truly understands everything there is to know about the software that keeps the airliner flying."). *See generally*, SAMUEL ARBESMAN, OVERCOMPLICATED: TECH. AT THE LIMITS OF COMPREHENSION 3 (2016) (discussing in particular computer enabled systems) ("[T]echnological complexity has eclipsed our ability to comprehend it.").

enough for a business to rely on, it is good enough for a jury.[65] So too here: if an entity relies on the validity of CGE in its day-to-day work, a jury is justified in making the same assumption of validity. Expecting much more would exclude CGE from our courts.

But this reasoning does not quite extend to justify the admission of computer simulations, which are purposely made for a trial—these are bespoke CGE. The techniques used are not employed in the run-of-the-mill business.[66]

Simulations and animations are not the same.[67] Animations are a sort of supporting evidence that serves only to illustrate other testimony, much as a drawing of a car accident scene by an eyewitness serves to illustrate and explain the witness testimony.[68] As demonstrative evidence, it may or may not be admissible,[69] but in any event it entirely depends on the primary testimony, and it is the human witness who is cross-examined.[70] No one really cares how the animation was made, for example, how the drawing program works, or how it calculates distances or other features, just as no one cares how a camera works when a witness testifies that picture is a fair representation of the scene she saw.

Simulations, by contrast, are introduced as primary or "substantive" evidence: they depend on accurate inputs, but their validity also, critically, depends on valid algorithms. Therefore, the

---

65. *E.g.*, United States v. Ary, 518 F.3d 775, 786 (10th Cir. 2008).

66. This is not always true. Some businesses do indeed rely on simulations for their quotidian work. Lapsley v. Xtek, Inc., 689 F.3d 802, 815 (7th Cir. 2012) ("[S]imulation is one of the most common of scientific and engineering tools. Around the world, computers simulate nuclear explosions, quantum mechanical interactions, atmospheric weather patterns, and innumerable other systems that are difficult or impossible to observe directly. A mathematical or computer model is a perfectly acceptable form of test."). These simulations may not be admissible under a lesser level of scrutiny, or that reliability might simply be easier to establish.

67. People v. Duenas, 283 P. 3d 887, 900 (Cal. 2012) ("Courts and commentators draw a distinction between computer animations and computer simulations."). For a detailed discussion, *see, e.g.*, CURTIS KARNOW, LITIGATION IN PRACTICE 31-33 (2017).

68. *E.g.*, People v. Hood, 62 Cal. Rptr. 2d 137, 139-40 (Cal. Ct. App. 1997).

69. That is, judges may not let the animation go to the jury room during deliberations, and it may not become part of the record sent to the court of appeal. But of course, the jury sees it, so in that less technical sense, the animation is admitted.

70. Betsy S. Fiedler, *Are Your Eyes Deceiving You?: The Evidentiary Crisis Regarding the Admissibility of Computer Generated Evidence*, 48 N.Y.L. SCH. L. REV. 295, 299 (2004) ("[T]he testifying witness must state that the CGE portrays the disputed subject matter fairly and accurately.").

validity of the algorithm, unlike the validity of an animation, is fair game for a challenge:

> Courts have compared computer animations to classic forms of demonstrative evidence such as charts or diagrams that illustrate expert testimony . . . . A computer animation is admissible if "it is a fair and accurate representation of the evidence to which it relates . . . ." A computer simulation, by contrast, is admissible only after a preliminary showing that any "new scientific technique" used to develop the simulation has gained "general acceptance . . . in the relevant scientific community."[71]

For this custom CGE, much more than the minimal threshold noted above must be presented. If challenged, the proponent must satisfy the much stronger test of showing that the methods used by the software are justified by science, for example, by a showing:

> [T]hat the facts and data upon which the simulation is based "are of a type reasonably relied upon by experts in the particular field," that the simulation is "the product of reliable principles and methods," and that the supporting expert witness "applied principles and methods reliably" when creating or using the simulation.[72]

As one commentator notes, "in the context of simulations, the computer itself is the expert."[73]

Simulations may, for example, analyze airplane crashes and the movement of groundwater and contaminants. The input consists of data such as records of radar returns, facts concerning the crash site including distances between pieces of the aircraft, so-called "black box" data including speed over time, whether flaps were deployed, and so on. These are fed to a program which reproduces a view of the accident from the pilots' perspective or provide a basis to conclude that a warning must have sounded and been ignored, or that the aircraft was at a certain angle of attack. A simulation of groundwater contamination might take inputs of data, including measurements of a toxin over time in a certain area and

---

71. *Duenas,* 283 P. 3d at 901, *quoting* People v. Kelly, 130 Cal. Rptr. 144, 148 (Cal. Ct. App. 1976) (internal citations omitted).

72. Victoria Webster & Fred E. (Trey) Bourn III, *The Use of Computer-Generated Animations and Simulations at Trial*, 83 DEF. COUNS. J. 439, 441 (2016) (notes omitted) (includes multi-circuit survey).

73. *Id.* at 440.

the movement rate of groundwater over that period, and then opine that the toxin must have been at a certain concentration at a point upstream at a specified earlier time.

In these situations, the validity, and hence the admissibility, of the simulation depends on the validity of the programming including calculations and underlying assumptions. The California Supreme Court has held that simulations are admissible if they are scientifically reliable, and "only after a preliminary showing that 'any new scientific technique' used to develop the simulation has gained general acceptance . . . in the relevant scientific community."[74] More established techniques must still be explained because, as expert systems, they are subject to the usual strictures including basic scientific reliability[75] and non-speculative connections between the conclusions (or output) of the opinion and the input.[76] But it is difficult to know what counts as a sufficient demonstration of reliability.

2. Interlude: Explaining Software

Judges and juries expect that, at some level, the operations of simulations can be explained, that the "heuristic basis" can be demonstrated.[77] For example, an expert might use commonly accepted formulae for the relationship between the pressure of a liquid and aperture of the container through which it is released, based on the classic Bernoulli equation, to compute the speed of the liquid or the pressure it would exert on its target.[78] The proponent of the model must establish this reliability:

---

74. *Duenas,* 283 P. 3d at 901, *quoting* People v. Kelly, 130 Cal. Rptr. at 148.

75. People v. Jackson, 376 P. 3d 528, 568 (Cal. 2016) ("[Expert] procedures and experiments must comply with the laws of physics, chemistry, and biology."). *See, e.g.*, Liquid Dynamics Corp. v. Vaughan Co., 449 F.3d 1209, 1221 (Fed. Cir. 2006) (simulations subject to analysis under classic *Daubert* criteria and deemed in this case to be reliable); Novartis Corp. v. Ben Venue Labs., Inc., 271 F.3d 1043, 1054 (Fed. Cir. 2001) ("[Valid] simulation . . . requires both a solid theoretical foundation and realistic input parameters to yield meaningful results. Without knowing these foundations, a court cannot evaluate whether the simulation is probative."); Lyondell Chem. Co. v. Occidental Chem. Corp., 608 F.3d 284, 294 (5th Cir. 2010) ("[W]e can gauge reliability by examining input values and requiring transparency from testifying experts.").

76. Sargon Enterprises, Inc. v. Univ. of S. Cal., 288 P. 3d 1237, 1252-53 (Cal. 2012).

77. William R. Swartout, *Explaining and Justifying Expert Consulting Programs*, *in* COMPUTER-ASSISTED MEDICAL DECISION MAKING 254-271 (James A. Reggia & Stanley Tuhrim eds., 1985).

78. Lapsley v. Xtek, Inc., 689 F.3d 802, 815 (7th Cir. 2012).

> Computer-generated simulations are based on mathematical models, and particular attention must be paid to the reliability and trustworthiness of the model. A model is a set of operating assumptions—a mathematical representation of a defined set of facts, or system. To be accurate, it must produce results that are identical or very similar to those produced by the physical facts (or system) being modeled. In order to do that, the model must contain all relevant elements—and reflect all relevant interactions—that occur in the real world.[79]

But it is not clear what is involved in any foundational requirement to explain the operation of software, including computer simulations.

