Harvard Data Science Review • Issue 2.1, Winter 2020

Should We Trust Algorithms?

David Spiegelhalter^{1,2}

¹Statistics Laboratory, Centre for Mathematical Sciences, University of Cambridge, Cambridge, England, United Kingdom,

²Centre for Science and Policy, University of Cambridge, Cambridge, England, United Kingdom

Published on: Jan 31, 2020

DOI: <https://doi.org/10.1162/99608f92.cb91a35a>

License: [Creative Commons Attribution 4.0 International License \(CC-BY 4.0\)](https://creativecommons.org/licenses/by/4.0/)

ABSTRACT

There is increasing use of algorithms in the health care and criminal justice systems, and corresponding increased concern with their ethical use. But perhaps a more basic issue is whether we should believe what we hear about them and what the algorithm tells us. It is illuminating to distinguish between the trustworthiness of claims made *about* an algorithm, and those made *by* an algorithm, which reveals the potential contribution of statistical science to both evaluation and ‘intelligent transparency.’ In particular, a four-phase evaluation structure is proposed, parallel to that adopted for pharmaceuticals.

Keywords: trustworthiness, evaluation, AI, criminal justice system, medical AI, transparency

1. Introduction

When on holiday in Portugal last year, we came to rely on ‘Mrs. Google’ to give us driving directions in her awful Portuguese accent. When negotiating the narrow streets in the ancient university town of Coimbra, she confidently told us to go left and so we obeyed her. But we were somewhat taken aback when the road abruptly turned into steps—we stopped in time, but after that we were not so trusting.

But it’s not just navigation algorithms that need caution. Large numbers of algorithms¹ of varying complexity are being developed within the health care and the criminal justice system, and include, for example, the U.K. HART (Harm Assessment Risk Tool) system (Oswald, Grace, Urwin, & Barnes, 2018) for assessing recidivism risk, which is based on a machine-learning technique known as a random forest. But the reliability and fairness of such algorithms for policing are being strongly contested: apart from the debate about facial recognition (Simonite, 2019), a recent report by rights-organization Liberty (Holmes, 2019) on predictive policing algorithms says that “their use puts our rights at risk.”

It is important not to be mesmerized by the mystique surrounding artificial intelligence (AI). The media (and politicians) are replete with credulous stories about machine learning and AI, but these stories are often based on commercial claims (Brennen & Nielsen, 2019). In essence, these programs simply take some data and use rules or mathematical formulae to come up with a response that is intended to be used to enhance professional judgment. The idea of algorithms in criminal justice is not new: it is rarely acknowledged that simple scoring systems for recidivism based on a statistical regression analysis have been used for decades (Copas & Marshall, 1998). Indeed, basic sentencing guidelines can be considered as algorithms designed to produce consistency, and provide a starting point that can be adjusted according to the judge’s discretion about a specific case (Sentencing Council, 2019).

Nevertheless, the Liberty report is just one example of increasing ethical concern, and it can seem that there are now more people working on the ethics of algorithms, AI, and machine learning than on the technology itself.

There are numerous checklists and initiatives, for example, Algo-care for policing (Oswald et al., 2018), while FATML (Fairness, Accountability and Transparency in Machine Learning), suggests a social impact statement (Fairness, Accountability, and Transparency in Machine Learning, 2019) for any algorithm, detailing:

- *Responsibility*: whom to approach when things go wrong.
- *Explainability*: to stakeholders in nontechnical terms.
- *Accuracy*: to identify sources of error and uncertainty.
- *Auditability*: to allow third parties to check and criticize.
- *Fairness*: to different demographics.

Within criminal justice, the [COMPAS system](#) (equivant, 2019) is widely used in the United States for predicting recidivism and informing bail decisions. It takes in 137 items of information, and comes up with a risk score from 1 to 10, which is classified into low/medium/high. But the procedure is proprietary and so acts as a complete black box, while COMPAS has been accused of racial bias (Angwin, 2016), although this analysis has been strongly contested (Corbett-Davies, Pierson, Feller, & Goel, 2016). An appeal against its use failed (*Harvard Law Review*, 2017), but COMPAS appears to perform poorly on most of the FATML criteria.

