When the frameworks don't work: data protection, trust and artificial intelligence

doi:10.1136/medethics-2022-108263

Zoë Fritz

With new technologies come new ethical (and legal) challenges. Often, we can apply previously established principles, even though it may take some time to fully understand the detail of the new technology - or the questions that arise from it. The International Commission on Radiological Protection, for example, was founded in 1928 and has based its advice on balancing the radiation exposure associated with X-rays and CT scans with the diagnostic benefits of the new investigations. They have regularly updated their advice as evidence has accumulated and technologies have changed,[1] and have been able to extrapolate from well-established ethical principles.

Other new technologies lend themselves less well to off-the-peg ethical solutions. In several articles in this edition the ethical challenges associated with the use of artificial intelligence (AI) in medicine are addressed. Although multiple ethical codes and guidelines have been written on the use and development of AI, Hagendorf noted that many of them reiterated a '*deontologically oriented, action-restricting ethic based on universal abidance of principles and rules*'.[2] Applying pre-existing ethical frameworks to artificial intelligence is problematic for several reasons. In particular, AI has two characteristics which are very different from the current clinical practice on which traditional medical ethics are based:

1. The so called 'black box' of deep learning, whereby a deep neural network is trained to iteratively adapt to make better and better interpretations across layers of complex and non-linear data.[3] The resulting (and ever adapting) algorithms are generally too complex to interpret or explain, meaning that part of the process that is being used is opaque even to the users.[4] This makes it difficult if not impossible to adhere to principles of transparency and informed consent, and restricts the autonomy of the users (both clinicians and patients).
2. Each element of AI has been developed to achieve a particular goal – set by its creators – but has no 'intent' beyond achieving that goal. Ethical analyses which include considerations of broader motives or virtuous qualities can therefore not be applied in relation to the AI.

These issues are highlighted by S Lee in a student essay.[5] Lee examines the NHS code of conduct for artificial intelligence-driven technology, and in particular looks at the conceptualisation of trust within this particular piece of ethical governance. He draws out the challenge of establishing a trust which is '*rationally justified on sound epistemological bases*' in the context of the 'black box' of deep learning. He notes that '*the Code assumes users are able to and will justify trust by weighing up risk and competence, where risk is the probability of an AI being incompetent at the function it is specified to fulfil, based on performative (ie, quantitative empirical data) information.*'

To fulfil this, he suggests, the data used to train the AI would need to be available to all users, in order for them to judge the risk of bias[6] and other built-in errors in the algorithms developed. This is clearly impractical. So he suggests that*: 'to foster trust, developers and decision makers should provide information of how they encapsulate the interests of users; they should show their values are aligned with the users.'* In other words we cannot apply the models of trust which are established in the doctor-patient relationship to AI, and so we need to turn to the developers and decision makers: we should assess their intent and competence when designing the system. Lee proposes that a seventh requirement is added to the code relating to the ethical conduct and history of the developers.

Turning away from broad ethical codes, Sorell *et al*[7] examine the tension between traditional data ethics and governance (where the emphasis is on minimisation of personal data collection, processing and sharing) and AI (whose success is dependent on maximal data). They focus their attention on Computational Pathology, where machine learning is applied to digitised whole slide images to improve pattern recognition of cancer presence, progression and prognosis. They draw attention to the mismatch between the motivations behind laws to protect both personal data and pathological samples (where the focus is on not using data or samples beyond direct benefit to the individual without their explicit consent) and the application of these laws to AI.

They argue that '*Stereotypical risks of privacy violation occur where data enables inferences about identifiable people's current health, wealth, sexual practices, political affiliations and friendships. These inferences may allow individuals or organisations to manipulate data subjects or make an economic gain from information about them*'. Where data has been aggregated, '*it cannot typically be used to identify the data subjects, or disadvantage them…So while deidentification may not amount to out and out anonymisation in the sense of GDPR, it may amount to anonymisation for most practical purposes.*' Here then, the standards applied to standard data sharing – of full transparency and of explicit consent - are almost impossible to achieve and are antithetical to the goal of improved population health. In fact '*the larger the data sets used for training and validation, the lower false positive and negative rates are likely to be, other things being equal, with corresponding clinical advantages.*' A different data governance framework is needed for the development of Computational Pathology and other AI dependent diagnostic tools, one which recognises the population benefits of data sharing in this context.

Finally, Kempt and Nagel[8] (and associated commentary authors) discuss the proposal to use artificial intelligent decision support systems (AI-DSS) as providers of second opinions in medical diagnostics, and again the issue of the 'black box' comes into play. The authors state: '*The difference in evidence-processing and lack of explainability renders an AI-DSS largely accurate but unchallengeable. Conflicts between human initial opinions and AI-second*

opinions, therefore, may not qualify as peer-disagreements, as its 'reasoning' is not reason-based like an expert's evaluation. Hence, we cannot peer-disagree with an AI-DSS, leaving a responsibility gap when trying to decide what to do in case of a conflict.' They propose a 'rule of disagreement' whereby if the AI-DSS confirms the initial opinion no further steps need to be taken, but, where there is disagreement, a second human opinion must be sought; the final arbitrator is the initial clinician.

These papers are rich in not only presenting the ethical challenges associated with various elements of AI but in proffering well-reasoned bespoke ethical solutions.

Funding ZF is funded by the Wellcome Trust (grant 208213/Z/17/Z). She is based in The Healthcare Improvement Studies Institute (THIS Institute), University of Cambridge. THIS Institute is supported by the Health Foundation, an independent charity committed to bringing about better health and healthcare for people in the UK.

Competing interests None declared.

Patient consent for publication Not applicable.

Ethics approval Not applicable.

Provenance and peer review Not commissioned; internally peer reviewed.

ORCID iD

Zoë Fritz http://orcid.org/0000-0001-9403-409X

REFERENCES

1 Bochud F, Cantone MC, Applegate K, *et al*. Ethical aspects in the use of radiation in medicine: update from ICRP task group 109. *Ann ICRP* 2020;49(1_suppl):143–53.
2 Hagendorff T. The ethics of AI ethics: an evaluation of guidelines. *Minds Mach* 2020;30(1):99–120.
3 Hinton G. Deep Learning-A technology with the potential to transform health care. *JAMA* 2018;320(11):1101–2.
4 Humphreys P. The philosophical novelty of computer simulation methods. *Synthese* 2009;169(3):615–26.
5 Lee SS. Philosophical evaluation of the conceptualisation of trust in the NHS' code of conduct for artificial intelligence-driven technology. *J Med Ethics* 2022;48:272–7.
6 Parikh RB, Teeple S, Navathe AS. Addressing bias in artificial intelligence in health care. *JAMA* 2019;322(24):2377–8.
7 Sorell T, Rajpoot N, Verrill C. Ethical issues in computational pathology. *J Med Ethics* 2022;48:278–84.
8 Kempt H, Nagel SK. Responsibility, second opinions and peer-disagreement: ethical and epistemological challenges of using AI in clinical diagnostic contexts. *J Med Ethics* 2022;48:222–9.