The dangers of faulty, biased, or malicious algorithms requires independent oversight

Ben Shneiderman[a,1]

The recent crash of a driverless car sends a clear warning about how algorithms can be deadly (1). Similarly, there are clear dangers in vital national services, such as communications, financial trading, healthcare, and transportation. These services depend on sophisticated algorithms, some relying on unpredictable artificial intelligence techniques, such as deep learning, that are increasingly embedded in complex software systems (2–4). As search algorithms, high-speed trading, medical devices, and autonomous aircraft become more widely implemented, stronger checks become necessary to prevent failures (5, 6).

What might help are traditional forms of independent oversight that use knowledgeable people who have powerful tools to anticipate, monitor, and retrospectively review operations of vital national services. The three forms of independent oversight that have been used in the past by industry and governments—planning oversight, continuous monitoring by knowledgeable review boards using advanced software, and a retrospective analysis of disasters—provide guidance for responsible technology leaders and concerned policy makers (7). Considering all three forms of oversight could lead to policies that prevent inadequate designs, biased outcomes, or criminal actions.

There is a long history of analyses of how poor design, unintentional bias, and malicious interventions can cause algorithms to trigger huge financial losses, promote unfair decisions, violate laws, and even cause

Proper independent oversight and investigation of flawed algorithms can help anticipate and improve quality, hence avoiding failures that lead to disaster. Image courtesy of Shutterstock/joloei.

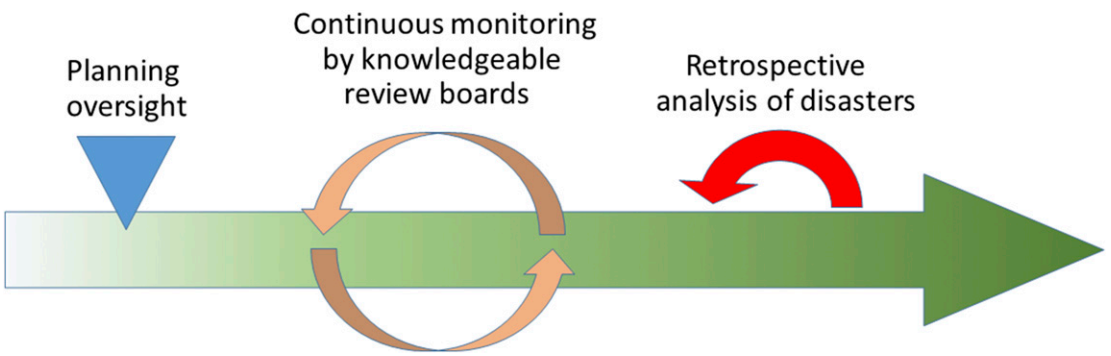
[a]Department of Computer Science, University of Maryland, College Park, MD 20742

Author contributions: B.S. wrote the paper.
The author declares no conflict of interest.

Any opinions, findings, conclusions, or recommendations expressed in this work are those of the author and have not been endorsed by the National Academy of Sciences.

[1]Email: ben@cs.umd.edu.

www.pnas.org/cgi/doi/10.1073/pnas.1618211113



Industry and governments have used three forms of independent oversight: planning oversight, continuous monitoring by knowledgeable review boards, and a retrospective analysis of disasters.

deaths (8). Helen Nissenbaum, a scholar who focuses on the impact of technology on culture and society, identified the sources of bugs and biases in software, complaining about the "systematic erosion of accountability" in computerized systems. She called for professional practices and legal changes to ensure "strict liability for defective consumer-oriented software" (9). Later, Friedman and Nissenbaum proposed a taxonomy of bias in computer systems that included preexisting biases based on social practices and attitudes, technical bias based on design constraints in hardware and software, and emergent bias that arises from changing the use context (10).

A Proactive Approach

Today's challenges are considerably more complex and require the implementation of a new system of checks to address. First is planning oversight. When major new or revised algorithmic systems are being developed, an independent oversight review could require implementers to submit an algorithms impact statement (11, 12). This document would be similar to the environmental impact statements that are now required for major construction programs. Standard questions about who the stakeholders are, and what the impacts might be, help ensure that implementers think carefully about potential problems and then propose reasonable solutions. Algorithm impact statements would document the goals of the program, data quality control for input sources, and expected outputs so that deviations can be detected. These algorithmic controls would act as surge protectors on electrical lines, which ensure that power spikes will not damage equipment.

