This article has been accepted for publication in *BMJ Quality and Safety, 2019* following peer review, and the Version of Record can be accessed online at http://dx.doi.org/10.1136/bmjqs-2019-009484

Governing the safety of artificial intelligence in healthcare

Prof Carl Macrae, Nottingham University Business School
Centre for Health Innovation, Leadership and Learning, University of Nottingham
Jubilee Campus, Wollaton Road, Nottingham NG8 1BB
carlmacrae@mac.com @CarlMacrae

Artificial Intelligence (AI) has enormous potential to improve the safety of healthcare, from increasing diagnostic accuracy,[1] to optimising treatment planning,[2] to forecasting outcomes of care.[3] However, integrating AI technologies into the delivery of healthcare is likely to introduce a range of new risks and amplify existing ones. For instance, failures in widely used software have the potential to quickly affect large numbers of patients;[4] hidden assumptions in underlying data and models can lead to AI systems delivering dangerous recommendations that are insensitive to local care processes;[5,6] and opaque AI techniques such as deep learning can make explaining and learning from failure extremely difficult.[7,8] To maximise the benefits of AI in healthcare, and to build trust amongst patients and practitioners, it will therefore be essential to robustly govern the risks that AI poses to patient safety.

In a recent review in this journal, Challen and colleagues present an important and timely analysis of some of the key technological risks associated with the application of machine learning in clinical settings.[9] Machine learning is a subfield of AI that focuses on the development of algorithms that are automatically derived and optimised through exposure to large quantities of exemplar 'training' data.[10] The outputs of machine learning

algorithms are essentially classifications of patterns that provide some sort of prediction—for instance, predicting whether an image shows a malignant melanoma or a benign mole.[11] Some of the basic techniques of machine learning have existed for half a century or more, but progress in the field has accelerated rapidly due to advances in the development of 'deep' artificial neural networks[12] combined with huge increases in computational power and the availability of enormous quantities of data. These techniques have underpinned recent public demonstrations of AI systems that display superhuman performance in a range of games, such as Go and chess, that have generated a mix of hope, hype and horror about the transformative potential—and risks—posed by AI.[13] Away from the headlines, intensive efforts are underway to apply machine learning to a variety of clinical tasks, particularly those involving the analysis of medical scans and other images,[4] and health systems are actively seeking to harness the benefits of AI while also beginning to define principles of appropriate conduct.[14,15]

Given this rapid pace of development, Challen *et al* analyse how machine learning may pose risks to patient safety in the short, medium and long term.[9] In the short term, they identify safety issues as primarily arising from the reliability and interpretability of predictions made by machine learning systems. These issues include mismatches between the data a system is trained on and the world it must make predictions about (the problem of 'distributional shift'), and the challenges of understanding and explaining how machine learning systems make predictions (the 'black box' problem). In the medium term, key safety issues are determined to be human over-reliance on the outputs of machine learning systems (the problem of 'automation complacency'), along with the challenges of needing large quantities of historical data to train machine learning systems, which will therefore struggle to accommodate rapid changes in practice or policy. In the long term, Challen *et al* point to the potential implications of some fundamental technical challenges being grappled with in the field of AI safety. These include, for instance, autonomous systems that independently

discover novel ways to meet the letter, but not the spirit, of an intended objective (the 'reward hacking' problem), and in so doing cause unintended harm.[9]

Challen and colleagues' review of these technological sources of risk posed by machine learning, combined with accelerating efforts to implement AI, raises a range of urgent questions regarding the governance of AI safety in healthcare. Answering these questions will require the development of a broad analytical and research framework that includes but also extends far beyond the technical issues faced in the operation of specific algorithms. It will involve looking backward to the practical activities of defining, developing, testing and deploying the models and data that underlie AI systems. And it will involve looking forward to the organisational and institutional contexts in which AI tools will be embedded—and in which they will sometimes fail.