So it all seems to come down to a simple question—can we trust algorithms?

2. Trust and Trustworthiness

In this age of misinformation and loud, competing voices, we all want to be trusted. But as the philosopher Onora O’Neill has said (O’Neill, 2013), organizations should not try to be trusted; rather they should aim to demonstrate *trustworthiness*, which requires honesty, competence, and reliability. This simple but powerful idea has been very influential: the revised Code of Practice for official statistics in the United Kingdom puts Trustworthiness as its first “pillar” (UK Statistics Authority, 2018).

It seems reasonable that, when confronted by an algorithm, we should expect trustworthy claims both:

1. *about the system*—what the developers say it can do, and how it has been evaluated, and
2. *by the system*—what it says about a specific case.

This is a complex topic, but statistical science can help—it has been contributing to communication and evaluation for decades. Let’s look at those two criteria in more detail.

2.1. The Trustworthiness of the Claims Made About the System

As documented in a recent report by the Reuters Institute (Brennen & Nielsen, 2019), there are many exaggerated claims about AI driven by commercial rather than scientific considerations. Eric Topol, in his authoritative review of medical AI, baldly asserts that “The state of AI hype has far exceeded the state of AI

science, especially when it pertains to validation and readiness for implementation in patient care” (Topol, 2019, p51).

The trustworthiness of claims about the overall system could be communicated by providing a social impact statement along the lines suggested by FATML. But there is one important consideration missing from that list. It seems taken for granted that algorithms will be beneficial when implemented and, since this is by no means assured, I would suggest adding:

- ***Impact: what are the benefits (and harms) in actual use?***

Statisticians have been familiar with structured evaluation for decades, ever since scandals such as the birth defects caused by thalidomide brought about a stringent testing regime for new pharmaceuticals. The established four-phase structure is summarized in Table 1, alongside a similar structure for algorithms based on longstanding similar proposals by me (Spiegelhalter, 1983) and Stead et al. (1994).

Table 1. Accepted phased evaluation structure for pharmaceuticals, with a proposed parallel structure for evaluation of algorithms.

	Pharmaceuticals	Algorithms
Phase 1	Safety: Initial testing on human subjects	Digital testing: Performance on test cases
Phase 2	Proof-of-concept: Estimating efficacy and optimal use on selected subjects	Laboratory testing: Comparison with humans, user testing
Phase 3	Randomized Controlled Trials: Comparison against existing treatment in clinical setting	Field testing: Controlled trials of impact
Phase 4	Post-marketing surveillance: For long-term side-effects	Routine use: Monitoring for problems

Nearly all attention in the published literature on both medical and policing algorithms has focused on Phase 1—claimed accuracy on digital data sets. But this is only the start of the evaluation process. There is a small but increasing number of Phase 2 evaluations in which performance is compared with human ‘experts,’ sometimes in the form of a Turing Test in which the quality of the judgments of both humans and algorithms are assessed by independent experts, who are blinded as to whether the judgment was made by a human or algorithm (Turing, 1950). For example, the medical AI company Babylon (Copestake, 2018) conducted a Phase 2 study comparing their diagnostic system with doctors, although this was subsequently strongly criticized in the *Lancet* (Fraser, Coiera, & Wong, 2018). Kleinberg, Lakkaraju, Leskovec, Ludwig, & Mullainathan (2018) also draw

the analogy between evaluating recidivism algorithms and the four-phase pharmaceutical structure, and model a Phase 2 comparison between human and algorithmic decisions.

Topol also reports that “There has been remarkably little prospective validation for tasks that machines could perform to help clinicians or predict clinical outcomes that would be useful for health systems” (Topol, 2019, p52). This means there have been very few Phase 3 evaluations that check whether a system in practice actually does more good than harm: even simple risk-scoring systems have rarely been evaluated in randomized trials, although a Cochrane Review of randomized trials (Karmali et al., 2017, p2) of risk scoring for the primary prevention of cardiovascular disease concluded that “providing CVD risk scores may slightly reduce CVD risk factor levels and may increase preventive medication prescribing in higher-risk people without evidence of harm”.

