Chapter 9

Fragility and Intelligibility of Deep Learning for Libraries

Michael Lesk
Rutgers University

Introduction

On February 7, 2018, Mounir Mahjoubi, then the "digital minister" of France (*le secrétariat d'État chargé du Numérique*), told the civil service to use only computer methods that could be understood (Mahjoubi 2018). To be precise, what he actually said to l'Assemblée Nationale was:

> Aucun algorithme non explicable ne pourra être utilisé.

I gave this to Google Translate and asked for it in English. What I got (on October 13, 2019) was:

> No algorithm that can not be explained can not be used.

That's a long way from fluent English. As I count the "not" words, it's actually reversed in meaning. But, what if I leave off the final period when I enter it in Google Translate? Then I get:

> No non-explainable algorithm can be used

Quite different, and although only barely fluent, now the meaning is right. The difference was only the final punctuation on the sentence.[1]

This is an example of the fragility of an AI algorithm. The point is not that both translations are of doubtful quality. The point is that a seemingly insignificant change in the input produced such a difference in the output. In this case, the fragility was detected by accident.

[1] In the months between my original queries in October 2019 and the final preparations for publication in November 2020, the algorithm has changed to produce the same translation with or without a period: "No non-explicable algorithm can be used."

Machine learning systems have a set of data for training. For example, if you are interested in translation, and you have a large collection of text in both French and English, you might notice that the word *truck* in English appears where the word *camion* appears in French. And the system might "learn" this translation. It would then apply this in other examples; this is called generalization. Of course if you wish to translate French into British English, a preferred translation of *camion* is *lorry*. And if the context of your English *truck* is a US discussion of the wheels and axles underneath railway vehicles, the better French word is *le bogie*.

Deep learning enthusiasts believe that with enough examples, machine learning systems will be able to generalize correctly. There can be various kinds of failures: we can discuss both (a) problems in the scope of the training data and (b) problems in the kind of modeling done. If the system has sufficiently general input data so that it learns well enough to produce reliably correct results on examples it has not seen, we call it *robust*; robustness is the opposite of fragility. Fragility errors here can arise from many sources—for example, the training data may not be representative of the real problem (if you train a machine translation program solely on engineering documents, do not expect it to do well on theater reviews). Or, the data may not have the scope of the real problem: if you train for "boat" based on ocean liners, don't be surprised if the program fails on canoes.

In addition, there are also modeling issues. Suppose you use a very simple model, such as a linear model, for data that is actually perhaps quadratic or exponential. This is called "underfitting" and may often arise when there is not enough training data. The reverse is also possible: there may be a lot of training data, including many noisy points, and the program may decide on a very complex model to cover all the noise in the training data. This is called "overfitting" and gives you an answer too dependent on noise and outliers in your data. For example, 1998 was an unusually warm year, but the decline in world temperature for the next few years suggests it was noise in the data, not a change in the development of climate.

Fragility is also a problem in image recognition ("AI Recognition" 2017). Currently the most common technique for image recognition research projects is the use of convolutional neural nets. Recently, several papers have looked at how trivial modifications to images may impact image classification. Here (figure 9.1) is an image taken from (Su, Vargas, and Sakurai 2019). The original image class is in black and the classifier choice (and confidence) after adding a single unusual pixel are shown in blue, with the extraneous pixel in white. The images were deliberately processed at low resolution—hence the pixellation—to match the input requirement of a popular image classification program.

The authors experimented with algorithms to find the quickest single-pixel change that would deceive an image classifier. They were routinely able to fool the recognition software. In this example, the deception was deliberate; the researchers searched for the best place to change the image.

Bias and mistakes

We have seen a major change in the way we do machine learning, and there are real dangers involved. The current enthusiasm for neural nets risks the use of processes which cannot be understood, as Mahjoubi warned, and which can thus conceal methods we would not approve of, such as discrimination in lending or hiring. Cathy O'Neil has described this in her book *Weapons of Math Destruction* (2016).

There is much research today that seeks methods to explain what neural nets are doing. See

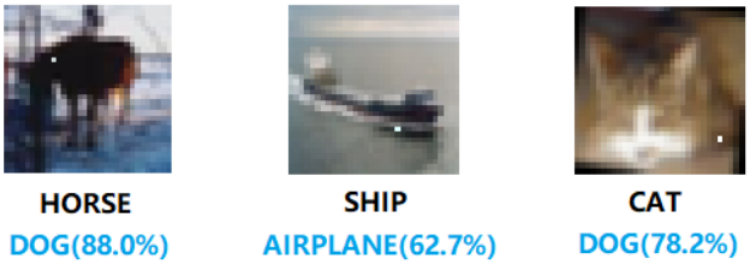

Figure 9.1: Examples of misclassification.