There are some issues of definition, such as whether the "program" includes operating systems, interfaces, and commonly available libraries.[80] Those issues can, to some extent, be defined away as not pertaining to the program. For example, judges probably do not want to hear about the general housekeeping functions, such as the runtime environment including operating systems, standard interfaces, and device drivers, or the language in which the program was written in (*e.g.*, C++, FORTRAN, etc.). These are aspects of base technology and ordinarily do not embody the decision-making processes that are at issue as a foundation is laid for a simulation. Juxtaposed with this base technology is the decision-making mechanism of interest to the court, which for convenience may be termed the "inference engine." This is the part of the program that manipulates data and generates conclusions.[81] It is the inference engine that embodies

---

79. Joseph, *supra* note 44, at 65.

80. A library contains "pre-written" code with functions that can be called on by the main executing program.

81. The inference engine's code may, however, be located in a variety of subprograms and libraries. The in-court proponent of the software may or may not be cognizant of the specific mechanisms of each piece of the inference engine because the engine might depend on components such as dynamic link libraries (DLLs) written by others, and the proponent may be wrong about what those DLLs do. *See* Andreas Björklund, et al., *DLL Spoofing in Windows*, UPPSALA UNIV. (Oct. 21, 2005) (unpublished student work), https://www.it.uu.se/edu/course/homepage/sakdat/ht05/assignments/pm/programme/DLL_Spoofing_in_Windows.pdf. Ordinary programs built with so-called objective oriented programming (OOP) tools, in effect, hide their basic functionality within the "objects" (components) sometimes built by others. *See* BROOKS, *supra* note 20, at 272. Practically speaking, no witness is likely to be able to explain the processing of all these components.

the central theories of the simulation, as opposed to more general theories of computation.

Setting issues of definition aside, more problematic concerns stem from the fact that software (including the inference engine) can be described at many levels of abstraction down to what some call "bare metal" (*i.e.*, the machine code that executes on the central processing unit (CPU)).

Courts apparently handle this problem of description in an ad hoc manner for there is little useful guidance in case law. There are suggestions that the foundation would include testimony "as to the accuracy of the equations used" in the simulation software,[82] or testimony on "a solid theoretical foundation and realistic input parameters,"[83] or that the proponent would "unravel his code and deduce the assumptions, algorithms, equations, and parameters that must be embedded within it," perhaps by "translat[ing] the foreign language of his computer model into a comprehensible language . . . ."[84] Some courts ask for a showing that "the input and underlying equations are sufficiently complete and accurate . . . and . . . the program is generally accepted by the appropriate community of scientists."[85] However, none of this tells us exactly what sort of explanation is enough to lay a foundation.

Judges certainly do not want to be led step by step through the bare metal code (*i.e.*, machine language[86]) or, just a bit less raw,

---

82. LAURIE L. LEVENSON, CAL. CRIM. PROC. § 22:26 (4th ed. 2016); *accord*, RUCKER & OVERLAND, *supra* note 49, at § 48:22.

83. BARBARA E. BERGMAN ET AL., WHARTON'S CRIM. EVID., § 16:22 (15th ed. 2016).

84. Novartis Corp. v. Ben Venue Labs., Inc., 271 F.3d 1043, 1054 (Fed. Cir. 2001).

85. Commercial Union Ins. Co. v. Boston Edison Co., 591 N.E. 2d 165, 168 (Mass. 1992).

86. In machine code, each instruction executes directly on the computer's CPU. Here's an example:

```
8020    78
8021    A9 80
8023    8D 15 03
8026    A9 2D
8028    8D 14 03
802B    58
802C    60
```

Here's another example:

```
00000000
00000001
00000010
00000100
00001000
```

assembly language,[87] which is understood by few individuals. Nor is it likely that a judge (or later, the jury) wants to be led through the next higher level of abstraction, source code, which most programmers use to write software.[88] At an even higher level of abstraction, there could be general flow-chart diagrams, but while these might summarize the components and processes of a system, they will not reflect most of the logical work or assumptions of the program. Those are *too* abstract.

Somewhere in between these levels of abstraction, there might be statistical formulas programmed into the inference engine, for example: [89]

00010000
00100000
01000000

87. This is an example:

```
Start:  .org $8020
   SEI
   LDA          #$80
   STA          $0315
   LDA          #$2D
   STA          $0314
   CLI
   RTS
   INC $D020
   JMP $EA31
   802D      EE 20 D0
   8030      4C 31 EA
```

88. Source code looks like this:

```
static void
print_cookies(CURL *curl)
{
  CURLcode res;
  struct curl_slist *cookies;
  struct curl_slist *nc;
  int i;
  printf("Cookies, curl knows:\n");
  res = curl_easy_getinfo(curl, CURLINFO_COOKIELIST, &cookies);
  if(res != CURLE_OK) {
    fprintf(stderr, "Curl curl_easy_getinfo failed: %s\n",
            curl_easy_strerror(res));
    exit(1);
  }
```

89. This is a formula used in a fuzzy logic inference engine expert system. *Supervised Learning and Fuzzy Logic Systems (Artificial Intelligence),* WHAT-WHEN-HOW, http://what-when-how.com/artificial-intelligence/supervised-learning-of-fuzzy-logic-systems-artificial-intelligence/ (last visited Nov. 14, 2017).

$$Z_1 = \frac{\sum_{l=1}^{M} y_k^l \left(\prod_{i=1}^{n} \mu A_l^i (x_i) \right)}{\sum_{l=1}^{M} y_k^l \left(\prod_{i=1}^{n} \mu A_l^i (x_i) \right)}$$

Together with the program itself, formulas such as this are good candidates for disclosure to the other party (that is, to the other side's expert) because they embody the statistical rules of the inference engine and, in effect, state the nature of the input and output. But on their own, they are of no help to the judge or the jury. While it possible to explain such formulas in plain English, other forms of representation are more useful.

As one commentator has suggested, this can be done in three ways: (i) propositional logic, (ii) fuzzy logic diagrams, and (iii) decision trees.[90] In propositional logic, the values of variables are stated. For example, A and B can be true or false or can be one of any specified range of values. The value could be numerical, or it could be something else. In the example provided by the commentator, A could have one of these values: {BUY, HOLD, SELL}. Then, relational operators, such as "less than" or "equal to," are used to compare the variables to other variables or values. Logical operators, such as And, Or, or But Not, may also be applied. Continuing with the example, imagine these variables as inputs: Price (P), Simple Moving Average (SMA), and Exponential Moving Average (EMA). The strategy might then look like this:[91]

IF (SMA > P) ∧ (EMA > P) THEN *BUY* ELSE

IF (SMA > P) ∧ (EMA < P) THEN HOLD

As for employ fuzzy logic, this provides a *range* over which a variable is true or belongs to a certain set. A program might conclude that a share of a company is a 20% BUY, 30% HOLD, and 50% SELL. Inputs too can be expressed over a range, expressing degrees of uncertainty, which corresponds at least a

90. Reid address neural networks, but his point is useful more generally. Stuart Reid, *10 Misconceptions About Neural Networks*, TURING FIN. (May 8, 2014), http://www.turingfinance.com/misconceptions-about-neural-networks/#blackbox.

91. "∧" means "and"; ">" means "greater than"; and "<" means "less than."

high level with the way in which layers in neural networks decide whether or not to pass on a finding to the next layer. Generated diagrams can then show that when the combined input from a series of sources exceeds a threshold, a decision is reached. For example, a medical diagnosis system might have degrees of certainty and uncertainty concerning inputs such as {has headaches - to some degree}, {has a rash - to some degree}, {is nauseous - to some degree}, {has difficulty breathing - to some degree}, and then express a result, such as {has Golem's Fever, with a specified level of certainty}.

Finally, decision trees show the impact of factors on a series of decisions. In flight simulators, for example, a series of models or subsystems, such as the aerodynamic, gear, weather, and engine models are inputs to equations which calculate motion, and those then in turn output to visual, sound, motion, instrument displays, and other outputs.[92] Each component model includes a series of equations. For example, a sophisticated engine model will produce figures for "engine thrust, fuel flows and engine pressures and rotation speeds . . . engine failure modes (*e.g.*. surge, stall or total failure) . . . [accounting for, *e.g.*,] engine characteristics [which] change considerably at low speeds and at very low altitude . . . ."[93] The number of equations in a flight simulator is far beyond what could possibly be addressed in a trial. Thus, here too the practical approach will distinguish and then ignore aspects of the program that are routine and presumably generally accepted from those that are novel. As to the latter, an expert can present a graphical representation of the decisions nodes,[94] the values for each which cause the node to make a decision one way or the other (*e.g.*, "if [engine temperature] > [5000 degrees], output ['explode']"), and a theory behind the figure, such as research that shows engines explode at certain temperatures.

There are two conclusions here. Importantly, for traditional expert systems, a human expert's competence in demonstrating, explaining, and justifying the theory behind a calculation is crucial. Both the judge determining admissibility and the jury determining weight look to the human expert to vouch for the simulation,[95] explain step-by-step the way in which the software works, state its

92. DAVID ALLERTON, PRINCIPLES OF FLIGHT SIMULATION 17 (2009).

93. *Id.* at 18.