Algorithms may have an impact through an unexpected mechanism. I was involved in a study of ‘computer-aided diagnosis’ back in the 1980s, when this meant having a large and clumsy personal computer in the corner of the clinic. In a randomized trial we showed that even a rather poor algorithm could improve clinical performance in diagnosing and treating acute abdominal pain—not because the doctors took much notice of what the computer said, but simply by encouraging them to systematically collect a good history and make an initial diagnosis (Wellwood, Johannessen, & Spiegelhalter, 1992).

There are, however, limitations to the analogy with evaluating pharmaceuticals. Prescription drugs act on individuals and, with the notable exceptions of overuse of antidepressants and opioids, rarely have an impact on society in general. In contrast, widespread use of an algorithm has the potential to have such an impact, and therefore the traditional individual-based randomized controlled trial may need supplementing by evaluation of the effect on populations. The UK Medical Research Council’s structure for the evaluation of complex medical interventions may be relevant; the original (very highly cited) proposal closely followed the staged pharmaceutical model described above (Campbell et al., 2000), but a revised version moved to a more iterative model with a reduced emphasis on experimental methods (Craig et al., 2008), and a further forthcoming update promises to broaden its perspective to other disciplines and further downplay randomized controlled trials (Skivington, Matthews, Craig, Simpson, & Moore, 2018).

An important consideration is that clinical algorithms are considered as medical devices for regulatory purposes, say by the European Union (Fraser et al., 2018) or the Food and Drug Administration (FDA) (Center for Devices and Radiological Health, 2019), and hence are not subject to the four-phase structure for pharmaceuticals shown in Table 1. Phase 3 randomized trials of impact are therefore not required for approval, with a strong emphasis played on the reliability of the technology or code itself. Again, this presupposes that algorithms shown to have reasonable accuracy in the laboratory must help in practice, and explicit evidence for this would improve the trustworthiness of the claims made about the system.

3. Trustworthiness of Claims Made by the System to the Recipients of Its Advice

When an individual is subject to an algorithm's claim, say, an assessment of the risk of recidivism or a medical diagnosis, it seems reasonable that they or their representatives should be able to get clear answers to questions such as:

- Is the current case within its competence?
- What was the chain of reasoning that drove this claim?
- What if the inputs had been different (counterfactuals)?
- Was there an important item of information that 'tipped the balance'?
- What is the uncertainty surrounding the claim?

There are many current ingenious attempts to make complex algorithms more explainable and less of a black box. For example, Google DeepMind's eye diagnosis system developed with Moorfields Eye Hospital is based on a deep-learning algorithm, but one that has been deliberately structured in layers to help visually explain intermediate steps between the raw image and diagnosis and triage recommendation (De Fauw et al., 2018).

While a deep-learning algorithm may be appropriate for automatic analysis of image data, when there are fewer inputs it may be possible to build a simpler, more interpretable model in the first place. Statistical science has mainly focused on linear regression models in which, essentially, features are weighted to lead to a scoring system, for example, Caruana and colleagues' work using generalized additive models to produce pneumonia risk scores (Caruana et al., 2015). It's often said that increased interpretability has to be traded off against performance, but this has been questioned for recidivism algorithms (Rudin, 2018). Indeed an online experiment showed that the untrained public were as good as COMPAS (65% accuracy), and that COMPAS performance could be matched by a simple rule-based classifier (Angelino, Larus-Stone, Alabi, Seltzer, & Rudin, 2017), and even a regression model with only two predictors (age and total previous convictions) (Dressel & Farid, 2018). Furthermore, assessments of uncertainty are a core component of statistical science.