Guidiotti et al. (2017) for a survey. There is also a 2018 DARPA program on "Explainable AI." Techniques used can include looking at the results over a range of input data and seeing if the neural net can be modeled by a decision tree, or modifying the input data to see which input elements have the greatest effect on the results, and then showing that to the user. For example, Mariusz Bojarski et al. describe a self-driving system that highlights what it thinks is important in what it is seeing (2017). However, this is generally research in progress, and it raises the question of whether we can trust the explanation generator.

Many popular magazines have discussed this problem; *Forbes,* for example, had an explanation of how the choice of datasets can produce a biased result without any deliberate attempt to do so (Taulli 2019). Similarly, the *New York Times* discussed the way groups of primarily young white men will build systems that focus on their data, and give wrong or discriminatory answers in more general situations (Tugend 2019). The MIT Media Lab hosts the Algorithmic Justice League, trying to stop organizations from building socially slanted systems. Similar thoughts come from groups like the Data and Society Research Institute or the AI Now Institute.

Again, the problems may be accidental or deliberate. The phrase "data poisoning" has been used to suggest malicious creation of training data or examples of data designed to deceive machine learning systems. There is now a DARPA research program, "Guaranteeing AI Robustness against Deception (GARD)," supporting research to learn how to stop trickery such as a demonstration of converting a traffic stop sign to a 45 mph speed limit with a few stickers (Eykholt et al. 2018). More generally, bias in systems deciding whether to grant loans may be discriminatory but nevertheless profitable.

Even if you want to detect AI mistakes, recognizing such problems is difficult. Often things will be wrong and we won't know why. And even hypothetical (but perhaps erroneous) explanations can be very convincing; people easily believe plausible stories. I routinely give my students a paper that concludes that prior ownership of a cat prevents fatal myocardial infarctions; its result implies that cats are more protective than statin drugs (Qureshi et al. 2009). The students are very quick to come up with possibilities like "petting a cat is relaxing, relaxation reduces your blood pressure, and lower blood pressure decreases the risk of heart attacks." Then I have to explain that the paper evaluates 32 possibilities (prior/current ownership $\times$ cats/dogs $\times$ 4 medical conditions $\times$ fatal/nonfatal) and you shouldn't be surprised if you evaluate 32 chances and one is significant at the 0.05 level, which is only 1 in 20. In this example, there is also the question of reverse causality: perhaps someone who is in ill health will decide he is too sick to take care of a

Figure 9.2: Panoramic landscape.

pet, so that the poor health is not caused by the lack of a cat, but rather the poor health causes the absence of a cat.

Sometimes explanations can help, as in a machine learning program that was deliberately trained to distinguish images of wolves and dogs but was trained using pictures of wolves that always contained snow and pictures of dogs that never did (Ribeiro, Singh, and Guestrin 2016). Without explaining that, 10 of 27 subjects thought the classifier was trustworthy; after pointing out the snow only 3 of 27 subjects believed the system. Usually you don't get such a clear presentation of a mis-trained system.

Recognition of problems

Can we tell when something is wrong? Here's the result of a Google Photo merge of three other photos; two landscapes and a picture of somebody's friend. The software was told to make a panorama and stitched the images together (Peng 2018). It looks like a joke, and even made it into a list of top jokes on reddit. The author's point was that the panorama system didn't understand basic composition: people are not the same scale as mountains.

Often, machine learning results are overstated. Google Flu Trends was acclaimed for several years and then turned out to be undependable (Lazer et al. 2014). A study that attempted to compare the performance of machine learning systems for medical diagnosis with actual doctors found that of over 20,000 papers analyzed, only a few dozen had data suitable for an evaluation (Liu et al. 2019). The results claimed comparable accuracy, but virtually none of the papers

presented adequate data to support that conclusion.

Unusually promising results are sometimes the result of overfitting (Brownlee 2018); this is what was wrong with Google Flu Trends. A machine learning program can learn a large number of special cases and then find that the results do not generalize. In other cases problems can result when using "clean" data for training, and then encountering messier data in applications. Ideally, training and testing data should be from the same dataset and divided at random, but it can be tempting to start off with examples that are the result of initial and higher quality data collection.