94. *E.g.*, David Madigan et al., *Graphical Explanation in Belief Networks*, 6 J. COMPUTATIONAL & GRAPHICAL STAT. 160, 160-81 (1997).

95. Elaine M. Chaney, *Computer Simulations: How They Can Be Used at Trial and the Arguments for Admissibility*, 19 IND. L. REV. 735, 743 (1986).

assumptions and the valid scientific theories on which it is based,[96] as well as the logic used to derive its results.[97] While opinions may be based on the results of programs, the human witness takes the credit, or suffers the impeachment, for the opinion. If the expert cannot explain the model—how the software works and why it uses the numbers or formulas it does—then the evidence is inadmissible.

Secondly, there is a limit to explanation. All evidence at trial assumes other facts are true. Courts do not ask the contractor to prove her measuring tapes are accurate or the doctor to prove the blood pressure cuff was accurate. While courts may require, as foundation for eyewitness testimony, evidence that the person was at the scene, they do not demand testimony on how the eye and brain work to record and recall the memory recited in court. As noted above,[98] only a minimal foundation is required for the routine operations of computers. It is a waste of time to reinvent the wheel; routine operations are a given. So too with most of the foundation for the admissibility of simulations. Courts will usually forego *all* explanation at the levels of greatest precision (*i.e.*, the source code level). They quickly pass by even high-level descriptions of most of the calculations and built-in assumptions. At most, they will seek (i) high-level explanations of a few central formulas, (ii) the foundation (studies, etc.) which justifies those formulas, and (iii) to some extent, the logic that links those two things. This is as it should be; given the constraints of time, the expertise of most judges and juries and the essential task of trial is to focus relentlessly on core material issues. But courts should not delude themselves: the trustworthiness of much evidence, including computer simulations, depends on a practically infinite network of unarticulated assumptions. Nevertheless, courts will say the evidence is reliable.

---

96. *In re* TMI Litig.,193 F.3d 613, 669 (3d Cir. 1999), *amended*, 199 F.3d 158 (3d Cir. 2000). As the court noted, wonderfully summarizing the distinction in tests applicable to accepted versus new scientific theories, the "use of standard techniques bolster the inference of reliability; nonstandard techniques need to be well-explained." *In re* Zoloft (Sertraline Hydrochloride) Prod. Liab. Litig., No. 16-2247, 2017 WL 2385279, at *6 (3d Cir. June 2, 2017) (note omitted).

97. DAVID BOIES ET AL., *Computer Generated Evidence—Admissibility of Computer Simulations*, *in* ABA, BUSINESS AND COMMERCIAL LITIGATION IN FEDERAL COURTS § 66:17 (4th ed. 2016). *See generally*, Novartis Corp. v. Ben Venue Labs., Inc., 271 F.3d 1043, 1051 (Fed. Cir. 2001) (requiring demonstration of the "assumptions made by [the expert] in his computer model, and ask whether they are supported by evidence in the record. These include both the theoretical principles that informed the model's design as well as the means by which its input parameters were derived.").

98. *See* text at *supra* note 57.

III. ADMITTING MACHINE OPINIONS

Humans cannot explain how neural networks make their decisions.[99] But they can still establish that their results are reliable: humans can explain how the nets are trained, how they were successful in the past, and how they are successful with new data. These are the features that make neural networks reliable in the real world, and these are the factors that should make the output of these networks admissible in court, for reliability in the field is a sign that neural nets should be reliable in the courtroom.

Presented here are four arguments in favor of the reliability of machine opinions. First, society generally recognizes and trusts tacit expertise, the bases for which cannot be fully articulated. Second, society generally trusts medications, sometimes with people's lives, even when no one knows how they work. Third, neural networks are statistical models, and judges commonly rely on statistical models. Fourth, reliability is importantly a function of the ability to test and cross-examine, and neural networks can, practically, be cross-examined.[100]

A. Tacit Expertise

In Malcolm Gladwell's *Blink*,[101] an art expert views a Greek statue offered to the Getty Museum for $10 million. The expert declares it a forgery. He cannot quite say why, but he is right. Much expertise is tacit: it cannot be clearly articulated. This is also true in sports (*e.g.*, how a professional hits a baseball travelling at one hundred miles per hour[102]), music, teaching, decisions by

---

99. "'We can build these models,' Dudley says ruefully, 'but we don't know how they work.'" Will Knight, *The Dark Secret at the Heart of AI*, MIT TECH. REV. (Apr. 11, 2017), https://www.technologyreview.com/s/604087/the-dark-secret-at-the-heart-of-ai/ (quoting a researcher: "It might just be part of the nature of intelligence that only part of it is exposed to rational explanation. Some of it is just instinctual, or subconscious, or inscrutable.").

100. Jennifer L. Mnookin, *Repeat Play Evidence: Jack Weinstein, 'Pedagogical Devices,' Technology, and Evidence*, 64 DEPAUL L. REV. 571, 577–78 (2015) (suggesting courts permit the "opposing party to replace given assumptions with alternative ones" to enable cross-examination) ("To be sure, we do not normally imagine that machine-generated evidence requires cross-examination, but it may be time to begin thinking in those terms.") (note omitted).

101. MALCOLM GLADWELL, BLINK: THE POWER OF THINKING WITHOUT THINKING (2005).

102. The batter has about 125 milliseconds to decide, far less time than it takes to blink. This makes the task impossible. But batters use unspecific information from the movements of the pitcher *before the pitch* to estimate a

administrative agencies,[103] perhaps even judging,[104] and many other domains[105] where expertise can be observed, but not described. As opposed to novices who use explicit step-by-step processes, experts tend to use more conceptual structures to solve problems, but it is difficult to use these structures to actually explain the work to others.[106]

The inarticulate basis for some expert opinion presents a challenge to the usual way in which tests for admissibility are considered. For example, under California's *Sargon* test, judges should be presented with the express logic of the reasoning between (A) the opinion and (B) its foundation, including (1) the facts of the case and (2) the general theory and techniques used, as demonstrably founded on studies or other sources.[107] The test puts

---

likely pitch. Alex Kuzoian, *Hitting a Major League Fastball Should Be Physically Impossible*, BUS. INSIDER (Apr. 15, 2017), http://www.businessinsider.com/science-major-league-fastball-brain-reaction-time-2016-4.

103. Jacob Gersen & Adrian Vermeule, *Thin Rationality Review*, 114 MICH. L. REV. 1355 (2016).

104. LEE EPSTEIN ET AL., THE BEHAVIOR OF FED. JUDGES 5 (2013); Chad M. Oldfather, *Of Judges, Law, and the River: Tacit Knowledge and the Judicial Role*, 2015 J. DISP. RESOL. 155, 156 (2015) ("[M]uch of what goes into the process of decision-making is inarticulable.").

105. Let us not forget the work of high-end travel agents. R. Buckley & A.C. Mossaz, *Decision Making by Specialist Luxury Travel Agents*, 55 TOURISM MGMT. 133, 133-38 (2016).

106. Pamela J. Hinds & Jeffrey Pfeffer, *Why Organizations Don't 'Know What They Know': Cognitive and Motivational Factors Affecting the Transfer of Expertise*, *in* SHARING EXPERTISE 3, 5 (Mark S. Ackerman et al. eds., 2003) (Regarding experts' "conceptual, abstract representations is that they appear to be simplified representations of the task. As experts begin to automate aspects of the task, details of the task become less salient and experts begin to view the task in an oversimplified way. In an experiment, Langer and Imber (1979) found that experts' lists of task components contained significantly fewer and less specific steps than did the lists of those with less expertise. Developing abstract, simplified representations of the task allows experts to process information more rapidly, view the task holistically, and avoid getting bogged down in details. As such, abstract and simplified representations generally serve experts well. However, there are situations in which these representations can interfere with experts' ability to share their expertise, particularly with others who have significantly less expertise.").