4. Transparency

Trustworthiness demands transparency, but not just 'fishbowl' transparency in which huge amounts of information are provided in indigestible form. Transparency does not necessarily provide explainability—if systems are very complex, even providing code will not be illuminating. Fortunately, Onora O'Neill has again made a major contribution in developing the idea of "intelligent transparency" (Royal Society, 2012), in which she argues that information should be

- *accessible: interested people should be able to find it easily.*
- *intelligible: they should be able to understand it.*
- *useable: it should address their concerns.*

- *assessable: if requested, the basis for any claims should be available.*

I feel the final criterion is essential: a trustworthy algorithm should be able to ‘show it’s working’ to those who want to understand how it came to its conclusions. While most users may be happy to take the algorithm’s claims ‘on trust,’ interested parties should be able to assess the reliability of such claims. In an experimental study of how much an algorithm adds to human accuracy, Lai & Tan (2019) found that providing an individualized explanation added as much as providing a generic assurance about the algorithm’s quality.

We have tried to live up to these aims in the interface we have constructed for the popular Predict program for women with newly diagnosed breast cancer (National Health Service, 2019), in which details of the disease and possible treatments are entered, and predictive information about the potential benefits and harms of post-surgical treatment is then communicated in text, numbers, and graphics. Explanation is provided at multiple levels and in multiple formats, and the full details of the algorithm, and even the code, are available for scrutiny if required. Of course, one problem of having a reliable algorithm that is packaged in a transparent and attractive manner is that it can lead to ‘overtrust,’ in which the output is treated as being precise and unquestionable. A truly trustworthy algorithm should be able to communicate its own limitations to ensure, rather ironically, that it is not trusted too much.

5. Conclusions

Developers need to demonstrate the trustworthiness of claims both *about* and *by* an algorithm, which requires phased evaluation of quality and impact based on strong statistical principles. In the context of clinical algorithms, Topol says “it requires rigorous studies, publication of the results in peer-reviewed journals, and clinical validation in a real-world environment, before roll-out and implementation” (Topol, 2019, p52). The same needs to be applied in the criminal justice system, where there is no FDA to license applications.

Finally, whenever I hear claims about any algorithm, my shortlist of questions I would like to ask include:

1. Is it any good when tried in new parts of the real world?
2. Would something simpler, and more transparent and robust, be just as good?
3. Could I explain how it works (in general) to anyone who is interested?
4. Could I explain to an individual how it reached its conclusion in their particular case?
5. Does it know when it is on shaky ground, and can it acknowledge uncertainty?
6. Do people use it appropriately, with the right level of skepticism?
7. Does it actually help in practice?

I feel that question 5 is particularly important. Being confidently told to drive down a set of steps reduced my trust in Mrs. Google, but on another occasion, she simply gave up and said “I cannot help you at the moment.” She soon recovered her composure, but to me this seemed to be trustworthy behavior—the algorithm knew when it didn’t know, and told us so. Such humility is rare and to be prized.

Disclosure Statement

David Spiegelhalter has no financial or non-financial disclosures to share for this article.

References

- Angelino, E., Larus-Stone, N., Alabi, D., Seltzer, M., & Rudin, C. (2017). Learning certifiably optimal rule lists. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 35–44). <https://doi.org/10.1145/3097983.3098047>
- Angelino, E., Larus-Stone, N., Alabi, D., Seltzer, M., & Rudin, C. (2017). Learning certifiably optimal rule lists. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 35–44). <https://doi.org/10.1145/3097983.3098047>
- Angwin, J. (2016, May 23). Machine bias. *ProPublica*. Retrieved August 8, 2019, from <https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing>
- Brennen, S., & Nielsen, R. (2019). An Industry-led debate: How UK media cover artificial intelligence. Retrieved August 8, 2019, from <https://reutersinstitute.politics.ox.ac.uk/our-research/industry-led-debate-how-uk-media-cover-artificial-intelligence>
- Campbell, M., Fitzpatrick, R., Haines, A., Kinmonth, A. L., Sandercock, P., Spiegelhalter, D., & Tyrer, P. (2000). Framework for design and evaluation of complex interventions to improve health. *BMJ*, 321, 694–696. <https://doi.org/10.1136/bmj.321.7262.694>
- Caruana, R., Lou, Y., Gehrke, J., Koch, P., Sturm, M., & Elhadad, N. (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* (pp. 1721–1730). <https://doi.org/10.1145/2783258.2788613>
- Center for Devices and Radiological Health. (2019). Artificial intelligence and machine learning in software as a medical device. Retrieved from <http://www.fda.gov/medical-devices/software-medical-device-samd/artificial-intelligence-and-machine-learning-software-medical-device>
- Copas, J., & Marshall, P. (1998). The offender group reconviction scale: A statistical reconviction score for use by probation officers. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 47(1), 159–171. <https://doi.org/10.1111/1467-9876.00104>
- Copestake, J. (2018, June 27). Chatbot claims to beat GPs at medical exam. *BBC*. <https://www.bbc.com/news/technology-44635134>