Second is the continuous monitoring by knowledgeable review boards using advanced software. Vital systems might be under review by in-house monitors, just as Food and Drug Administration meat and pharmaceutical inspectors continuously check on production. This is expensive but has proven to be effective. Inspectors become familiar with production methods and communicate with peers so as to learn about problems elsewhere. Regular tests would help ensure stability of the algorithms, and help cope with problems that come when fresh training data has different distributions (13). Often the inspectors make helpful suggestions that increase safety, with the potential to lower costs and raise quality. Because inspectors eventually become too close to the system maintainers, regular rotation of inspectors is helpful to ensure continuing independence.

Third is the retrospective analysis of disasters. This might be carried out by a National Algorithms Safety Board, much like the National Transportation Safety Board, whose investigators fly in to study airplane, ship, and train accidents (www.ntsb.gov/Pages/default.aspx). Accident reviews often lead to improved designs that promote future safety. Algorithmic accidents are less visible than transportation crashes, so logging and monitoring techniques will have to be improved and widely applied. Like the National Transportation Safety Board, the National Algorithms Safety Board could be an independent board, outside of any government agency, with only power to investigate accidents and no authority to regulate. An industry-led voluntary approach could initiate a National Algorithms Safety Board to establish best practices for open analyses of disasters, so that all parties could learn from past failures (https://www.partnershiponai.org/tenets/).

As algorithms grow in importance and complexity, new software techniques will be needed to ensure that monitoring is built in, not added on.

Well-designed independent oversight boards constitute collected wisdom in specialized technologies. Still, they will have to work hard to earn the trust of algorithm developers, operators, and users who will be appropriately concerned about their liability. Effective independent oversight boards have sufficient legal power to investigate accidents and necessary knowledge to raise concerns. Members of a National Algorithms Safety Board will also need to be able to revisit their report recommendations to see that appropriate changes have been made.

Complex Problem, Creative Solutions

The detailed analyses of real systems by Friedman and Nissenbaum (10) led to two remedies: careful scrutiny to avoid preexisting biases, and thoughtful consideration of how diverse use contexts could introduce emergent biases. However, despite their worthy efforts, algorithms continue to fail, suggesting that stronger professional and legal changes are needed. Teaching students and professionals about ethical practices and professional codes of ethics is a key step (14). Another useful approach is to present well-documented case studies of how developers gain trust from operators and then engage with them to repair problems, while identifying possible improvements (15).

Newer strategies for reducing bias are being developed in the data mining community (16; see also www.fatml.org/resources) and proposals for machine-assisted oversight may also prove useful (17). The European Union is already moving strongly to ensure that algorithms will come with human-understandable explanations and that people will have a right to see a report on how the input variables triggered an unfavorable rating by an algorithm (18).

As algorithms grow in importance and complexity, new software techniques will be needed to ensure that monitoring is built in, not added on. In my view, technology advances are most effective when they are accompanied by increased clarity about responsibility for failures and liability for damages. Another productive shift would be to replace existing software contracts that limit liability through "hold harmless" provisions with clearer statements about the responsibility of developers, maintainers, and operators of algorithms (19). An ally in this difficult legal shift is likely to be the insurance industry, which has been an effective advocate of protective technologies in construction, transportation, and healthcare.

Careful logging of algorithm executions will also help. These detailed logs, such as those collected by aircraft flight data recorders, will enable National Algorithm Safety Board investigators to study exactly what happened. As best practices for logging in each industry become widely accepted, reviewers will be able to more reliably assign responsibility for failures while making compelling evidence-based recommendations for improvements (20).

There are many legitimate concerns about this proposal, such as who pays for it, which projects are big enough to warrant review, and how the independent oversight would mesh with existing review boards. Further discussion and study is warranted. But there's little doubt that because algorithms are increasingly vital to national economies, defense, and healthcare systems, some independent oversight will be helpful. Adding proactive technical, legal, and social mechanisms to support independent oversight processes will make them safer and better.