107. SIMONS, *supra* note 45, at § 4:22. *See also* Jennings v. Palomar Pomerado Health Sys., Inc., 8 Cal. Rptr. 3d 363, 369 (Cal. Ct. App. 2003) ("[W]hen an expert's opinion is purely conclusory because unaccompanied by a reasoned explanation connecting the factual predicates to the ultimate conclusion, that opinion has no evidentiary value because an expert opinion is worth no more than the reasons upon which it rests.") (internal quotes omitted). *See generally*, CURTIS KARNOW, *Expert Witness:* Sargon *and the Science of Reliable Experts*, *in* LITIGATION IN PRACTICE 161-67 (2017).

a high premium on the articulation of the connection or "logic" between (a) studies and other foundations that establish a general theory or technique and (b) those theories (or techniques) and the facts of the case, on the one hand, and the ultimate opinion on the other. The reasoning or "logic" should be express. This allows the judge evaluating admissibility to determine that each step in the process is reliable.[108]

But this cannot be entirely right. Experts with "special knowledge, skill, experience, training, or education" can testify,[109] even though their experiences or skills may not be able to be described in minute detail. Their skills are generally built up from many years of experience as a banker or landowner testifying to value of property, or years of experience as a carpenter, plumber, or tile layer.[110] For some of these experts, there is only so much they can say about the foundation of their opinions. In contrast, there may be "even . . . a witness whose expertise is based purely on experience, say, a perfume tester able to distinguish among 140 odors at a sniff, whether his preparation is of a kind that others in the field would recognize as acceptable."[111]

The results in two relatively recent cases, one from the California Court of Appeal and one from the Ninth Circuit, may be explained at least in part by the notion of tacit expertise. To the surprise of some trial judges (presumably including the highly respected jurists reversed in these two cases), the trial courts' meticulous examination of the articulated foundations of the experts' testimony, which led to their exclusion, was set aside by the appellate courts. The appellate panels found that the trial judges had, in each case, glossed over the basic reliability of the opinion, as demonstrated by the experts' credentials and extensive experience.

In the state case, *Cooper*,[112] the trial judge had examined each study relied upon by the expert and found many problems. But

---

108. *E.g.*, *In re* Paoli R.R. Yard PCB Litig., 35 F.3d 717, 745 (3d Cir. 1994) (requiring "conclusions supported by good grounds for each step in the analysis . . . [such that] *any* step that renders the analysis unreliable under the *Daubert* factors" is exposed).

109. Cal. Evid. Code § 720(a) (Deering 2017).

110. *See* MICHAEL H. GRAHAM, 5 HANDBOOK OF FED. EVID. § 702:6 at n.7 (7th ed. 2016) (listing, extensively, occupations which qualify by virtue of experience).

111. Kumho Tire Co. v. Carmichael, 526 U.S. 137, 151 (1999), as noted by GRAHAM, *supra* note 110, at n.24.

112. Cooper v. Takeda Pharm. Am., Inc., 191 Cal. Rptr. 3d 67, 72-73 (Cal. Ct. App. 2015).

the appellate court said that the expert, a cancer specialist, had looked at all the studies together and based on his experience found them as a whole to be an adequate foundation. Importantly, the appellate court seems to have gone out of its way to cite the doctor's credentials and experience at length.[113] There were other issues in *Cooper* regarding the trial judge's approach, but the tenor of the opinion is that the witness was unquestionably an expert in the field, and if he found a basis for his opinions, then the trial judge was in no position to second-guess him.

In *Wendell*, the Ninth Circuit also chastised the trial court for:

> look[ing] too narrowly at each individual consideration, without taking into account the broader picture of the experts' overall methodology. It improperly ignored the experts' experience, reliance on a variety of literature and studies, and review of [the] medical records and history, as well as the fundamental importance of differential diagnosis by experienced doctors treating troubled patients.[114]

Here too, the appellate court gave a good deal of space to the experts' credentials and remarkable experience in the relevant fields, noting that the doctors used the same techniques used in their quotidian work for their court opinions.[115] The court made this point: "Nothing in *Daubert*, or its progeny, properly understood, suggests that the most experienced and credentialed doctors in a given field should be barred from testifying based on a differential diagnosis."[116]

In both cases, the trial courts' crusade to analyze each part of the foundation and each step in the logical progression from foundation to opinion—although seemingly called for by state and federal supreme court precedent—foundered on the rock of the witnesses' more general expertise, measured by their credentials such as their education, years of experience, experience in treating patients, list of publications, and the like.

These cases, and the fact that skilled experts' testimony is admissible even though it may be impossible to fully articulate the foundation for it, suggest that admissibility of expert opinion is often a function of the *general* reliability of the source, perhaps

113. *Id.* at 73-74.
114. Wendell v. GlaxoSmithKline LLC, 858 F.3d 1227, 1233 (9th Cir. 2017).
115. *Id.* at 1234.
116. *Id.* at 1235.

more so than the express articulation of the individuated reasons employed and steps taken in reaching the opinion. Courts already recognize tacit expertise, and neural networks have tacit expertise.[117]

B. The Drug Analogy

Even when prescription drugs are approved, not all the side effects, or the factors on which they depend, may be known. So too with their benefits: not all factors affecting the efficacy of a drug are known. Drugs are evaluated after they are first approved, and warnings may change over time. As the Food and Drug Administration (FDA) notes, "[i]n the end, no matter how much data are available, we often have to make a judgment call, weighing the known benefits against known risks and the potential—and possibly unknown—risks."[118] Importantly for present purposes, all that may be known about the approved drug is that it has certain benefits and other effects, while the details of why that is so may be unknown. The FDA may approve a drug for use, either over the counter or by prescription only, despite the fact that its mechanism is unknown. Examples include "acetaminophen for pain relief, penicillin for infections, and lithium for bipolar disorder, [which] continue to be scientific mysteries today."[119] A 2011 study of seventy-five drugs found that only seventeen were derived from a "detailed understanding of how the disease worked."[120] In short, after enough trials with a representative

---

117. *See generally*, Jason Millar & Ian Kerr, *Delegation, Relinquishment, and Responsibility: The Prospect of Expert Robots*, ROBOT L. 102, 109-13 (R. Calo et al., eds. 2016). Millar and his co-author suggest that (i) much "expert" knowledge is tacit, (ii) machine intelligence can or will manifest such knowledge, and (iii) we should defer to the conclusions of machine intelligence when it demonstrably does a better job than humans in given domains.

118. *How FDA Evaluates Regulated Products: Drugs*, FOOD & DRUG ADMIN., https://www.fda.gov/aboutfda/transparency/basics/ucm269834.htm (last visited Oct. 24, 2017).

119. Carolyn Y. Johnson, *One Big Myth About Medicine: We Know How Drugs Work*, WASH. POST (July 23, 2015), https://www.washingtonpost.com/news/wonk/wp/2015/07/23/one-big-myth-about-medicine-we-know-how-drugs-work/.

120. David C. Swinney & Jason Anthony, *How Were New Medicines Discovered?*, 10 NAT. REV. DRUG DISCOVERY 507 (2011). *See also*, Carmen Drahl, *How Does Acetaminophen Work? Researchers Still Aren't Sure,* 92 SCI. 31, 31-32 (2014); Tanya Lewis, *Mystery Mechanisms*, THE SCIENTIST (Jul. 29, 2016), http://www.thescientist.com/?articles.view/articleNo/46688/title/Mystery-Mechanisms/ ("Scientists still don't know exactly how some commonly used drugs work.").

population, the benefits and detriments of drugs may be adequately known to allow them to be used, even it is not known why they work.

In related contexts, society is comfortable with the fact that while we may not know the exact mechanisms, we know enough, generally through statistical studies, to find that a putative cause (such as a drug) has a certain effect (such as a birth defect).[121] To be sure, there is a difference between accepting an expert opinion on causation without knowing the mechanism of causation where, on the one hand, there is a demonstrated statistical basis, and, on the other, as in the case of a neural network, where there is not a demonstration of the specific statistical basis. But the truth is that neural networks *are* statistical analyses, and their reliability can be demonstrated through validation.

C. The Statistical Framework and Explaining the Logic

As the drug analogy demonstrates, statistics are used to make very serious decisions. Statistics are widely used in courts, and indeed judges may take judicial notice[122] of certain statistical facts.[123] Statistics are used to support and combat class certification decisions;[124] to evaluate DNA evidence;[125] to show and disprove racial disparity[126] and age discrimination;[127] to prove Fourth

121. Daubert v. Merrell Dow Pharm., *Inc.*, 43 F.3d 1311, 1314 (9th Cir. 1995) (cited in Wendell v. GlaxoSmithKline LLC, 858 F.3d 1227, 1233 (9th Cir. 2017)). In California courts, medical causation depends on expert testimony that there is a "reasonably probable causal connection" between the injury and alleged cause, Jones v. Ortho Pharm. Corp., 209 Cal. Rptr. 456, 461 (Cal. Ct. App. 1985), *i.e.*, greater than 50% odds. *See e.g.*, Uriell v. Regents of Univ. of California, 184 Cal. Rptr. 3d 79, 86-87 (Cal. Ct. App. 2015) (discussion of the state's "more probable than not" test); Cooper v. Takeda Pharm. Am., Inc., 191 Cal. Rptr. 3d 67, 85 (Cal. Ct. App. 2015) (same).