Governing the development of AI

Many of the key safety issues identified by Challen *et al* both arise from, and need to be addressed during, the initial design and development of AI systems. Activities such as defining requirements and objectives, collecting and cleaning data, training and testing models, and producing user-facing interfaces all involve difficult decisions, necessary trade-offs and fine-grained human judgements that can have considerable implications for safety.[10] Understanding these decisions, and the human and organisational processes that underpin them, is critically important to developing a complete picture of the safety of an AI system. For example, to properly govern the safety of a machine learning algorithm, it will be important to not only ascertain that a system 'errs on the side of caution' in its predictions.[9] It will also be important to understand how any asymmetric cost ratio that

underlies this behaviour was defined and agreed upon,[10] what evidence was drawn on to determine this, which stakeholders were involved and on what basis, and whether the balance between false positive and false negative errors is appropriate in relation to the work systems and safety controls in place in the setting in which the AI tool will be embedded. These are also, of course, the sorts of questions that will need to be answered in safety investigations when AI systems are involved in future adverse events.

Grappling with opacity and inscrutability

A more fundamental challenge to assuring the safety of machine learning systems highlighted by Challen *et al* is that of opacity: the basis of predictions made by some machine learning systems, particularly deep neural networks, are effectively an inscrutable 'black box', taking the form of a distribution of weights over a network rather than a logic that might (in principle) be explainable by—and to—humans. Technical solutions, such as saliency maps[9] and variable importance plots[10] may help in future. But while the precise workings of an algorithm itself may remain inscrutable, it should still be possible to examine and explain such things as the design decisions, functional requirements, training activities and input data that produce and constrain an AI system and its behaviour—as long as these are appropriately documented and auditable.[16] A relevant patient safety analogue might be found in current approaches to learning from unexpected adverse events in clinical care. To explain these, investigators do not (at present) seek to understand the precise neural basis of clinicians' actions. Rather, in an ideal world they aim to understand what information was available at key points, if there were any constraints on decision making or cognition, whether there were gaps in knowledge or training, how decisions and actions were supervised or checked, and a whole range of other sociotechnical factors.[17] Governing the safety of AI technologies in healthcare will require analytical and explanatory frameworks that can explain these broader sociotechnical processes, as much as the underlying

mechanics of the algorithms themselves. Despite that, irreducible technical inscrutability—along with the need for large quantities of retrospective training data best suited to a largely stable world[18]—may still render some approaches to machine learning unsuitable for safety-critical applications.[19]

Integrating AI into clinical systems

In addition to safety issues associated with the design and development of AI systems, Challen *et al* importantly draw attention to some risks associated with practical implementation. Ensuring that AI technologies 'fail safe' when they are unable to make reliable predictions is critical. However, 'failing safe' can not be fully understood as simply a technical property of an AI system declining to provide a prediction when its confidence is low,[9] but rather is a sociotechnical property of the entire work system that a technology is embedded in. For example, the introduction of AI technologies will likely need to be accompanied by targeted strategies to maintain the practical expertise of clinicians and avoid long-term deskilling in relation to tasks that become routinely performed by AI.[20] An AI tool that occasionally refrains from providing a prediction and hands a task back to a deskilled clinician is unlikely to constitute a safe clinical system. Other industries have long faced these sociotechnical challenges of automation, as illustrated by the loss of Air France 447 over the Atlantic Ocean in 2009–a disaster in part initiated by the simple disconnection of the aircraft's autopilot in level cruise due to uncertain data inputs, coupled with the crew's limited experience of manually recovering from flight upsets at high altitude, related to broader deskilling across the industry.[21]