Sometimes in the past we had a choice between modeling and data for predictions. Consider, for example, the problem of guessing what the weather will be tomorrow. We now do this based on a model of the atmosphere that uses the Navier-Stokes equations; we use supercomputers and derive tomorrow's atmosphere from today's (Christensen 2015). What did we do before we had supercomputers? Solving those equations by hand is impractical. One of the methods was "prediction by analogy": find some day in the past whose weather was most similar to today. Suppose that day is Oct. 20, 1970. Then use October 21, 1970 as tomorrow's prediction. Prediction by analogy doesn't require you to have a model or use advanced mathematics. In this case, however, it doesn't work as well—partly because we don't have enough past days to choose from, and we only get new days at the rate of one per day.

In fact, Huug van den Dool estimated the number of days of data needed to make accurate predictions as 10^{30} years, which is far more than the age of the universe (Wilks 2008). The underlying problem is that the weather is very random. If your state lottery is properly run, it should be completely pointless to look at past winning numbers and try to guess the next one. The weather is not that random but it has too much variation to be solved easily by analogy. If your problem is very simple (tic-tac-toe) you could indeed write down each position and what the best next move is; there are only about 255,000 games.

To deal with more realistic problems, much of machine learning research is now focused on obtaining larger training sets. Instead of trying to learn more about the characteristics of a system that is being modeled, the effort is driven by the dictum, "more data beats better algorithms." In a review of the history of speech recognition, Xuedong Huang, James Baker, and Raj Reddy write, "The power of these systems arises mainly from their ability to collect, process, and learn from very large datasets. The basic learning and decoding algorithms have not changed substantially in 40 years" (2014). Nevertheless, speech recognition has gone from frustration to useful products such as dictation software or home appliances.

Lacking a model, however, means that we won't know the limits of the calculations being done. For example, if you have some data that looks quadratic, but you fit a linear model, any attempt at extrapolation is fraught with error. If you are using a "black box" system, you don't know when this is happening. And, regrettably, many of the AI software systems are sold as black boxes where the purchasers and users do not have access to the process, even if they are imagined to be able to understand it.

What's changing

Many AI researchers are sensitive to the risks, especially given the publicity over self-driving cars. As the hype over "deep learning" built up, writers discussed examples such as a Pittsburgh medical system that proposed to send patients with both pneumonia and asthma home, because the computer had not understood that patients with both problems were actually being sent to the ICU (Bornstein 2016; Caruana et al. 2015).

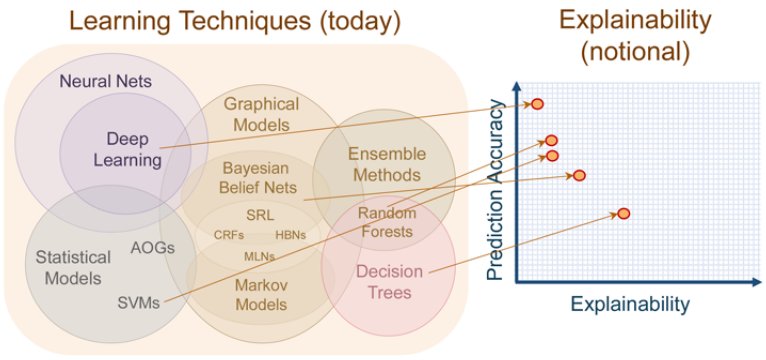

Figure 9.3: Explainability.

Many people work on ways of explaining or presenting neural net software (Harley 2015). Most important, perhaps, are new EU regulations that prohibit automated decision making that affects EU citizens, and provides a "right of explanation" (Metz 2016).

We recognize that systems which don't rely on a mathematical model may be cheaper to build than one where the coders understand what is going on. More serious is that they may be more accurate. This image is from the same article on understandability (Bornstein 2016).

If there really is a tradeoff between what will solve the problem and what can be explained, we know that many system builders will choose to solve the problem. And yet even having explanations may not be an answer; a key paper on interpretability discusses the complexities of meaning related to explanation, causality, and modeling (Lipton 2018).

Arend Hintze has noted that we do not always impose a demand for explanation on people. I can write that the New York Public Library main building is well proportioned and attractive without anyone expecting that I will recite its dimensions or the source of the marble used to construct it. And for some problems that's fine: I don't care how my camera decides on the focus distance to the subject. Where it matters, however, we often want explanations; the hard ethical problem, as noted before, is if better performance can be achieved in an inexplicable way.