Corbett-Davies, S., Pierson, E., Feller, A., & Goel, S. (2016). A computer program used for bail and sentencing decisions was labeled biased against blacks. It's actually not that clear. *The Washington Post*. Retrieved October 28, 2019, from <https://www.washingtonpost.com/news/monkey-cage/wp/2016/10/17/can-an-algorithm-be-racist-our-analysis-is-more-cautious-than-propublicas/>

Craig, P., Dieppe, P., Macintyre, S., Michie, S., Nazareth, I., & Petticrew, M. (2008). Developing and evaluating complex interventions: The new Medical Research Council guidance. *BMJ*, 337(7676), Article a1655. <https://doi.org/10.1136/bmj.a1655>

De Fauw, J. D., Ledsam, J. R., Romera-Paredes, B., Nikolov, S., Tomasev, N., Blackwell, S., ... Ronneberger, O. (2018). Clinically applicable deep learning for diagnosis and referral in retinal disease. *Nature Medicine*, 24(9), 1342–1350. <https://doi.org/10.1038/s41591-018-0107-6>

Dressel, J., & Farid, H. (2018). The accuracy, fairness, and limits of predicting recidivism. *Science Advances*, 4(1), Article eaao5580. <https://doi.org/10.1126/sciadv.aao5580>

equivant. (2019). *The Northpointe Suite*. Retrieved August 8, 2019, from <https://www.equivant.com/northpointe-suite/>

Fairness, Accountability, and Transparency in Machine Learning. (2019). *Principles for accountable algorithms and a social impact statement for algorithms*. Retrieved August 8, 2019, from <https://www.fatml.org/resources/principles-for-accountable-algorithms>

Fraser, A. G., Butchart, E. G., Szymański, P., Caiani, E. G., Crosby, S., Kearney, P., & de Werf, F. V. (2018). The need for transparency of clinical evidence for medical devices in Europe. *The Lancet*, 392(10146), 521–530. [https://doi.org/10.1016/S0140-6736\(18\)31270-4](https://doi.org/10.1016/S0140-6736(18)31270-4)

Fraser, H., Coiera, E., & Wong, D. (2018). Safety of patient-facing digital symptom checkers. *The Lancet*, 392(10161), 2263–2264. [https://doi.org/10.1016/S0140-6736\(18\)32819-8](https://doi.org/10.1016/S0140-6736(18)32819-8)

Harvard Law Review (2019) *State v Loomis*. *Harvard Law Review*, 130(5). <https://harvardlawreview.org/2017/03/state-v-loomis/>

Holmes, L. (2019, February 1). *Report: Policing by machine*. Retrieved August 8, 2019, from Liberty Human Rights website: <https://www.libertyhumanrights.org.uk/policy/report-policing-machine>

Karmali, K. N., Persell, S. D., Perel, P., Lloyd-Jones, D. M., Berendsen, M. A., & Huffman, M. D. (2017). Risk scoring for the primary prevention of cardiovascular disease. *Cochrane Database of Systematic Reviews*, (3), Article CD006887. <https://doi.org/10.1002/14651858.CD006887.pub4>