Governing machines that learn: AI as sociotechnical systems

AI has the potential to transform healthcare in exciting and important ways but will also pose new risks to patient safety. To effectively manage these risks, the analysis and governance of AI safety must be treated as a fundamentally sociotechnical problem that considers the entire human, social and organisational infrastructure that AI technologies both emerge from and are embedded in. The field of AI is developing rapidly and in some instances new technologies, such as patient-facing symptom checkers, have already been deployed on a large scale with limited evaluation[22] in a light-touch regulatory environment.[23] Healthcare rapidly needs to develop governance systems, institutions and specialists with the expertise and resources to develop robust sociotechnical safety requirements and conduct integrated safety analysis of AI systems.[24] In particular, mechanisms are needed to detect, analyse and learn from the failures of emerging AI technologies, and to take preventative action before serious failures can cause harm to patients. At a minimum, AI systems should be subject to rigorous testing and require the publication of detailed safety cases[25] that explain and evidence how risks to patient safety have been managed both in the technical development and organisational implementation of a system—before those systems are used in patient care. Open publication of these reports will be particularly important in building trust and acceptance of new AI technologies. Even with rigorous testing, technologies still fail. AI systems should therefore incorporate the equivalent of airline 'black box' recorders to capture data related to safety events[13,19], and existing regulatory requirements, institutional structures and social agreements will need to be reformulated to ensure that all relevant data and information about AI safety issues can be openly shared, robustly investigated and systematically analysed to support accountability and learning. Due to the rapid pace of development in AI and its varied applications, as well as the continually-learning nature of some AI technologies,[26] systems of AI safety governance themselves will need to be adaptive, flexible and able to rapidly learn from

experience—and failure. Institutional infrastructures of learning will need to be built around these new technologies of learning.[27]

The development of safety governance systems for emerging AI technologies will need to be supported by a broad-based programme of interdisciplinary research. Some of the safety challenges associated with AI will echo similar challenges encountered in medical technology, health informatics, complex automation and organisational safety, but others will be novel and unique. Four initial lines of inquiry seem particularly urgent. First, it will be important to understand and survey the forms of sociotechnical risks that AI systems are likely to introduce into different arenas of healthcare, to develop a coherent map of the safety landscape. Second, it will be necessary to identify, develop and adapt models of sociotechnical safety—and associated analytical methods—that are well suited to explaining and governing the patient safety risks arising from AI. Third, it will be critical to explore patient, public and practitioner perspectives on the risks, benefits and acceptability of AI systems in healthcare, and the social and institutional mechanisms that may be required for acceptance and to build trust. Fourth, it will be necessary to move beyond high level principles and examine the practical regulatory functions and concrete governance mechanisms that are best suited to assuring the safety of emerging AI systems, and that can accommodate the unique challenges of governing technologies that may have the potential to act autonomously and continuously learn and adapt. Many of these challenges are being actively explored in other sectors, such as the nascent autonomous vehicle industry [27,28] which has already experience high-profile safety failures and fatalities[29]. It will therefore be important for research and policy in healthcare to both contribute to and learn from these broader experiments and debates in AI safety and governance. By drawing attention to some key technological risks posed by machine learning, Challen *et al*[9] provide an important step forward and help to start a new conversation about the governance of AI safety in healthcare. A core premise of this debate must be that to effectively govern the safety of AI it

will be necessary to understand the vulnerabilities of the human and organisational systems that create and interact with AI—just as it will be necessary to understand the vulnerabilities of AI technologies themselves.

References

1. Bahl M, Barzilay R, Yedidia A, Locascio N, Yu L, Lehman C. High-Risk Breast Lesions: A Machine Learning Model to Predict Pathologic Upgrade and Reduce Unnecessary Surgical Excision. *Radiology*. 2018;286(3):810-818.

2. DeepMind. Applying Machine Learning to Radiotherapy Planning for Head & Neck Cancer. https://deepmind.com/blog/applying-machine-learning-radiotherapy-planning-head-neck-cancer/ (Accessed August 2018).

3. Norgeot B, Glicksberg BS, Trupin L, et al. Assessment of a Deep Learning Model Based on Electronic Health Record Data to Forecast Clinical Outcomes in Patients With Rheumatoid Arthritis. *JAMA Netw Open*. 2019;2(3):e190606.

4. Topol EJ. High-performance medicine: the convergence of human and artificial intelligence. *Nat Med*, 2019;*25*(1):1–13.

5. Crawford K, Calo R. There is a blind spot in AI research. *Nature*. 2016;538(7625):311–313.

6. Cabitza F, Rasoini R, Rasoini R, Gensini GF, Gensini GF. Unintended Consequences of Machine Learning in Medicine. *JAMA*. 2017;318(6):517-518.