Recommendations

2017 saw the publication of the "Asilomar AI principles" (2017). Two of these principles are:

- **Safety:** AI systems should be safe and secure throughout their operational lifetime, and verifiably so where applicable and feasible.
- **Failure Transparency:** If an AI system causes harm, it should be possible to ascertain why.

The problem is that the technology used to build many systems does not enable verifiability and explanation. Similarly the World Economic Forum calls for protection against discrimination but notes many ways in which technology can have unanticipated and undesirable effects as a result of machine learning ("How to Prevent" 2018).

Historically there has been and continues to be too much hype. An important image recognition task is distinguishing malignant and benign spots on mammograms. There have been promises for decades that computers would do this better than radiologists. Here are examples from 1995 ("computer-aided diagnosis can improve radiologists' observational performance") (Schmidt and Nishikawa) and 2009 ("The Bayesian network significantly exceeded the performance of interpreting radiologists") (Burnside et al.). A typical recent AI paper to do this with convolutional neural nets reports 90% accuracy (Singh et al. 2020). To put this in perspective, the problem is complex, but some examples are more straightforward, and even pigeons can reach 85% (Levenson et al. 2015). A serious recent review is "Diagnostic Accuracy of Digital Screening Mammography With and Without Computer-Aided Detection" (Lehman et al. 2015). Very recently there was another claim that computers have surpassed radiologists (Walsh 2020); we will have to await evaluation. As with many claims of medical progress, replicability and evaluation are needed before doctors will be willing to believe them.

What should we do? Software testing generally is a decades-old discipline, and many basic principles of regression testing apply here also:

- Test data should cover the full range of expected input.
- Test data should also cover unexpected and even illegal input.
- Test data should include known past failures believed cleared up.
- Test data should exercise all parts of the program, and all important paths (coverage).
- Test data should include a set of data which is representative of the distribution of actual data, to be used for timing purposes.

It is difficult to apply these ideas in parts of the AI world. If the allowed input is speech, there is no exhaustive list of utterances which can be sampled. If a black-box commercial machine learning package is being used, there is no way to ask about coverage of any number of test cases. If a program is constantly learning from new data, there is no list of previously fixed failures to be collected that reflects the constantly changing program.

And obviously the circumstances of use matter. We may well, as a society, decide that forcing banks evaluating loan applications to use decision trees instead of deep learning is appropriate, so that we know whether illegal discrimination is going on, even if this raises the costs to the banks. We might also believe that the safest possible railway operation is important, even if the automated train doesn't routinely explain how it balanced its choices of acceleration to achieve high punctuality and low risk.

What would I suggest?

Organizationally:

- Have teams including both the computer scientists and the users.
- Collaborate with a statistician: they've seen a lot of these problems before.
- Work on easier problems.

As examples, I watched a group of zoologists with a group of computer scientists discussing how to improve accuracy at identifying animals in photographs. The discussion indicated that

you needed hundreds of training examples at a minimum, if not thousands, since the animals do not typically walk up to the camera and pose for a full-frame shot. It was important to have both the people who understood the learning systems and the people who knew what the pictures were realistically like. The most amusing contribution by a statistician happened when a computer scientist offered a program that tried to recognize individual giraffes, and a zoologist complained that it only worked if you had a view of the right-hand side of the giraffe. Somebody who knew statistics perked up and said "it's a 50% chance of recognizing the animal? I can do the math for that." And it is simpler to do "is there any animal in the picture?" before asking "which animal is it?" and create two easier problems.

Technically:

- Try to interpolate rather than extrapolate: use the algorithm on points "inside" the training set (thinking in multiple dimensions).
- Lean towards feature detection and modeling rather than completely unsupervised learning.
- Emphasize continuous rather than discrete variables.

I suggest using methods that involve feature detection, since that tells you what the algorithm is relying on. For example, consider the Google Flu Trends failure; the public was not told what terms were used. As David Lazer noted, some of them were just "winter" terms (like 'basketball'). If you know that, you might be skeptical. More significant are decisions like jail sentences or college admissions; knowing that racial or religious discrimination are not relevant can be verified by knowing that the program did not use them. Knowing what features were used can sometimes help the user: if you know that your loan application was downrated because of your credit score, it may be possible for you to pay off some bill to raise the score.

Sometimes you have to use categorical variables (what county do you live in?) but if you have a choice of how you phrase a variable, asking something like "how many minutes a day do you spend reading?" is likely to produce a better fit than asking people to choose "how much do you read: never, sometimes, a lot?" A machine learning algorithm may tell you how much of the variance each input variable explains; you can use that information to focus on the variables that are most important to your problem, and decide whether you think you are measuring them well enough.