- Kleinberg, J., Lakkaraju, H., Leskovec, J., Ludwig, J., & Mullainathan, S. (2018). Human decisions and machine predictions. *The Quarterly Journal of Economics*, 133(1), 237–293. <https://doi.org/10.1093/qje/qjx032>
- Lai, V., & Tan, C. (2019). On human predictions with explanations and predictions of machine learning models: A case study on deception detection. In *FAT* '19: Proceedings of the Conference on Fairness, Accountability, and Transparency* (pp. 29–38). <https://doi.org/10.1145/3287560.3287590>
- National Health Service. (2019). *Predict breast*. Retrieved August 8, 2019, from <https://breast.predict.nhs.uk/>
- O'Neill, O. (2013, June). *What we don't understand about trust* [Video]. TED Conferences. https://www.ted.com/talks/onora_o_neill_what_we_don_t_understand_about_trust/transcript?language=en
- Oswald, M., Grace, J., Urwin, S., & Barnes, G. C. (2018). Algorithmic risk assessment policing models: Lessons from the Durham HART model and “Experimental” proportionality. *Information & Communications Technology Law*, 27(2), 223–250. <https://doi.org/10.1080/13600834.2018.1458455>
- Royal Society. (2012). *Science as an open enterprise*. Retrieved May 17, 2017, from <https://royalsociety.org/topics-policy/projects/science-public-enterprise/report/>
- Rudin, C. (2018). Please stop explaining black box models for high stakes decisions. *arXiv*. <https://doi.org/10.48550/arXiv.1811.10154>
- Sentencing Council. (2019). Sentencing guidelines for use in Crown Court. Retrieved August 8, 2019, from <https://www.sentencingcouncil.org.uk/crown-court/>
- Simonite, T. (2019, August 6). Facial recognition is suddenly everywhere: Should you worry? *Wired*. Retrieved from <https://www.wired.com/story/facial-recognition-everywhere-should-you-worry/>
- Skivington, K., Matthews, L., Craig, P., Simpson, S., & Moore, L. (2018). Developing and evaluating complex interventions: Updating Medical Research Council guidance to take account of new methodological and theoretical approaches. *The Lancet*, 392, Article S2. [https://doi.org/10.1016/S0140-6736\(18\)32865-4](https://doi.org/10.1016/S0140-6736(18)32865-4)
- Spiegelhalter, D. J. (1983). Evaluation of clinical decision-aids, with an application to a system for dyspepsia. *Statistics in Medicine*, 2(2), 207–216. <https://doi.org/10.1002/sim.4780020215>
- Stead, W. W., Haynes, R. B., Fuller, S., Friedman, C. P., Travis, L. E., Beck, J. R., ... Abola, E. E. (1994). Designing medical informatics research and library--resource projects to increase what is learned. *Journal of the American Medical Informatics Association*, 1(1), 28–33. <https://doi.org/10.1136/jamia.1994.95236134>
- Topol, E. J. (2019). High-performance medicine: the convergence of human and artificial intelligence. *Nature Medicine*, 25(1), 44–56. <https://doi.org/10.1038/s41591-018-0300-7>

Turing, A. M. (1950). I.—Computing machinery and intelligence. *Mind*, *LIX*(236), 433–460.
<https://doi.org/10.1093/mind/LIX.236.433>

UK Statistics Authority. (2018). *Code of practice for statistics*. Retrieved July 8, 2019, from
<https://www.statisticsauthority.gov.uk/code-of-practice/the-code/>

Wellwood, J., Johannessen, S., & Spiegelhalter, D. J. (1992). How does computer-aided diagnosis improve the management of acute abdominal pain? *Annals of The Royal College of Surgeons of England*, *74*(1), 40–46.
<https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2497469/>

©2020 David Spiegelhalter. This article is licensed under a Creative Commons Attribution (CC BY 4.0) [International license](#), except where otherwise indicated with respect to particular material included in the article.

Footnotes

1. I am using ‘algorithm’ in a fairly general sense to mean the actual formula, coded model, or even guideline, that is to be used in a new situation. I am not referring to the way in which this procedure was built, whether using machine-learning, standard statistical analysis, judgment, or other methods. [↩](#)