Acknowledgments

Thanks for thoughtful comments on drafts by my colleagues: Hal Daume, Nick Diakopoulos, Sorelle, Friedler, Simson Garfinkel, Jennifer Golbeck, Eric Horvitz, Jimmy Lin, Jennifer Preece, and Stuart Shapiro.

1 Singhvi A, Russell K (July 12, 2016) Inside the self-driving Tesla fatal accident, *The New York Times*. Available at www.nytimes.com/interactive/2016/07/01/business/inside-tesla-accident.html. Accessed August 1, 2016.
2 AI100 Standing Committee and Study Panel (2016) One hundred year study on artificial intelligence. Available at, https://ai100.stanford.edu/2016-report. Accessed November 10, 2016.
3 National Science and Technology Council, Committee on Technology (October 12, 2016) Preparing for the Future of Artificial Intelligence. Available at https://www.whitehouse.gov/sites/default/files/whitehouse_files/microsites/ostp/NSTC/preparing_for_the_future_of_ai.pdf. Accessed November 10, 2016.
4 National Science and Technology Council, Networking and Information Technology Research and Development Subcommittee (October 13, 2016) The National Artificial Intelligence Research and Development Strategic Plan. Available at https://www.whitehouse.gov/sites/default/files/whitehouse_files/microsites/ostp/NSTC/national_ai_rd_strategic_plan.pdf. Accessed November 10, 2016.
5 O'Neil C (2016) *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy* (Crown Publishers, New York).
6 Markoff J, Miller CC (June 16, 2014) As robotics advances, worries of killer robots rise, *The New York Times*. Available at _www.nytimes.com/2014/06/17/upshot/danger-robots-working.html. Accessed November 10, 2016.
7 National Academies Committee on Technical and Privacy Dimensions of Information for Terrorism Prevention and Other National Goals (2008) *Protecting Individual Privacy in the Struggle Against Terrorists: A Framework for Program Assessment* (National Academies Press, Washington, DC). Available at https://www.nap.edu/catalog.php?record_id=12452. Accessed October 14, 2016.
8 Leveson NG, Turner CS (1993) An investigation of the Therac-25 accidents. *IEEE Computer* 26(7):18–41.
9 Nissenbaum H (1994) Computing and accountability. *Commun ACM* 37(1):72–80.
10 Friedman B, Nissenbaum H (1996) Bias in computer systems. *ACM Trans Inf Syst* 14(3):330–347.
11 Diakopoulos N (2016) Accountability in algorithmic decision-making. *Commun ACM* 59(2):56–62.
12 Diakopoulos N, Koliska M (2016) Algorithmic transparency in the news media. *Digital Journalism*, 10.1080/21670811.2016.1208053.
13 Lazer D, Kennedy R, King G, Vespignani A (2014) Big data. The parable of Google Flu: Traps in big data analysis. *Science* 343(6176):1203–1205.
14 Don Gotterbarn KM, Rogerson S (1999) Software engineering code of ethics is approved. *Commun ACM* 42(10):102–107.
15 Lee JD, See KA (2004) Trust in automation: Designing for appropriate reliance. *Hum Factors* 46(1):50–80.
16 Feldman M, Friedler SA, Moeller J, Scheidegger C, Venkatasubramanian S (2015) Certifying and removing disparate impact, *Proceedings of the 21st ACM Conference on Knowledge Discovery and Data Mining* (Association for Computing Machinery, New York) pp 259–268.
17 Etzioni A, Etzioni O (2016) Designing AI systems that obey our laws and values. *Commun ACM* 59(9):29–31.
18 Goodman B, Flaxman S (2016) European Union regulations on algorithmic decision-making and a "right to explanation." Available at https://arxiv.org/pdf/1606.08813.pdf. Accessed November 10, 2016.
19 Pasquale F (2015) *The Black Box Society: The Secret Algorithms That Control Money and Information* (Harvard Univ Press, Cambridge, MA).
20 Shneiderman B (2007) Human responsibility for autonomous agents. *IEEE Intell Syst* 22(2):60–61.