Why not extrapolate? Sadly, as I write this in early April 2020, we are seeing all sorts of extrapolations of the COVID-19 epidemic, with expected US deaths ranging from 30,000 to 2 million, as people try to fit various functions (Gaussians, logistic regression, or whatever) with inadequately precise data and uncertain models. A simpler example is Mark Twain's: "In the space of one hundred and seventy-six years the Lower Mississippi has shortened itself two hundred and forty-two miles. That is an average of a trifle over one mile and a third per year. Therefore, any calm person, who is not blind or idiotic, can see that in the 'Old Oolitic Silurian Period,' just a million years ago next November, the Lower Mississippi River was upwards of one million three hundred thousand miles long, and stuck out over the Gulf of Mexico like a fishing-rod. And by the same token any person can see that seven hundred and forty-two years from now the Lower Mississippi will be only a mile and three-quarters long, and Cairo and New Orleans will have joined their streets together, and be plodding comfortably along under a single mayor and a mutual board of aldermen" (1883).

Finally, note the advice of Edgar Allan Poe: "Believe nothing you hear, and only one half that you see."

References

"AI Recognition Fooled by Single Pixel Change." *BBC News*, November 3, 2017. https://www.bbc.com/news/technology-41845878.

"Asilomar AI Principles." 2017. https://futureoflife.org/ai-principles/.

Bojarski, Mariusz, Larry Jackel, Ben Firner, and Urs Muller. 2017. "Explaining How End-to-End Deep Learning Steers a Self-Driving Car." NVIDIA Developer Blog. https://devblogs.nvidia.com/explaining-deep-learning-self-driving-car/.

Bornstein, Aaron. 2016. "Is Artificial Intelligence Permanently Inscrutable?" *Nautilus* 40 (1). http://nautil.us/issue/40/learning/is-artificial-intelligence-permanently-inscrutable.

Brownlee, Jason. 2018. "The Model Performance Mismatch Problem (and What to Do about It)." Machine Learning Mastery. https://machinelearningmastery.com/the-model-performance-mismatch-problem/.

Burnside, Elizabeth S., Jessie Davis, Jagpreet Chhatwal, Oguzhan Alagoz, Mary J. Lindstrom, Berta M. Geller, Benjamin Littenberg, Katherine A. Shaffer, Charles E. Kahn, and C. David Page. 2009. "Probabilistic Computer Model Developed from Clinical Data in National Mammography Database Format to Classify Mammographic Findings." *Radiology* 251 (3): 663–72.

Caruana, Rich, Yin Lou, Johannes Gehrke, Paul Koch, Marc Sturm, and Noemie Elhadad. 2015. "Intelligible Models for HealthCare: Predicting Pneumonia Risk and Hospital 30-day Readmission." In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (KDD '15), 1721–30. New York: ACM Press. https://doi.org/10.1145/2783258.2788613.

Christensen, Hannah. 2015. "Banking on better forecasts: the new maths of weather prediction." *The Guardian,* 8 Jan 2015. https://www.theguardian.com/science/alexs-adventures-in-numberland/2015/jan/08/banking-forecasts-maths-weather-prediction-stochastic-processes.

Eykholt, Kevin, Ivan Evtimov, Earlence Fernandes, Bo Li, Amir Rahmati, Florian Tramèr, Atul Prakash, Tadayoshi Kohno, and Dawn Song. 2018. "Physical Adversarial Examples for Object Detectors." 12th USENIX Workshop on Offensive Technologies (WOOT 18).

Guidiotti, Riccardo, Anna Monreale, Salvatore Ruggieri, Franco Turini, Giannotti Fosca, and Dino Pedreschi. 2018. "A Survey of Methods for Explaining Black Box Models." *ACM Computing Surveys* 51 (5): 1–42.

Halevy, Alon, Peter Norvig, and Fernando Pereira. 2009. "The Unreasonable Effectiveness of Data." *IEEE Intelligent Systems* 24 (2).

Harley, Adam W. 2015. "An Interactive Node-Link Visualization of Convolutional Neural Networks." In *Advances in Visual Computing*, edited by George Bebis et al., 867–77. Lecture Notes in Computer Science. Cham: Springer International Publishing.

"How to Prevent Discriminatory Outcomes in Machine Learning." 2018. White Paper from the Global Future Council on Human Rights 2016–2018, World Economic Forum. https://www.weforum.org/whitepapers/how-to-prevent-discriminatory-outcomes-in-machine-learning.