122. Facts that are the subject of judicial notice are those which are not reasonably the subject of dispute, and those "capable of immediate and accurate determination by resort to easily accessible sources of indisputable accuracy . . . ." Weaver v. United States, 298 F.2d 496, 498 (5th Cir. 1962). *See also, e.g.*, Cal. Evid. Code §§ 450 - 460; ROBERT I. WEIL ET AL., CAL. PRACTICE GUIDE: CIV. PRO. BEFORE TRIAL ¶ 7:13 (The Rutter Group 2017).

123. Envtl. Law Found. v. Beech-Nut Nutrition Corp., 185 Cal. Rptr. 3d 189, 203 n.7 (Cal. Ct. App. 2015).

124. Mies v. Sephora U.S.A., Inc., 184 Cal. Rptr. 3d 446 (Cal. Ct. App. 2015); Duran v. U.S. Bank Nat'l Ass'n, 68 Cal. Rptr. 2d 644 (Cal. Ct. App. 2014).

125. *See, e.g.*, People v. Venegas, 954 P.2d 525 (Cal. 1998).

126. Alston v. City of Madison, 853 F.3d 901, 908 (7th Cir. 2017); Paige v. California, 291 F.3d 1141 (9th Cir. 2002).

127. Karlo v. Pittsburgh Glass Works, LLC, 849 F.3d 61 (3d Cir. 2017).

Amendment violations;[128] in labor litigation;[129] to attack the practices at the Patent and Trademark Office;[130] to fix damages allocation in environmental clean-up actions;[131] and in many other situations. Courts often use regression analysis to estimate the impact of illegal acts.[132] Judges rely on statistical tools such as the STATIC-99 to evaluate the risk of recidivism of registerable sex offenders,[133] and many courts use survey results and their statistical conclusions in sentencing,[134] setting bail, and deciding issues such as the risk of recidivism and likelihood that defendants will appear for their next hearings.[135]

---

128. United States v. Soto-Zuniga, 837 F.3d 992, 1002 (9th Cir. 2016).

129. Nat'l Labor Relations Bd. v. Lily Transportation Corp., 853 F.3d 31 (1st Cir. 2017).

130. Ethicon Endo–Surgery, Inc. v. Covidien LP, 826 F.3d 1366, 1368 (Fed. Cir. 2016) (Newman, J., dissenting from denial of rehearing en banc).

131. Lyondell Chem. Co. v. Occidental Chem. Corp., 608 F.3d 284, 292 (5th Cir. 2010).

132. *E.g.*, *In re* Se. Milk Antitrust Litig., 739 F.3d 262, 285 (6th Cir. 2014), *cert. denied sub nom. See also* Dean Foods Co. v. Food Lion, LLC, 135 S. Ct. 676 (2014); Werdebaugh v. Blue Diamond Growers, No. 12–CV–2724–LHK, 2014 WL 2191901 (N.D. Cal. May 23, 2014) (proving damages under UCL, FAL, and CLRA); Kleen Prods. LLC v. Int'l Paper, 306 F.R.D. 585, 602 (N.D. Ill. 2015).

133. *Static-99/Static-99R*, STATIC99 CLEARINGHOUSE, http://www.static99.org (last visited Oct. 19, 2017); *see* CAL. PENAL CODE §§ 290.003-008 (Deering 2017).

134. Adam Liptak, *Sent to Prison by a Software Program's Secret Algorithms*, N.Y. TIMES (May 1, 2017), https://www.nytimes.com/2017/05/01/us/politics/sent-to-prison-by-a-software-programs-secret-algorithms.html?hp&action=click&pgtype=Homepage&clickSource=story-heading&module=first-column-region&region=top-news&WT.nav=top-news (describing C.J. Roberts' apparent reference to risk assessment software used in sentencing: "'Can you foresee a day,' asked Shirley Ann Jackson, president of the college in upstate New York [Rensselaer Polytechnic Institute], 'when smart machines, driven with artificial intelligences, will assist with courtroom fact-finding or, more controversially even, judicial decision-making?' The Chief Justice's answer was more surprising than the question. 'It's a day that's here,' he said, 'and it's putting a significant strain on how the judiciary goes about doing things.'")

135. *See, e.g.*, MARTIN FRISHER ET AL., PREDICTIVE FACTORS FOR ILLICIT DRUG USE AMONG YOUNG PEOPLE: A LITERATURE REVIEW (2007); Kristin Bechtel et al., *Identifying the Predictors of Pretrial Failure: A Meta-Analysis*, 75 FED. PROB. 78 (2011); Peter J. Henning, *Is Deterrence Relevant in Sentencing White-Collar Criminals?*, 61 WAYNE L. REV. 27, 38 (2015); Curtis Karnow, *Setting Bail for Public Safety*, 13 BERKELEY J. CRIM. L. (2008); Arthur L. Kellermann et al., *Gun Ownership as a Risk Factor for Homicide in the Home*, 329 NEW ENG. J. MED. 1084, 1084 (1993); Jason Tashea, *Kentucky Tests New Assessment Tool to Determine Whether to Keep Defendants Behind Bars*, ABA J. (2015); Marie VanNostrand & Gena Keebler, *Pretrial Risk Assessment in the Federal Court*, 73 FED. PROB. 1 (2009); Angèle Christin et al., *Courts and*

Neural networks are, in effect, statistical models.[136] A valid (or "statistically significant") result shows a certain degree of correlation; it does not prove causation. But with traditional statistical studies, at some point either the strength of a study, or better, many studies,[137] is enough to conclude that, given some predicate or sample, a more general conclusion is very likely to be true. Statistics use samples to tell us something about the world; they provide, based on partial data, an inference about general patterns. This how predictive coding works: given a sample of documents on which the system has been trained, some of which are privileged (or relevant), the system determines which of the larger collection are privileged (or relevant). AlphaGo works, in part, the same way: it first extrapolates the present board pattern to an intermediate but limited set of possible patterns, given a series of possible moves. Then, it compares the intermediate set to all the patterns it has experienced.[138] Knowing which of that larger set led to victory, AlphaGo then chooses the best intermediate pattern.[139]

Predictive Algorithms, DATA & C.R.: A NEW ERA OF POLICING & JUST. (Oct. 27, 2015), http://www.datacivilrights.org/pubs/2015-1027/Courts_and_Predictive_Algorithms.pdf; TAYLOR TILLMAN, RISK FACTORS PREDICTIVE OF JUVENILE OFFENDER RECIDIVISM (May 2015), https://scholarworks.umt.edu/etd/4495/.

136. For a technical discussion, *see* Raul Rojas, *Statistics and Neural Networks*, *in* NEURAL NETWORKS 229, 229–63 (1996).

137. Because individual studies may reflect cherry-picking and other problems, studies that review the results of many studies, known as metastudies, are preferred. Ben Goldacre, *Listen Carefully, I Shall Say This Only Once*, THE GUARDIAN (Oct. 25, 2008), https://www.theguardian.com/commentisfree/2008/oct/25/medical-research-science-health (discussing problems with issuing multiple reports of what is, in fact one study; contrasting results of the "one" study with true metastudy results); *AllTrials*, ALLTRIALS, http://www.alltrials.net (importance of meta studies). For discussion of a leading effort in this regard, *see* Cochrane, *What Are Systematic Reviews,* YOUTUBE (Jan. 27, 2016), https://www.youtube.com/watch?v=egJlW4vkb1Y; *Reporting Biases*, COCHRANE, http://methods.cochrane.org/bias/reporting-biases (last visited Nov. 10, 2017) (Cochrane furthers transparency in research and publication, and use of metastudies); *What is Cochrane Evidence and How Can It Help You?*, COCHRANE, http://www.cochrane.org/what-is-cochrane-evidence (last visited Nov. 10, 2017).

138. This comprises *millions* of games, orders of magnitude more games than any human could play in a lifetime. *See Full Length Games for Go Players to Enjoy*, DEEPMIND, https://deepmind.com/research/alphago/alphago-vs-alphago-self-play-games (last visited Nov. 9, 2017).