7. Hutson M. Has artificial intelligence become alchemy? *Science*, 2018;360(6388), 478–478.

8. Burrell J. How the machine “thinks”: Understanding opacity in machine learning algorithms. *Big Data Soc*. 2016;3(1):1-12.

9. Challen R, Denny J, Pitt M*, et al* Artificial intelligence, bias and clinical safety. *BMJ Qual Saf*. 2019;28:231-237.

10. Lehr D, Ohm P. Playing with the Data: What Legal Scholars Should Learn About Machine Learning, *UC Davis L Rev* 2017;51(2):653-717.

11. Esteva A, Kuprel B, Novoa RA, *et al*. Dermatologist-level classification of skin cancer with deep neural networks. *Nature* 2017;542:115–8.

12. Esteva A, Robicquet A, Ramsundar B, Kuleshov V, DePristo M, Chou K, *et al*. A guide to deep learning in healthcare. *Nat Med*, 2018;*25*(1):1–6.

13. Bryson J, Winfield A. Standardizing Ethical Design for Artificial Intelligence and Autonomous Systems. *Computer*, 2017;*50*(5):116–119.

14. Joshi I, Morley J. Ethics is our competitive advantage: how the NHS can lead the world in AI based healthtech. Department of Health and Social Care Technology in the NHS Blog https://healthtech.blog.gov.uk/2019/02/20/ethics-is-our-competitive-advantage-how-the-nhs-can-lead-the-world-in-ai-based-healthtech/ (Accessed Feb 2019)

15. Department of Health and Social Care. Code of conduct for data-driven health and care technology. https://www.gov.uk/government/publications/code-of-conduct-for-data-driven-health-and-care-technology/initial-code-of-conduct-for-data-driven-health-and-care-technology (Accessed March 2019)

16. Kroll JA. The fallacy of inscrutability. *Phil Trans R Soc A*. 2018;376:1-14.

17. Macrae C, Vincent C. Learning from failure: the need for independent safety investigation in healthcare. *J R Soc Med*. 2014;107(11):439–43.

18. Marcus G. Deep learning: a critical appraisal. arXiv [cs.AI]. 2018. Available: https://arxiv.org/abs/1801.00631

19. Winfield AFT, Jirotka M. Ethical governance is essential to building trust in robotics and artificial intelligence systems, *Philos Trans A Math Phys Eng Sci* 2018;376.

20. Susskind R, Susskind D. The Future of the Professions: How Technology Will Transform the Work of Human Experts. Oxford University Press, 2015.

21. Langeweische L. The human factor. Vanity Fair, Sep 2014. https://www.vanityfair.com/news/business/2014/10/air-france-flight-447-crash (Accessed Feb 2019)

22. Fraser H, Coiera E, Wong D. Safety of patient-facing digital symptom checkers. *Lancet*. 2018;*392*(10161):2263–2264.

23. McCartney M. AI in medicine must be rigorously tested. BMJ 2018;361:k1752.

24. Coiera EW, Kidd MR, Haikerwa MC. A call for national e-health clinical safety governance. *Med J Aust*. 2012;196(7):430–431.

25. Sujan MA, Habli I, Kelly TP, Pozzi S, Johnson CW. Should healthcare providers do safety cases? Lessons from a cross-industry review of safety case practices. *Saf Sci 2016*;*84*:181–189.

26. Yu KH, Kohane IS. Framing the challenges of artificial intelligence in medicine. BMJ Qual Saf. 2019;28(3):238-241.

27. Stilgoe J. Machine learning, social learning and the governance of self-driving cars. *Soc Stud Sci* 2018;*48*(1):25–56.

28. Cummings, M. The Brave New World of Driverless Cars. *TR News*. 2017;308:34-37.

29. National Transportation Safety Board. Preliminary report for Crash Involving Pedestrian, Uber Technologies, Inc., Test Vehicle, March 18, 2018. Washington, DC: NTSB, 2018.