Huang, Xuedong, James Baker, and Raj Reddy. 2014. "A Historical Perspective of Speech Recognition." *Communications of the ACM* 57 (1): 94–103.

Lazer, David, Ryan Kennedy, Gary King, and Alessandro Vespignani. 2014. "The Parable of Google Flu: Traps in Big Data Analysis." *Science* 343 (6176): 1203–1205.

Lehman, Constance, Robert Wellman, Diana Buist, Karl Kerlikowske, Anna Tosteson, and Diana Miglioretti. 2015. "Diagnostic Accuracy of Digital Screening Mammography with and without Computer-Aided Detection." *JAMA Intern Med* 175 (11): 1828–1837.

Levenson, Richard M., Elizabeth A. Krupinski, Victor M. Navarro, and Edward A. Wasserman. 2015. "Pigeons (*Columba livia*) as Trainable Observers of Pathology and Radiology Breast Cancer Images." *PLoS One,* November 18, 2015. https://doi.org/10.1371/journal.pone.0141357.

Lipton, Zachary. 2018. "The Mythos of Model Interpretability." *ACM Queue* 61 (10): 36–43.

Liu, Xiaoxuan et al. 2019. "A Comparison of Deep Learning Performance against Health-Care Professionals in Detecting Diseases from Medical Imaging: a Systematic Review and Meta-Analysis." *Lancet Digital Health* 1 (6): e271–97. https://www.sciencedirect.com/science/article/pii/S2589750019301232.

Mahjoubi, Mounir. 2018. "Assemblée nationale, XV^{e} législature. Session ordinaire de 2017–2018." Compte rendu intégral, Deuxième séance du mercredi 07 février 2018. http://www.assemblee-nationale.fr/15/cri/2017-2018/20180137.asp.

Metz, Cade. 2016. "Artificial Intelligence Is Setting Up the Internet for a Huge Clash with Europe." *Wired,* July 11, 2016. https://www.wired.com/2016/07/artificial-intelligence-setting-internet-huge-clash-europe/.

O'Neil, Cathy. 2016. *Weapons of Math Destruction.* New York: Crown.

Peng, Tony. 2018. "2018 in review: 10 AI failures." *Medium,* December 10, 2018. https://medium.com/syncedreview/2018-in-review-10-ai-failures-c18faadf5983.

Qureshi, A. I., M. Z. Memon, G. Vazquez, and M. F. Suri. 2009. "Cat ownership and the Risk of Fatal Cardiovascular Diseases. Results from the Second National Health and Nutrition Examination Study Mortality Follow-up Study." *Journal of Vascular and Interventional Neurology* 2 (1): 132–5. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3317329.

Ribeiro, Marco Tulio, Sameer Singh, and Carlos Guestrin. 2016. " 'Why Should I Trust You?': Explaining the Predictions of Any Classifier." In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (KDD '16), 1135–1144. New York: ACM Press.

Schmidt, R. A. and R. M. Nishikawa. 1995. "Clinical Use of Digital Mammography: the Present and the Prospects." *Journal of Digital Imaging* 8 (1 Suppl 1): 74–9.

Singh, Vivek Kumar et al. 2020. "Breast Tumor Segmentation and Shape Classification in Mammograms Using Generative Adversarial and Convolutional Neural Network." *Expert Systems with Applications* 139.

Su, Jiawei, Danilo Vasconcellos Vargas, and Kouichi Sakurai. 2019. "One Pixel Attack for Fooling Deep Neural Networks." *IEEE Transactions on Evolutionary Computation* 23 (5): 828–841.

Taulli, Tom. 2019. "How Bias Distorts AI (Artificial Intelligence)." *Forbes*, August 4, 2019. https://www.forbes.com/sites/tomtaulli/2019/08/04/bias-the-silent-killer-of-ai-artificial-intelligence/#1cc6f35d7d87.

Twain, Mark. 1883. *Life on the Mississippi.* Boston: J. R. Osgood & Co.

Tugend, Alina. 2019. "The Bias Embedded in Tech." *The New York Times,* June 17, 2019, section F, 10.

Walsh, Fergus. 2020. "AI 'outperforms' doctors diagnosing breast cancer." *BBC News*, January 2, 2020. https://www.bbc.com/news/health-50857759.

Wilks, Daniel S. 2008. Review of *Empirical Methods in Short-Term Climate Prediction*, by Huug van den Dool. *Bulletin of the American Meteorological Society* 89 (6): 887–88.