139. David Silver et al., *Mastering the Game of Go with Deep Neural Networks and Tree Search*, 529 NATURE 484 (2016); Christof Koch, *How the Computer Beat the Go Master*, SCI. AM. (Mar. 19, 2016), https://www.scientificamerican.com/article/how-the-computer-beat-the-go-master; David Silver, *AlphaGo: Mastering the Ancient Game of Go with Machine Learning*, GOOGLE

There is, of course, a difference between the operations of neural networks and the traditional statistical expert because the human expert "shows his work." He writes out his calculations, which can then be inspected for errors.

The work of a neural network can similarly be checked. Evaluations or checks can sort fallacious from reliable inferences by testing a neural network on new data, validating the system. This sort of testing can show that a statistical correlation is invalid, *i.e.*, that it is just a random product.[140] Correlations can be found among almost any set of facts,[141] but this is cherry-picking and does not reflect a hypothesis that is then tested on new data. Thus, these untested correlations are fallacious. They are random correlations selected after the fact—that is the cherry picking—because in isolation they appear to present a pattern.

A fundamental element of training (whether human-supervised or not) neural networks is a feedback loop that tests whether the learned correlations play out correctly on *new* data, just as networks used in business should have the predicted result constantly compared to real-world events. This step is sometimes referred to as *validation.*[142] Sometimes validation is performed by using a portion of the data available when the system was first trained: this data is split into training and testing data, and the testing data is used to validate. In TAR, for example, the system might be tested on a sample of the general production documents.

But performing well just on this limited set of testing data may not be a sufficient foundation to trust the system more generally. If the testing data are few, or not homogenous (*i.e.*, not identically distributed), the test may not show much.[143]

RES. BLOG (Jan. 27, 2016), https://research.googleblog.com/2016/01/alphago-mastering-ancient-game-of-go.html.

140. A favorite example of fallacious induction is a Thanksgiving Day turkey, which extrapolates from just under a year's worth of daily good feeding that Thursday, November 23, 2018, will be a good day. It will not be.

141. Such as (i) the number of people who drown by falling into a pool and films Nicolas Cage appeared in, or (ii) per capita cheese consumption and the number of people who died entangled in bedsheets; and so on and so forth. *See* Tyler Vigen, *Spurious Correlations*, http://www.tylervigen.com/spurious-correlations (last visited Nov. 9, 2017); for more on bad or fallacious inferences, *see* CURTIS KARNOW, *Statistics & Probability: Bad Inferences and Uncommon Sense*, *in* LITIGATION IN PRACTICE 43 (2017).

142. *See, e.g.*, BRIAN CHRISTIAN & TOM GRIFFITHS, ALGORITHMS TO LIVE BY 159 (2016).

143. *See e.g.*, Sylvain Arlot & Alain Celisse, *A Survey of Cross-Validation Procedures for Model Selection*, 4 STAT. SURVEYS 40, 52 (2010), https://projecteuclid.org/euclid.ssu/1268143839; *see also* Damjan Krstajic et al.,

AlphaGo has been cross-validated and then proven reliable in the field: it beat the human champions. By the same token, networks can be tested against known results. And a human witness can discuss, or challenge, the performance of a network against new data, both cross-validated against data that was part of the initial dataset *and* later in the field on entirely new data. The proponent of the machine opinion (*e.g.*, for facial recognition or medical diagnosis) reports the performance of the network against new data and states that the program had correct results a certain percentage of the time. The party opposing admissibility or later disputing the weight to be given to the opinion could report results on his own set of new data. This process requires the software to be provided to all parties in order to allow for this sort of "cross-examination." The selection of data by the party opposing admissibility should be generated to detect flaws, such as training on inapposite data.

Furthermore, even though the technical calculation of the opinion is not available to any human, the proponent of the machine opinion should be able to provide to the judge or jury an abstracted view of the logic flow that produces the opinion, similar to that available for a more traditional expert software with propositional logic, fuzzy logic diagrams, or decision trees.[144] The point here is not just that these three approaches can be used generally to illustrate machine decision making, but rather that there are tools to extract these illustrations *from the specific neural network* (*i.e.*, its inputs and outputs) at issue.[145] While, again, these illustrations are not and cannot be descriptions of the actual mechanism of the hidden layers, nor justifications for them, they

---

Cross-Validation Pitfalls When Selecting and Assessing Regression and Classification Models, 6 J. CHEMINFORMATICS 10, 10 (2014), https://link.springer.com/article/10.1186/1758-2946-6-10 ("In an ideal situation we would have enough data to train and validate our models (training samples) and have separate data for assessing the quality of our model (test samples). Both training and test samples would need to be sufficiently large and diverse in order to be representative [sic].").

144. *See* Part III.C.2, *infra*.

145. Stuart Reid, *10 Misconceptions About Neural Networks*, TURING FIN. (May 8, 2014), http://www.turingfinance.com/misconceptions-about-neural-networks/#blackbox. Researchers continue to develop tools used to at least illustrate the details of training of specific networks. *See, e.g.*, Jason Yosinski et al., *Understanding Neural Networks Through Deep Visualization*, DEEP LEARNING WORKSHOP, 32ND INT'L CONF. ON MACHINE LEARNING 4 (2015), https://arxiv.org/pdf/1506.06579.pdf ("We describe and release a software tool that provides a live, interactive visualization of every neuron in a trained convnet as it responds to a user-provided image or video.").

are probably as detailed as any descriptions provided to judges and juries in connection with more traditional inference engines.[146] That is, it often does not take much to provide as much detail as judges and juries really want, or need, in the evaluations of the proponent's foundation, because the real test for reliability, discussed next, is measured by the ability of the opponent to *challenge* the machine opinion.

D. Risks & Cross-Examination

The ability to cross-examine is the classic test of reliability, and reliability is the cornerstone of admissibility.[147] Recall the *sine qua non* of cross-examination, which is that the program must be made available to the opposing side in order to be tested against new data.

To a shocking degree, many ordinary "expert" systems are in fact never tested against new data or real-world results; their predictions are never analyzed, that is, they are not validated. They are used because they are convenient, or because they are cheaper than using humans or give the appearance of objectivity or infallibility. They may be used because they deflect responsibility from whoever would otherwise be the human agent, or to save time. Companies use software to hire and fire but never determine whether the results were as predicted. Colleges use a variety of criteria to admit students and use tests to measure competence in academic areas, but these decisions may or may not have ever been validated. For example, did the students with higher scores actually perform better in college? Did algorithms used to pick stocks actually do better than human decision-makers with the same information? As a matter of fact, what was the performance of loans when a program made the lending decision? Algorithms are used to suggest products on Amazon, movies on Netflix, plots for movies, patterns for room-cleaning robots,[148] and for meeting people on dating sites online. Some of these *are* validated, especially, as in the case of companies such as Amazon and Netflix, because the accuracy of the algorithm translates to millions of dollars in revenue.

---

146. See text at Chaney, *supra* note 95, at 743.

147. This is true under both federal and California law. *See, e.g.*, Sargon Enters. v. Univ. of S. Cal., 288 P.3d 1237, 1252 (Cal. 2012).

148. Kevin Slavin, *How Algorithms Shape Our World*, TED (July 2011), https://www.ted.com/talks/kevin_slavin_how_algorithms_shape_our_world/transcript?language=en.

But many programs are never validated. The problem is sufficiently serious and pervasive that an entire book could be written about it. Indeed, it has been.[149] Worse, at least from the point of view of those in the court system, many programs and tests used in criminal trials are of dubious validity because there are no accepted validation benchmarks, no validation tests are used, or the level of precision announced to the jury is far in excess of the true value.[150]

Thus, one risk in using programs is a failure of validation; that is, either none was performed, or the validation was conducted on unrepresentative data. And while this Article presses the notion that software on which worldly activities (such as businesses) rely is generally sufficiently trustworthy to be used in court, this is a critical caveat, an exception to the rule that may swallow it. "Unvalidated" software is used all the time in the real world, but, as is true of any of the fallacies that infect daily life, has no place in court.

The critical importance of testing out a model, or any predictive system, on new data is exemplified by an algorithm

---

149. CATHY MCNEIL, WEAPONS OF MATH DESTRUCTION (2016); for a discussion of the issues, *see, e.g.*, Mary-Ann Russon, *The Dangers of Big Data: How Society Is Being Controlled by Mathematical Algorithms*, INT'L BUS. TIMES (Sept. 13, 2016), http://www.ibtimes.co.uk/dangers-big-data-how-society-being-controlled-by-mathematical-algorithms-1581174. McNeil discussed the lack of feedback mechanisms (*i.e.*, validation) in many areas, such as using algorithms to hire and fire teachers, *id.* at 7, 138, evaluating other potential employees, *id.* at 7, 111, issuing credit rating, *id.* at 146, and so on.

150. PRESIDENT'S COUNCIL OF ADVISORS ON SCI. & TECH. (PCAST), FORENSIC SCIENCE IN CRIMINAL COURTS: ENSURING SCIENTIFIC VALIDITY OF FEATURE-COMPARISON METHODS (Sept. 2016), https://obamawhitehouse.archives.gov/sites/default/files/microsites/ostp/PCAST/pcast_forensic_science_report_final.pdf (highlighting problems with tests regarding certain DNA, bite-mark, firearms, hair comparison, fingerprint, and footwear). For serious problems with latent fingerprint testimony, *see* United States v. Llera Plaza, 179 F. Supp. 2d 492, 494 (E.D. Pa. 2002) *withdrawn from bound volume, opinion vacated and superseded on reconsideration*, 188 F. Supp. 2d (E.D. Pa. 2002) (Pollak, J.). In almost all criminal trials for a period over twenty years in which a group of FBI fingerprint experts testified, including trials of thirty-two defendants sentenced to death, the FBI experts gave flawed testimony; twenty-six experts overstated forensic matches in ways that favored prosecutors in over ninety-five percent of the 268 trials reviewed as of April 2015. *See* Spencer S. Hsu, *FBI Admits Flaws in Hair Analysis Over Decades*, WASH. POST (Apr. 18, 2015), https://www.washingtonpost.com/local/crime/fbi-overstated-forensic-hair-matches-in-nearly-all-criminal-trials-for-decades/2015/04/18/39c8d8c6-e515-11e4-b510-962fcfabc310_story.html?utm_term=.e63ad6b8db16.

from *Market Watch*'s Gary Smith.[151] His algorithm shows a remarkable eighty-eight percent correlation between predicted and actual stock prices for 2015, including an almost perfect match of the drop in the third quarter:

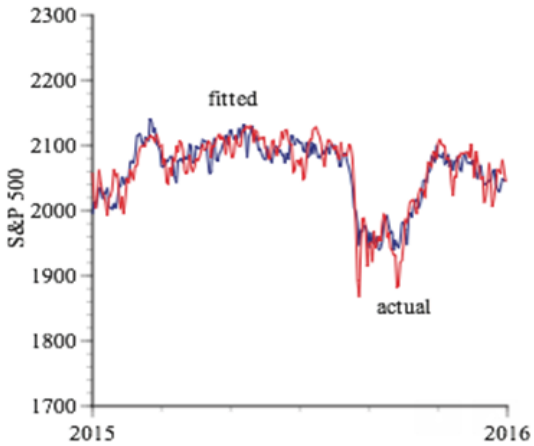


Figure 1 My Model of Stock Prices

But tested on new data—prices in 2016—it was a complete failure:

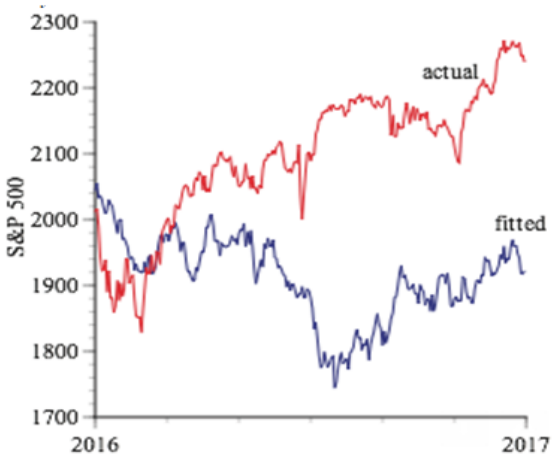


Figure 2 My Model with Fresh Data

And as with any validation, whether of a drug, a test, or some other screening device, validation must be conducted on the

151. Gary Smith, *Opinion: This Experiment Shows the Danger in Black-Box Investment Algorithms*, MARKETWATCH (June 17, 2017), http://www.marketwatch.com/story/this-experiment-shows-the-danger-in-black-box-investment-algorithms-2017-06-13.

relevant population and the relevant data. When testing drugs for childhood cancer, was the drug tested on seventy-year-olds with cancer because those subjects were more easily located? When testing an algorithm for recidivism and examining factors such as type of crime, income, or whether job history correlates with new crimes, or when looking at a program which predicts loan failures, was the validation population from the same type of locale (*i.e.*, area of country, rural versus inner city) as the population on which the algorithm is to be used?

Underlying this issue is the problem of what *counts* as validation. The matter is relatively obvious with AlphaGo, because it keeps beating every new opponent. So too with predictive coding of documents: lawyers can examine the program's decisions on new data and score accuracy. And then they spot-check the final result. In these cases, the selection of a "new" population of items used for validation is easy. But it is less certain what the new data (used for cross-validation) might be for systems designed, for example, to do handwriting analysis, facial recognition, medical diagnosis, or predict recidivism.

For facial recognition systems, validation data might include photographs taken under a variety of lighting conditions and of various angles of the face, some of which will reveal few of the facial features. "Successful" testing is likely to depend on which data used. Similarly, with handwriting analysis, validation data might include a wide variety of legible and illegible scrawls, initials, small and large groupings of letters. "Success" is likely to depend on which of these are used. Medical diagnosis too depends on an unconstrained or arbitrary number of inputs, from a few to a very large number, such as body temperature, blood chemistry, as well as a range of vaguely reported conditions such as nausea, pain, skin tone, and extent of bruising. Recidivism may depend on different factors, such as geographic or socio-economic distribution. In all of these situations, and doubtless in others as well, "success" with one group of validation data may or may not be persuasive.

More technically, with respect to neural networks, the validation data must meet certain criteria, such as that it not be the data used during training.[152] With a large dataset, a chunk of it, perhaps twenty percent, can be set aside specifically for validation testing and is never used for the initial training.[153] When there is a small dataset, or some question that the initial training dataset and

152. GOODFELLOW ET AL., *supra* note 6, at 118.

153. *Id.*

the testing set are not similar, various methods can make a series of passes through the system with randomly selected parts of the dataset, a process termed as "cross-validation."[154] But while there many cross-validation techniques, they all assume that the dataset is representative of the data on which the system will be let loose after its training. And in practice, if not in theory, datasets are limited. Thus, even with sophisticated, competent cross-validation, the performance of the system in the "field," as it were, may not match that in the laboratory.

With programs that have extensive past experiences (*i.e.*, they have worked on exceedingly large sets of data in the past) and have been tested on very large sets of validation data, these concerns will tend to dissipate. The fact that AlphaGo has played millions of games obviates concerns that it may not be successful in the next game against a top professional player. The extent of the training and of the validation data in effect tells us that the next test—the one in the "field" for which it provides the opinion in issue—is not "unexpected," not an outlier.

It is no coincidence that neural networks have made their mark just as "big data" erupted. It is commonplace to remark that there is a stunning amount of data,[155] as a function not only of efforts to

---

154. *Id.* at 118-19. *See generally, e.g.*, G. Varoquaux, et al., *Assessing and Tuning Brain Decoders: Cross-Validation, Caveats, and Guidelines*, 145 NEUROIMAGE 166 (2017), https://arxiv.org/pdf/1606.05201.pdf; Ron Kohavi, *A Study of Cross-Validation and Bootstrap for Accuracy Estimation and Model Selection*, 2 PROCEEDINGS OF THE 14TH INT'L JOINT CONF. ON ARTIFICIAL INTELLIGENCE 1137 (1995), https://pdfs.semanticscholar.org/0be0/d781305750b37acb35fa187febd8db67bfcc.pdf; Anders Krogh & Jesper Vedelsby, *Neural Network Ensembles, Cross Validation, and Active Learning*, *in* ADVANCES IN NEURAL INFORMATION PROCESSING SYSTEMS 231 (1995), http://papers.nips.cc/paper/1001-neural-network-ensembles-cross-validation-and-active-learning.pdf; Andrew W. Moore & Mary S. Lee *Efficient Algorithms for Minimizing Cross Validation Error*, 1994 PROCEEDINGS OF THE 11TH INT'L CONF. ON MACHINE LEARNING 190, 190 (1994), https://pdfs.semanticscholar.org/352c/4ead66a8cf89b91f9de5ac86bc69f17b29d0.pdf.

155. Any measure of currently available data exceeds the mind's ability to understand, but for those of us who love words like petabytes and (even better) zettabytes, there is this: "The total amount of data in the world was 4.4 zettabytes in 2013. That is set to rise steeply to 44 zettabytes by 2020. To put that in perspective, one zettabyte is equivalent to 44 trillion gigabytes." Every day about 2.5 exabytes are produced, equivalent to 250,000 Libraries of Congress. Mikal Khoso, *How Much Data Is Produced Every Day?*, LEVEL (May 13, 2016), http://www.northeastern.edu/levelblog/2016/05/13/how-much-data-produced-every-day/. *See*, *e.g., The Zettabyte Era: Trends and Analysis*, CISCO (June 7, 2017), http://www.cisco.com/c/en/us/solutions/collateral/service-provider/visual-networking-index-vni/vni-hyperconnectivity-wp.html (discussing the interesting

digitize past records but also the recordation of communications such as email, texts, searches, and social media which have taken the place of unrecorded oral communications of the past. The accumulation of this data has not only made it imperative to have software capable of digesting it, but is the very basis for the tools—neural networks—needed to do so. The exceedingly large data set makes it reasonable to trust validation tests that use that data, all without knowing why it is that the validation is successful, that is, without either having a theory of correlation nor knowing the details of the underlying mechanism that explains the found correlation.[156]

There are two other dangers in the use of neural networks, which should be the subject of future academic review. The first, the use of proxies, is closely related to the makeup of the validation discussed previously. Assume a network looks at the relationship between a series of factors and fluctuations in fish stock.[157] A deeper review of the system might reflect not the designer's decision to measure the fish stock directly but some proxy for it, such as fish caught or consumed. Or a system designed to opine on earthquake damage might use simplified input of proxies of certain soil conditions.[158] A network might provide an opinion on the valuation of initial public offerings but actually use a proxy, such as valuation of certain stock one day or one week after the initial offering date. These may all be reasonable, but the underlying assumptions should be made manifest; sometimes, it may indicate a mismatch between the training data and the validation data on the one hand and the proposed input for the specific opinion at issue on the other.

Finally, there is bias. The cartoon conceit is that algorithms are unbiased; the computer is neutral, and free of prejudice. Without

---

predictions by Cisco, which as a manufacturer of Internet servers and related equipment, presumably should know) ("It would take more than 5 million years to watch the amount of video that will cross global IP networks each month in 2021.").

156. *See* Chris Anderson, *The End of Theory: The Data Deluge Makes the Scientific Method Obsolete*, WIRED (June 23, 2008), https://www.wired.com/2008/06/pb-theory/.

157. *Cf., e.g.*, D. G. Chen & D. M. Ware, *A Neural Network Model for Forecasting Fish Stock Recruitment*, 56 CANADIAN J. FISHERIES & AQUATIC SCI. 2385 (1999), https://doi.org/10.1139/f99-178.

158. C. Salameh et al., *Estimation of Damage Level at Urban Scale from Simple Proxies Accounting for Soil and Building Dynamic Properties*, 2017 PROCEEDINGS OF THE 16TH WORLD CONF. ON EARTHQUAKE ENG'G 2049 (2017), https://hal.archives-ouvertes.fr/hal-01461198/document.

programming, the empty computer surely is. But most neural networks, even those which improve with self-training, begin their existence trained by humans. They literally model themselves on human choice and predilection. Some of these systems are in effect told that success is doing things the way humans would, and failure is diverging from those human choices. In this way, human biases become embedded in the very fabric of the system's decisions. The impact may be the most significant in what appear to be complex, subjective decisions such as hiring, evaluating written essays, and face recognition.[159] An interesting study of the way thirty thousand images were used to train networks to recognize content found that human stereotypes on gender and race—*i.e.*, prejudices—were routinely derived from the human-tagged dataset.[160] The danger is obvious: courts—and society—must not rely on software that reproduces human cognitive failures, but those very human cognitive failures make it difficult to discern the software's failure.

IV. CONCLUSION

The results of well-trained neural networks are trusted in the world, and they can be trusted in the courts. Perfection is not guaranteed,[161] but neither is it guaranteed with already routinely

---

159. Joy Buolamwini, *How I'm Fighting Bias in Algorithms*, TED (Nov. 2016), https://www.ted.com/talks/joy_buolamwini_how_i_m_fighting_bias_in_algorithms/transcript?language=en. *See generally*, Nanette Byrnes, *Why We Should Expect Algorithms to Be Biased*, MIT TECH. REV. (June 24, 2016), https://www.technologyreview.com/s/601775/why-we-should-expect-algorithms-to-be-biased/; Claire Miller, *When Algorithms Discriminate*, N.Y. TIMES (July 9, 2015), https://www.nytimes.com/2015/07/10/upshot/when-algorithms-discriminate.html.

160. Emiel van Miltenburg, *Stereotyping and Bias in the Flickr30K Dataset*, 2016 PROCEEDINGS OF THE WORKSHOP ON MULTIMODAL CORPORA: COMP. VISION AND LANGUAGE PROCESSING 1-4 (2016), https://arxiv.org/pdf/1506.06579.pdf. *See also Human Prejudices Sneak into Artificial Intelligence*, NEUROSCIENCE NEWS (Apr. 14, 2017), http://neurosciencenews.com/artificial-intelligence-human-prejudice-6411/ (discussing Aylin Caliskan et al., *Semantics Derived Automatically from Language Corpora Contain Human-Like Biases*, 356 SCI. 183 (2017)). For a report on possible racial bias in software, which may *not* be a trained neural network, used for sentencing and bail decisions, see Julia Angwin et al., *Machine Bias*, PROPUBLICA (May 23, 2016), https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing.

161. GOODFELLOW ET AL., *supra* note 6, at 193.

accepted testimony, such as eyewitness evidence[162] confessions, which may have peculiar reliability problems,[163] nor other routine testimony, which may be false or misremembered.[164]

General admissibility rules are not meant to be onerous. The default is that all relevant evidence is admissible,[165] and if an opinion is reliable and relates to a contested fact, it is surely relevant. The foundations of expert testimony typically must be explained to the judge and jury, but this Article demonstrates that, in the case of opinions generated by neural networks, the fact that the specific basis for the opinion cannot be demonstrated or articulated should not block the admission of the opinion, because the opinion may yet be fundamentally reliable and remains subject to meaningful cross-examination. It may be that in the first case or two in which true machine opinion is offered, a so-called *Kelly* hearing may be warranted, because a court may find that the neural network is in this context an "unproven technique or procedure [used] . . . to provide some definitive truth which the expert need only accurately recognize and relay to the jury."[166] To avoid presenting the jury with a "misleading aura of certainty,"[167] the court may examine the technology. It may find that the basic technology is sound, widely used and accepted in the real world; that it is reliable because correct scientific procedures (*e.g.*, accepted statistical algorithms) were used to build and train the program; and that the network is helpful to the jury. With all

---

162. Cross-racial witness identification can pose serious problems. *See* New Jersey's approach in State v. Henderson, 208 N.J. 208, 267 (2011), *holding modified by* State v. Chen, 208 N.J. 307, 327 (2011); State v. Romero, 191 N.J. 59, 68 (2007). *See also* United States v. Langford, 802 F.2d 1176, 1182 (9th Cir. 1986).

163. *E.g.*, People v. McCurdy, 59 Cal. 4th 1063, 1109 (2014); Campos v. Stone, 201 F. Supp. 3d 1083, 1099 (N.D. Cal. 2016). *See generally*, Steven A. Drizin & Richard A. Leo, *The Problem of False Confessions in the Post-DNA World*, *in* 82 N.C. L. REV. 891 (2004); Welsh S. White, *False Confessions and the Constitution: Safeguards Against Untrustworthy Confessions*, 32 HARV. C.R.-C.L. L. REV. 105, 119 (1997) (standard interrogation guidelines may induce false confessions).

164. *Cf.*, Trear v. Sills, 69 Cal. App. 4th 1341, 1345 (1999); F. LEE BAILEY & KENNETH J. FISHERMAN, CRIM. TRIAL TECHNIQUES § 58:11 (2d ed. 1996).

165. Cal. Evid. Code § 350 (Deering 2017).

166. People v. Jackson, 1 Cal. 5th 269, 316 (2016), *quoting* People v. Stoll, 49 Cal. 3d 1136, 1155–56 (1989).

167. People v. Kelly, 17 Cal. 3d 24, 32 (1976), *quoting* Huntingdon v. Crowley, 64 Cal. 2d 647, 656 (1966). *See generally* SIMONS, *supra* note 45, at §4:27.

parties being well-informed and able to validate functionality, machine opinions may provide insight no human can offer.